Les membres du jury qui ont évalué ce mémoire ont demandé des corrections mineures.

HEC Montréal

Modélisation par analyse de sujets et analyse de graphes dans le cadre du droit du travail québécois

Par

Adam Paul Ihaddadene-Salzgeber

Science de la gestion

(Intelligence d'affaires)

Mémoire présenté en vue de l'obtention

Du grade de maîtrise ès sciences (M. Sc.)

Gilles Caporossi

HEC Montréal

Directeur de recherche

Mai 2025 : date de dépôt du mémoire

Remerciements

Je suis reconnaissant envers Gilles Caporossi pour sa disponibilité, sa patience et les précieux conseils qu'il m'a dispensés au cours de ma recherche. Je souhaite aussi souligner le rôle de Florian Carichon qui a éveillé mon intérêt pour ce mémoire via le cours d'analyse des données textuelles. Ensuite, il est important de souligner le rôle du groupe de recherche en NLP *(Natural Language Processing*) qui m'a permis d'approfondir mes recherches sur mon sujet. Finalement, j'adresse des remerciements à mes proches qui m'ont soutenu dans ce projet, même dans les moments les plus difficiles de ce dernier.

Table des matières

I) Introduction

Le monde des affaires est de plus en plus confronté aux progrès et aux évolutions rapides des normes quant au droit du travail. Par exemple, dans les dernières années, la crise de la COVID-19 et l'introduction de masse du télétravail sont tant de facteurs qui ont requestionné le cadre d'un accident au travail et la responsabilité. En outre, dans un rapport annuel de 2022, le gouvernement du Canada reconnait une hausse des accidents invalidants du travail de 10,9 % (gouvernement du Canada, 2022) en comparaison avec l'année antérieure pour les emplois relevant de la compétence fédérale. Ces éléments amènent alors les équipes juridiques des entreprises ainsi que les syndicats représentant les travailleurs à accroître leurs travaux pour prévenir et plaider leurs cas devant les instances. Dans ce sens, *Thomson Reuters* (Thomson Reuters, 2022), dans un mémoire de 2022, souligne que 65 % des personnes qu'ils ont interrogées, travaillant dans des départements légaux d'entreprises, remarquent une augmentation de la charge de travail. Dans cette même étude, les répondants regrettent un retard dans la technologie qui les assiste dans leur travail. Si l'avocat ou le juge joue un rôle prépondérant dans le processus juridique, une partie essentielle du travail repose sur la recherche en vue de constituer des arguments pour la défense ou la prévention d'infractions potentielles. Alors que des progrès majeurs dans les technologies du langage sont à souligner au cours des dernières décennies, il est légitime de se demander comment nous pourrions assister les parties prenantes avec des outils adaptés à leurs besoins.

En science des données, depuis les années 1950, une approche à l'intersection entre la linguistique et l'intelligence artificielle est utilisée pour traiter les données textuelles : il s'agit du *natural language processing* (NLP) (P. M. Nadkarni, L. Ohno-Machado, W. Chapman, 2011). Si, à l'origine, le NLP se concentrait simplement sur la recherche d'informations, aujourd'hui, la communauté de chercheurs a développé de nombreuses méthodes permettant d'effectuer plusieurs tâches reliées à l'analyse de textes. Parmi celles

qui nous intéressent, on retrouve la classification de documents via des algorithmes non supervisés. Ainsi, les différents outils de recherche de documentation au Canada et au Québec, comme CaNLII ou SOQUIJ, s'appuient sur des algorithmes de recherche basés sur les principes du NLP. Cependant, l'utilisation de ces moteurs de recherche peut se montrer technique et complexe pour les recherchistes juridiques.

L'objectif de ce travail est de combiner plusieurs techniques pour proposer une recherche et une visualisation des jugements permettant de retrouver plus facilement l'information pertinente. En prime, nous regardons si nos méthodes peuvent contribuer à révéler des informations clés aidant les preneurs de décision en matière de normes de santé et sécurité au travail. Si nous avons opté pour une représentation des sujets couverts par les textes via l'allocation latente de Dirichlet (LDA) et la modélisation de réseau de graphes des citations du corpus, nous expliquons dans ce mémoire les raisons derrière ces choix.

Pour aborder ces tâches, nous devons commencer par regarder l'état de la recherche sur le sujet. Dans cette optique, nous regardons tant des articles généralistes sur le NLP que des articles plus précis sur les données textuelles juridiques. Dans cette revue de littérature, nous voyons les différentes approches existantes et nous nous inspirons de ces dernières pour créer nos modèles.

Ensuite, nous expliquons notre méthode de travail en détaillant plus spécifiquement les dimensions que nous devons mesurer et explorer pour rendre ce dernier pertinent. Dans cette partie, nous nous attardons aussi sur les spécificités du domaine légal québécois et les défis que cela soulève. Aussi, nous regardons notre méthode de collecte de données en soulignant les potentiels biais que peut comporter cette dernière.

Après cela, nous procédons à l'exposition de nos analyses pour chaque modèle. Nous commençons ici par décrire nos données via une exploration de ces dernières. Ensuite, nous expliquons en détail comment nous avons procédé à notre modélisation de sujet, comment nous avons nettoyé nos données et procédé à la sélection du nombre de sujets. Enfin, nous présentons nos données de réseau de graphes pour mieux exposer la structure de ce dernier.

Nous poursuivons alors ce travail en présentant les résultats de chaque méthode. Pour la modélisation via le LDA, nous regardons comment sont composés les différents sujets qui composent notre corpus et différents aspects relatifs à ces derniers. Puis, nous regardons comment le graphe complet est structuré en nous attardant plus particulièrement sur les communautés de nœuds qui en font partie en les interprétant. Finalement, nous essayons de réunir les deux méthodes via une démarche que nous qualifions d'approche mixte.

Nous finirons ce mémoire par une partie sur les limites et les améliorations futures à apporter à ce travail. Ce chapitre nous permet de détailler l'importance et le potentiel que porte l'intelligence légale assistée par les données. C'est alors après cela que nous concluons ce travail en résumant les principaux éléments que ce travail a amenés.

II) Revue de littérature

La grande majorité des données issues du système judiciaire sont de nature textuelle. De ce constat, c'est naturellement que les outils hérités du NLP se sont imposés comme une évidence pour apporter des instruments d'intelligence artificielle dans le domaine juridique. Si on peut remonter au début de la deuxième moitié du XXe siècle pour voir apparaitre les premières approches computationnelles du droit (Kort, 1957 ; Ulmer, 1963 ; Nagel, 1965 ; Segal, 1984 ; Gardner, 1984), c'est dans les dernières décennies que l'on a observé une accélération de la recherche sur le sujet. Ce phénomène peut s'expliquer par l'augmentation considérable des capacités de calcul, qui s'est accompagnée d'avancées majeures dans le champ de l'apprentissage machine.

Plusieurs applications visant à réduire le temps passé à effectuer des tâches fastidieuses pour les experts ou les novices sont envisagées. Dans l'article *How Does NLP Benefit Legal System : A Summary of Legal Artificial Intelligence* (Zhong, 2020), les auteurs suggèrent un cadre pour classifier les approches dans le domaine. Ils proposent deux grandes familles de méthodes pour regrouper les différentes recherches dans le domaine. La première s'appuie sur les méthodes basées sur les symboles (symbol-based methods). Dans le NLP, cette méthodologie s'appuie sur l'analyse des règles explicites et des

représentations formelles pour comprendre et analyser le langage. La deuxième repose sur les modèles d'embeddings, qui sont des représentations denses et continues des mots ou des phrases sous forme de vecteurs. Ici, on considère les mots ou les phrases comme des variables discrètes. Fréquemment liée à l'apprentissage automatique, cette méthodologie permet d'obtenir des résultats remarquables. Cependant, contrairement aux méthodes basées sur les symboles, l'interprétation de ces dernières s'avère complexe, voire irréalisable. Si cette catégorisation constitue une amorce de cadre conceptuel, nous proposons d'ajouter une autre approche méthodologique, l'approche probabiliste. Cette méthode quantitative propose d'analyser les textes via l'étude de la fréquence et les probabilités des mots, phrases ou autres unités issues des textes. Une variété de techniques appartient à cette catégorie, par exemple TF-IDF (*term frequency-inverse document frequency*), qui, à travers la fréquence des termes dans un corpus, permet l'extraction ou la recherche d'informations ou encore les tâches de résumé automatique (Pand, 2019 ; Basu, 2024). On peut aussi mettre de l'avant l'allocation latente de Dirichlet (LDA), l'analyse sémantique latente (LSA) ou la factorisation par matrices non négatives (NMF), qui, quoiqu'ayant des contraintes mathématiques différentes, proposent différentes manières de modéliser les sujets d'un texte (Didwania, 2024 ; Bermylle, 2022 ; Budahazy, 2021). Nous proposons cette classification, car de nombreux articles continuent à être publiés utilisant ces méthodes statistiques. Ces démarches résistent à l'avènement du *machine learning* grâce à leurs coûts de calcul raisonnables et à leurs interprétabilités plus faciles. Si cette topologie nous permet de poser un cadre d'analyse, il est important de souligner que chacune de ces catégories n'est pas pour autant exclusive. En effet, on retrouve plusieurs articles où des méthodes *d'embeddings* sont mélangées avec des approches basées sur les symboles ou des approches statistiques ; on peut même parfois voir les trois s'entremêler (Wehnert, 2021 ; Basu, 2024).

Le cadre conceptuel étayé précédemment regroupe l'ensemble des outils convoqués par la recherche afin de développer des applications pour le domaine. Dans l'article *The Law and NLP : Bridging Disciplinary Disconnect* (Mahari, 2023), les auteurs reprennent le cadre de Zhong (Zhong, 2020) pour apporter une approche critique de l'état de

l'art actuel dans le domaine. Les auteurs dénoncent le manque d'utilisation des outils développés par la recherche dans la pratique actuelle du droit. Afin de comprendre l'écart entre les connaissances et la pratique, ils analysent les articles cités dans cette publication de référence. De cette analyse, ils résument le cadre des applications du NLP dans le droit en trois classes.

La première est relative aux applications existantes d'utilité pour la communauté de professionnels. Dans cette catégorie, on compte plusieurs tâches que la recherche a déjà explorées. De cela, on peut voir plusieurs cas d'applications qui ont fait leurs preuves dans le temps.

On peut évoquer la création de documents juridiques ou d'analyses. En effet, une grande partie du travail des juristes réside dans la production de documents officiels ou de résumés. C'est pour cela que la recherche s'est penchée sur cette question en utilisant un mélange d'approches statistiques et *d'embeddings* (Basu, 2024 ; Devaraj, 2023). Il est important de souligner qu'avec les nouveaux modèles de langue généraux basés sur des *embeddings* comme GPT-4, on voit des performances convenables pour l'extraction d'informations afin de générer des résumés cohérents (Deroy, 2024 ; Greco, 2024).

Ensuite, les outils de recherche sémantique sont aujourd'hui régulièrement utilisés par la communauté dans les moteurs de recherches juridiques, par exemple. En effet, la création de plaidoyers ou d'arguments juridiques repose sur des précédents. La recherche de jugements et de citations constitue alors une tâche importante des équipes juridiques. La recherche propose plusieurs approches pour assister les professionnels. Que ce soit par la modélisation de sujets dans les documents (Didwania, 2024 ; Bermylle, 2022 ; Budahazy, 2021) ou alors l'extraction de citations dans des décisions antérieures (Huang, 2021 ; Tang, 2021), ces applications répondent à un besoin concret du système de justice.

Finalement, les applications qui cherchent à vulgariser le langage juridique sont importantes dans l'activité des documentalistes et des avocats. Le droit est composé d'un vocabulaire et d'un jargon spécifiques qui demandent une connaissance approfondie des sujets abordés. Si l'utilisation des graphes de connaissances (*knowledge graphs*) est une

approche utilisée dans de nombreux domaines, aucun travail n'existe à cette fin dans le droit (Zhong, 2021). Dans ce cadre, nous retrouvons les outils de résumé automatique. Ici, le cas d'utilisation est différent, car l'objectif est de simplifier l'analyse d'un jugement ou d'un contrat, par exemple (Huang et al., 2021 ; Tang et Clematide, 2021). D'autres applications de *chatbot* ou de *Q&A* sont aussi envisagées, permettant de clarifier un concept rapidement (Khazaeli, 2021). On peut aussi inclure dans cette classe les applications qui s'inspirent des méthodes basées sur les symboles. En effet, les tâches de détection d'éléments (Akbik, 2019 ; Kuru, 2016), d'extraction de temporalité d'évènements (Nguyen, 2018 ; 2016), ou encore l'extraction de relation (Zeng, 2015 ; Christopoulou, 2018). Ces outils permettent aux utilisateurs de mieux appréhender un ensemble de documents complexes, tant pour la communauté de pratique que pour les novices dans le domaine.

La deuxième classe d'applications regroupe plusieurs articles qui n'ont pas d'écho dans la communauté de pratique juridique (Devaraj, 2023). Les raisons majeures avancées qui expliquent ce manque d'impact sont l'inexplicabilité des modèles, le manque de transparence sur les données d'entraînement de ces derniers et l'absence de besoins de la communauté quant aux tâches résolues par ces applications. La principale application concernée par ces failles est la prévision d'issues de jugements. En effet, les avocats ou juges ne cherchent souvent pas à prévoir l'issue d'un jugement. En revanche, ils cherchent des arguments pour orienter le jugement sur un fait plutôt qu'un autre. Concernant les données d'entrainement, quand on a accès aux données d'entrainement de ces recherches, on peut se questionner sur les biais que peuvent présenter ces dernières. En effet, les éléments ethniques, sociaux et culturels peuvent avoir une colinéarité avec des issues de jugements. Ainsi, l'utilisation d'une machine affectée par ce biais constitue un danger pour la justice. Finalement, il est important de souligner que la confiance en la justice est un élément crucial de nos démocraties. De ce fait, les modèles « boîtes noires » peuvent s'assimiler à des modes de décision arbitraires pour le grand public, ce qui endommage la confiance dans nos institutions. Cependant, si ces applications n'ont pas d'utilité avérée dans la pratique du droit, l'aspect expérimental que portent ces démarches demeure important pour la communauté scientifique. On peut souligner que ces tâches constituent

un défi important qui peut faire avancer la recherche dans le NLP (Chalkidis, 2022 ; Guha, 2023). Outre cela, si certaines applications ont peu d'utilités pour les juristes, elles peuvent nous fournir des points de vue intéressants sur le domaine, comme la détection de biais dans le langage (Rice, 2019).

La troisième et dernière catégorie concerne les applications n'ayant pas eu d'attention suffisante pour le moment. Ces cas d'applications sont inspirés par une étude de l'école de droit de Harvard qui a interrogé plusieurs avocats de pays différents en leur demandant quelles applications du NLP leur seraient utiles. Cette étude révèle plusieurs besoins des praticiens du droit qui peuvent être explorés par la recherche actuelle. L'article révèle trois cas d'applications qui sont sous-utilisés ou non exploités dans le paysage des outils juridiques.

L'argumentation étant au centre de la démarche de plaidoirie, il est pertinent de s'intéresser aux outils permettant d'identifier les arguments les plus pertinents à la plaidoirie. Pour cela, il est nécessaire d'intégrer des variables comme la composition du jury, le juge qui s'occupe du cas ou encore la nature du cas. Tous ces facteurs peuvent avoir un impact sur les arguments les plus pertinents à solliciter en tant que membre des équipes juridiques. Si des études en NLP sur l'argumentation contextuelle ont été proposées, aucun travail n'a été adapté au droit. Si cette piste constitue une voie de développement pour de futurs travaux, notre travail ne se penche pas sur ce cas spécifique.

Dans le cadre de la pratique, les cabinets d'avocats constituent de larges bases de données constituées de documents passés. Dans ces bases de données privées, la recherche d'information afin de générer de nouveaux documents est une tâche non assistée souvent effectuée par les avocats eux-mêmes. L'enjeu ici est de retrouver les documents pertinents tout en conservant la confidentialité d'information sensible à ne pas divulguer à de tierces parties ou au public. Le *retrieval-augmented generation* est une méthode qui semble adaptée, mais nécessite des ajustements pour se conformer au cadre juridique.

Finalement, la dernière tâche concerne les outils de recherches juridiques orientés vers la pratique des avocats. Afin de préparer les dossiers, les avocats doivent identifier les

cas précédents, la jurisprudence, qui permette d'appuyer leurs plaidoiries. L'enjeu réside dans la grande quantité de jurisprudence disponible qui nécessite une recherche avancée afin de construire l'argumentaire le plus pertinent. Une bonne décision est une décision qui est citée régulièrement, qui correspond le plus possible au cas que nous soutenons et finalement qui n'est pas renversée par une autre instance. En relation avec cela, trouver des passages précis dans la jurisprudence afin de diminuer la taille de l'échantillon de recherche pour les professionnels. Dans l'article, les auteurs soulignent des travaux de prédiction de passages légaux qui, combinés avec des outils de résumés automatiques, sont des pistes pour effectuer cette tâche. Dans son article de 2021, Robert Mahari (Mahari, 2021) propose d'affiner le modèle préentraîné proposé par Chalkidis *Legal-BERT* (Chalkidis et al., 2020) pour répondre à une tâche de prédiction de citations de textes importants. Utilisant des données annotées, ils parviennent à identifier ces passages en fonction de leurs pertinences. Dans un autre article de 2023, Robert Mahari (Mahari, 2023) propose de cumuler cette approche avec les outils de résumé automatique pour simplifier la recherche des avocats. La recherche a proposé des algorithmes de résumé automatique évalués avec la métrique ROUGE qui ont été appliqués au cas du droit américain (Bauer, 2023). Cependant, ces approches n'ont jamais été combinées pour proposer une application permettant d'accélérer la recherche d'arguments. Malgré les résultats prometteurs que nous promet l'utilisation de ces méthodes, les auteurs reconnaissent eux-mêmes les limites éthiques de ces approches supervisées. En effet, un modèle comme BERT rend impossible la compréhension du cheminement pour le choix de citations. Outre cela, les auteurs reconnaissent que les données d'entraînement amènent plusieurs biais. Parmi ces biais, on peut souligner que l'entraînement de ce modèle repose sur des opinions passées des tribunaux ignorant les facteurs sociologiques, idéologiques, etc. Ainsi, il pourrait véhiculer des opinions passées et les confirmer sans contester la posture des juges ou de la société sur certains sujets. De plus, ce conservatisme n'inciterait pas les avocats à poser de nouvelles questions de droit aux instances, ce qui est un rôle important de la justice dans notre société. Enfin, les avocats ont besoin de certitudes pour présenter des arguments face aux instances. De ce fait, ne pas comprendre le processus de décision constitue un élément

de défiance pour les avocats vis-à-vis de ces outils. La confiance est essentielle pour l'utilisation d'outils juridiques, car des décisions majeures sont en jeu pour les parties.

Notre travail s'intéresse aux outils de recherches de jurisprudence dans le droit du travail québécois pour des cas de maladies professionnelles. Si la recherche actuelle se concentre fortement sur les approches supervisées, nous voulons affirmer l'importance des approches non supervisées, notamment vis-à-vis de la neutralisation des biais liés à l'étiquetage des données. Au cours de l'élaboration de cette revue de littérature, nous avons remarqué que l'utilisation du LDA pour modéliser les sujets dans les jugements est efficace (Celard et al., 2024 ; Didwania, 2024). Cependant, ce modèle ne se concentre que sur une analyse sémantique des corpus. Pour compléter l'étude de corpus, nous proposons d'y adjoindre une analyse de graphe basée sur le réseau de références des citations de jurisprudences et de lois présent au sein du corpus. La combinaison de ces deux méthodes nous permet de proposer une nouvelle approche de la tâche d'appariement de documents juridiques au Québec, permettant de retrouver plus facilement les textes similaires dans le corpus. Au cours de l'élaboration de notre revue de littérature, nous avons trouvé peu de travaux s'intéressant à l'analyse de graphes pour les corpus juridiques. Malgré cela, quelques recherches proposent des méthodes de graphe ingénieuses pour révéler de l'information dans le cadre du droit. (Qiqi Wang et al., 2022) Proposent de regarder les liens entre les individus dans des cas de droit afin de simplifier l'analyse des juristes. Pour modéliser ce réseau, ils comparent des modèles de classification issus du *machine learning*, du *deep learning* et, finalement, des modèles préentrainés comme BERT. Cette extraction permet d'isoler les faits, les sujets, les liens de causalité et tout autre élément de nature logique. Cette méthode aboutit alors à la constitution de plusieurs types de graphes qui permettent la modélisation des interactions entre les entités et les actions, les entités et des concepts clés, et les éléments de droit. Dans le paysage des recherches sur le sujet, on retrouve aussi des démarches s'intéressant plutôt aux connexions entre les jugements via les citations de lois ou de précédents en leur sein. C'est dans ce sens que la recherche proposée par Paul Shounak (Paul Shounak et al., 2022) s'intéresse à l'identification du statut légal des documents juridiques, via l'utilisation d'un réseau de graphes des citations et des

références en soutien avec la structure des textes. Pour arriver à ce terme, ils utilisent des jugements de droit criminel indien, qu'ils associent avec un graphe de la structure du Code criminel. Ils proposent une structure hiérarchique de ce code de loi, qui est une imbrication d'actes, de chapitres, d'articles, de sujets, de sections et de faits. Via une analyse des jugements, ils s'attellent à isoler les faits de ces derniers avec le système d'encodeur *hierarchical attention network* (HAN) (Yang et al. 2016). HAN a la particularité d'intégrer la structure hiérarchique des textes dans son fonctionnement. En plus de cela, il porte attention aux mots et aux phrases dans la construction de la représentation dense des textes. La combinaison de l'analyse structurelle du texte et par le réseau est réalisé via une fonction de score qui nous permet d'obtenir un score d'attribut qui permet d'attribuer *l'embedding* du fait à une section, puis un score structurel qui attribue *l'embedding* de la structure du fait à une section, et finalement un score d'alignement qui attribue à *l'embedding* structurel des faits à la structure de *l'embedding* des sections. Cette expérience révèle de meilleurs scores de performances que des méthodes concurrentes, comme *LegalBERT*, par exemple. On peut aussi citer le travail de Yanrang Tang et al. (Yanran et al., 2023) qui, dans la même lignée, cherche à améliorer l'association de textes juridiques avec un modèle de *graph neural network* (GNN). En utilisant des données de la cour fédérale canadienne, ils appliquent un GNN sur les textes afin d'extraire la structure d'argumentaire des cas. Ces méthodes se penchent sur la structure interne des jurisprudences, mais ne regardent pas la structure des références entre les textes et les lois. Or, notamment dans le champ des publications scientifiques, nous avons vu plusieurs articles qui proposent d'utiliser les réseaux de citations pour plusieurs tâches, comme notamment la recommandation d'articles connexes (Hanwen et al., 2019 ; Masaki, 2019). Nous proposons ici d'utiliser une approche similaire pour le cas du droit québécois. Pour cela, nous expliquerons notre méthodologie et les particularités que portent nos données dans la prochaine partie.

III) Méthodologie

Dans cette section, nous allons expliquer en détail la méthodologie de ce mémoire. Pour cela, nous regardons le cas spécifique du droit du travail au Québec et les enjeux reliés à la modélisation dans ce contexte. Dans cette section, nous expliquons la raison du choix de ce domaine de droit spécifique et nous détaillerons plus en détail le fonctionnement des genres de droit québécois. N'ayant pas de bases de données prêtes à l'usage pour ce travail, nous détaillons dans une deuxième partie notre méthode de collecte de données en expliquant les possibles biais et enjeux que cette dernière peut comporter. Finalement, nous proposons de formaliser la tâche que nous voulons effectuer en expliquant les méthodes mathématiques et informatiques utilisées pour la résoudre.

III.1) Le cas du droit du travail québécois

Le droit occidental se divise en deux grandes familles, le droit civil et le droit de *common law*. Si nous ne décrivons pas en détail les éléments qui composent ces systèmes, nous devons expliquer la principale différence qui réside dans la manière de créer et d'appliquer la loi. Le droit civil s'appuie sur des codes de lois issus du système législatif. Il est important de notifier que les tribunaux ne sont pas liés entre eux par leurs décisions. Ainsi, si les précédents sont importants pour l'argumentation, ils ne constituent pas une injonction à agir d'une manière ou d'une autre de la part de l'instance qui juge. En *common law*, les tribunaux créent la loi par leurs décisions et ils sont liés par celles-ci. De ce fait, pour assurer une cohérence dans la prise de décision, les juges doivent suivre les décisions des instances supérieures. Au Québec, notre système est hybride et s'appuie sur les éléments du droit civil et de la *common law* en fonction de la juridiction. Cependant, la culture de *common law* a infusé la pratique judiciaire, et les précédents ont une importance cruciale dans la prise de décision des juges. Dans ce contexte, la recherche et l'identification de cas pertinents pour la plaidoirie constituent une tâche essentielle pour la préparation de dossiers. L'objectif de cette tâche est de citer les jugements et les passages pertinents de ces derniers pour convaincre les juges d'agir en cohérence avec leurs décisions passées. Si aujourd'hui les avocats ont des moteurs de recherche comme CanLII ou SOQUIJ pour

trouver des cas via des mots-clés, affiner le choix des cas les plus pertinents repose souvent sur les connaissances acquises des juristes. La problématique réside dans le temps attribué à l'identification des jugements les plus importants pour constituer un argumentaire persuasif. Le droit du travail est quant à lui administratif au Québec. Cette catégorie de droit concerne l'ensemble des règles qui régissent les relations entre les administrés, qui sont les employeurs et les employés, et l'administration publique en cause. L'administration qui est responsable des recours et des droits en matière de santé et sécurité au travail est la Commission des normes de l'équité de la santé et de la sécurité du travail (CNESST). Son rôle principal est de régir les relations entre les employeurs et leurs salariés afin de s'assurer que les normes sont respectées. En cas de litige ou de contestation, le Tribunal administratif du travail (TAT) est le tribunal administratif spécialisé dans ces matières. Afin d'assurer une juste représentation des intérêts des parties, le TAT s'appuie sur un ensemble de lois, de règlements et de précédents permettant de répondre aux contestations qui lui sont adressées. Il est essentiel pour les parties de s'appuyer sur ces éléments afin d'étayer et de défendre ses intérêts et de maximiser les chances de victoire face aux juges administratifs. Il est important de noter que les administrés ont toujours la possibilité de faire appel de ces décisions devant des cours d'instances supérieures en suivant un processus précis. Cependant, la première instance reste le TAT au Québec en matière de contestation en santé et sécurité au travail.

Cette brève description du cadre juridique québécois nous révèle l'importance de la tâche d'appariement des cas pour constituer le plaidoyer et la veille stratégique pour les parties. Si, au cours de l'élaboration du sujet de ce mémoire, la question de l'analyse de cas juridiques via le NLP s'est avérée être le sujet central, nous avons vite réalisé que la source des jugements était large. En effet, de nombreux domaines du droit existent avec des logiques et des vocabulaires propres à chacun d'entre eux. De ce fait, baser notre corpus sur une grande diversité de domaines pourrait mener à des problèmes relatifs à la grande diversité du domaine. Afin de simplifier notre travail, nous avons décidé de nous baser sur le droit du travail et plus spécifiquement sur les accidents et les maladies professionnelles pour plusieurs raisons. La première est liée à l'importance que revêtent ces dossiers pour

les entreprises. Notre travail se déroule dans le cadre du programme en intelligence d'affaires, il est essentiel que notre angle de recherche soit en lien avec des enjeux d'affaires auxquels font face les entreprises. La santé et la sécurité au travail sont un domaine qui touche particulièrement les entreprises et les travailleurs, et ce, de manière potentiellement durable. Proposer d'obtenir une meilleure intelligence sur ces cas s'inscrit dans une volonté de mettre en place de meilleurs outils de plaidoiries, certes, mais aussi de proposer de nouvelles données pour être proactif dans le secteur afin d'assurer une meilleure veille stratégique sur le sujet. Ces éléments peuvent amener des gains économiques et des gains de bien-être au sein des entreprises. La deuxième raison réside dans les ressources limitées à notre disposition. Devant constituer nos données par nos propres moyens, il était impossible de traiter des centaines de milliers de textes pour constituer notre corpus. Les jugements étant de longs documents, les moyens de calcul nécessaires au traitement des données peuvent vite devenir exponentiels. En recentrant notre travail sur un secteur spécifique, cela nous permet de limiter la dimensionnalité de notre jeu de données, facilitant l'utilisation d'algorithmes. La troisième raison concerne l'homogénéité du vocabulaire de notre corpus. En effet, en choisissant ce domaine, on s'assure d'avoir une homogénéité des cas nous permettant de mieux saisir les logiques du tribunal sur les décisions prises. De plus, le cadre des maladies professionnelles est bien défini par la loi sur les accidents du travail (LATMP), ce qui facilite l'interprétation de nos résultats. Finalement, le fait qu'une instance centralise les contestations des décisions de la CNESST, le TAT, facilite l'obtention des données. En effet, la loi instituant le tribunal administratif du travail (LITAT), via ses articles 4 et 6, donne des compétences exclusives au TAT en matière de contestation concernant les maladies professionnelles vis-à-vis des décisions de la CNESST. Si des appels sont cependant possibles sur des instances supérieures, on peut considérer que la majorité des cas qui se rapportent à ces sujets sont reçus par cette instance. Cela facilite la construction de nos données pour nos analyses en ne considérant que les décisions émises par ce tribunal.

Si ces choix ont été pris pour faciliter notre travail, ils comportent aussi des limites liées à la taille limitée de l'échantillon dans le temps, qui sont discutées plus tard dans notre section sur les limites et les travaux futurs.

Ces éléments endémiques à notre objet d'analyse étant étayés, nous procédons à la collecte de nos données pour constituer notre corpus d'étude. La prochaine partie expose notre approche pour constituer nos données.

III.2) Collecte de données

Comme décrit précédemment, nous avons décidé de nous concentrer sur les contestations dans les cas de maladies professionnelles au TAT. Deux grands sites internet proposent de mettre en ligne des données publiques de jugements canadiens, CanLII et SOQUIJ. Ces deux plateformes proposent des moteurs de recherche basés sur des requêtes par mots-clés pour rechercher des jurisprudences dans leurs bases de données. En plus de cela, ces moteurs proposent des outils de filtre permettant de choisir des juridictions, des dates, des numéros de cas, etc. Nous avons ici opté pour SOQUIJ du fait de notre expérience passée avec cette plateforme et de l'accès facilité via le portail institutionnel de HEC Montréal. En utilisant la requête "maladie professionnelle", nous obtenons plus de 200 000 jugements toutes juridictions confondues et plus de 100 000 pour le tribunal du travail uniquement. Afin de diminuer ce chiffre, nous avons décidé d'appliquer des filtres à notre recherche. Notre choix de filtre s'est porté sur une fourchette de dates dans laquelle nous allons prendre nos cas. Pour toutes considérations pratiques, nous avons décidé de prendre une date antérieure à la crise de la COVID-19. En effet, cette période peut contenir des biais sur les données. On peut par exemple suggérer que le rythme des audiences a été ralenti à cause des différentes mesures de confinement et d'écartement social. Plus généralement, afin d'éviter tout biais relié à cette période particulière, nous sélectionnons une plage temporelle de recherche à une période antérieure à 2020. À des fins de cohérence, nous voulons que ces données soient issues d'une année civile de jugement. Par conséquent, nous cherchons des données publiées entre le 1er janvier d'une année en cours et le 31 décembre de l'année étudiée. Avec cet ensemble de filtres, nous cherchons une année

qui contient suffisamment de données pour que les modèles que nous utilisons soient robustes. Néanmoins, nous ne voulons pas avoir une année avec un trop grand nombre de données, car les données juridiques sont des données textuelles longues en comparaison des données souvent utilisées en NLP, comme des commentaires de forum, par exemple. Finalement, nous souhaitons tout de même collecter des données relativement récentes pour travailler sur des décisions le plus possible en accord avec l'état du droit actuel. De ce fait, nous décidons de ne pas considérer des données antérieures à 2013. En prenant en compte toutes ces contraintes, nous nous sommes arrêtés sur l'année 2017, qui comporte 5 403 jugements relatifs aux maladies professionnelles pour le tribunal administratif du travail dans la division santé et sécurité. Nous avons donc procédé au téléchargement des données au format PDF pour notre travail. À la suite du téléchargement des données, nous avons dû travailler sur leurs formats, car les PDF ne sont pas utilisables en l'état pour Python. Nous avons donc élaboré un script en utilisant le paquet Py2pdf pour les convertir en données txt. Ces manipulations ont comporté des enjeux reliés à la mise en page des documents qui ne se transcrivent pas correctement en format texte. Ce problème est apparu dans les textes qui contiennent des tableaux. Nous les avons isolés pour les traiter à part. Finalement, nous les avons classés selon la nomenclature du tribunal, qui suit la norme suivante : 2017QCTAT1 soit l'année du cas, la juridiction et le numéro de ce cas.

Si cette méthode nous permet d'obtenir des données exploitables, nous devons tout de même souligner que nous sommes passés par un algorithme de recherche pour les obtenir. Cela implique que nous travaillons sur des données qui ont déjà été triées par un modèle. De plus, nous ne connaissons pas le fonctionnement détaillé de cet algorithme, ce qui constitue un biais. Cependant, notre travail n'est pas centré sur la recherche de cas relevant des maladies professionnelles, mais sur la proximité des cas dans ce domaine. Outre cela, faute de meilleure méthode de collecte de données, nous avons pris les outils disponibles les plus adaptés à notre cas. Ainsi, si nous soulignons ce facteur, son impact sur notre travail est minime, quoique nous devions signifier son existence.

III.3) Formalisation de la tâche

Comme énoncé précédemment, nous souhaitons appareiller les textes les plus proches les uns des autres. Nous considérons deux grandes dimensions dans cette tâche, une dimension sémantique et une dimension relative à la logique juridique. Mesurer ces deux dimensions constitue des sous-tâches nécessaires à la résolution de notre problème de recherche.

III.3.a) La dimension sémantique

La sémantique est l'étude de la signification linguistique dans l'expression des langues. Cette branche linguistique se penche sur la construction des phrases et des idées via les mots. À ces fins, Ferdinand de Saussure propose d'étudier les signes dans les langues comme un système interconnecté qui porte le sens des concepts évoqués par l'auteur (Beneviste, 1964). Dans notre cas, cela nous permet d'étudier les sens que portent nos textes et les thèmes abordés par le biais de l'analyse des mots portés en leurs seins. À partir de cette théorie, on peut envisager plusieurs approches pour la modélisation sémantique de textes.

La première postule qu'un document est porteur d'un sens général et les mots qui le composent sont les signes qui définissent un sens ; en d'autres termes, chaque document porte un sens linguistique. Par ce postulat, on peut modéliser les mots des documents du corpus via des vecteurs de mots, puis appliquer des algorithmes non supervisés, comme le K-means ou DBSCAN afin de regrouper les textes les uns avec les autres en fonction de leurs similarités. Dans la grande majorité des cas, ces algorithmes nécessitent de déterminer le nombre de groupes en amont ou des hyperparamètres spécifiques, comme le rayon du point voisin pour DBSCAN.

La deuxième approche postule que les sujets étant des concepts, ils sont représentés par des associations de mots. De ce fait, un document peut avoir plusieurs sujets abordés et porter plusieurs significations. Dans ce cadre, on modélise les documents par des vecteurs de mots, puis on estime la probabilité de chaque document du corpus d'appartenir à un sujet. Cependant, la distribution de Dirichlet propose de contrôler la forme

de la distribution des thèmes via un paramètre alpha, car on postule en linguistique que, si plusieurs thèmes peuvent être évoqués, on retrouve fréquemment une dominance d'un sujet sur les autres dans les textes. Pour cette forme de modélisation, l'algorithme le plus utilisé est le LDA, qui intègre toutes ces contraintes pour proposer une modélisation de sujets via des mots en couple avec une modélisation des textes via leurs probabilités d'appartenir à un sujet.

Dans le cas de ce mémoire, nous avons choisi d'utiliser la modélisation de sujet pour mesurer la dimension sémantique de notre corpus. Plusieurs raisons ont motivé ce choix, relatives aux forces et faiblesses que ces approches comportent. Nous les détaillons ici. Dans le cas des documents juridiques, nous avons des documents relativement longs. Dans notre corpus, le plus court texte contient 987 mots, alors que le plus long en compte plus de 120 000. À partir de ce constat, nous proposons l'hypothèse qu'il est probable que plusieurs sujets soient abordés dans ces textes. De plus, la modélisation des thèmes via des vecteurs de mots rend la lisibilité de l'analyse plus simple que dans le cas de groupes qui demandent une analyse plus profonde pour comprendre la raison de leurs regroupements. En lien avec cela, notre corpus est composé de 5403 textes. Par conséquent, notre connaissance préalable du contenu de ce corpus est mince ; déterminer le nombre de regroupements est plus complexe, n'ayant aucune intuition des classes représentées par ces textes. En outre, nous avons pris en compte des considérations futures pour ce travail, qui inclurait de nouvelles données. Comme vu précédemment, une grande quantité de jugements est disponible pour l'analyse. Ainsi, il est important d'anticiper la taille de futurs jeux de données qui peuvent devenir rapidement imposants. Plusieurs variations du LDA proposent des méthodes incrémentales d'entraînement du LDA afin d'intégrer de futures données au modèle (Uray et al., 2007; Wenchao et al., 2022; Gerasimenko et al., 2023). En somme, la potentielle évolutivité du LDA, son interprétabilité et sa facilité d'utilisation avec nos données sont les raisons qui nous poussent à utiliser ce modèle pour extraire la sémantique de notre corpus. Finalement, nous devons souligner l'existence d'autres algorithmes que le LDA pour modéliser les sujets s'appuyant sur les nouvelles avancées de l'apprentissage profond, comme BERTopic ou récemment TopicGPT. Or, ces modèles semblent moins bien

performer sur des textes de grandes tailles que le LDA à cela, il faut ajouter une difficulté accrue à affiner les performances de ces algorithmes via le nettoyage des textes à cause de leurs sensibilités au retrait de variables.

Comme expliqué plus haut, le LDA propose d'analyser les sujets d'un corpus en analysant les vecteurs de mots qui constituent des sujets, puis propose d'attribuer une probabilité d'appartenir à un ou plusieurs sujets à un texte. Ci-dessous, nous proposons des diagrammes en assiette représentant le fonctionnement du LDA.

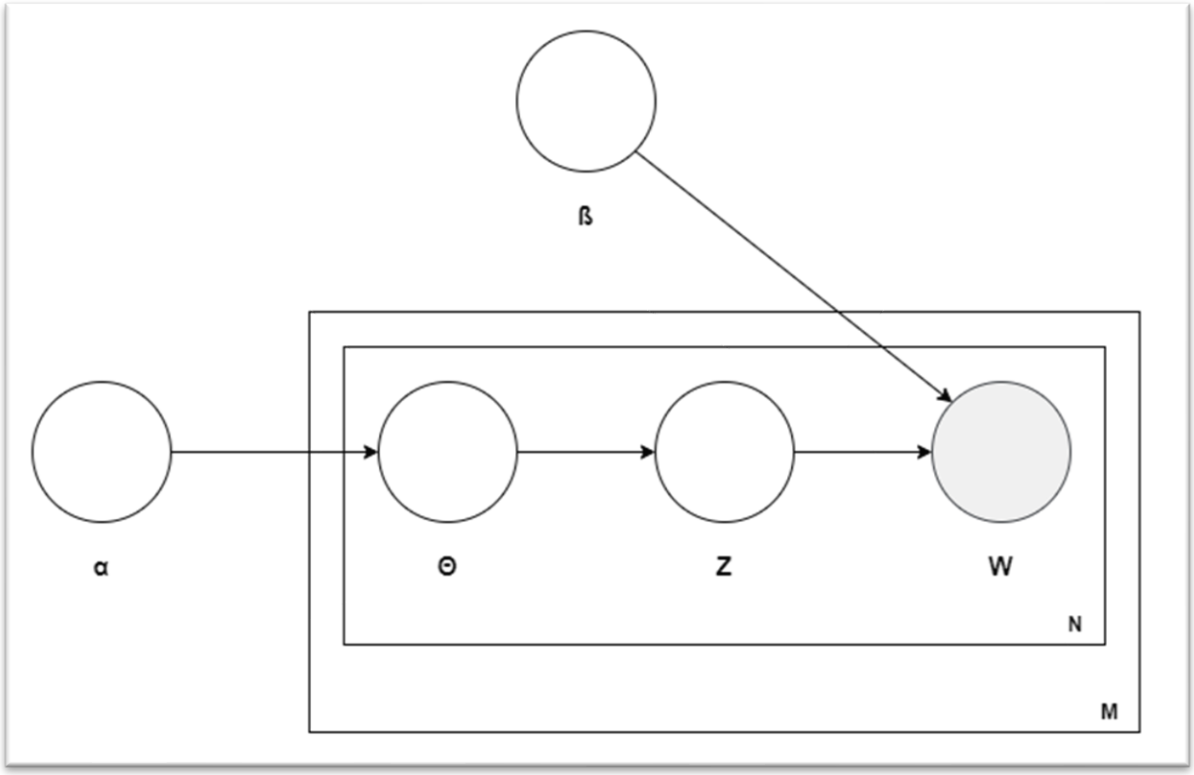


Figure 1 : représentation en assiette du modèle LDA (Saura, 2019)

Dans la figure 1, nous avons plusieurs variables illustrées par des lettres. M est l'ensemble des documents qui composent notre corpus, N est l'ensemble des mots qui représentent les documents du corpus M, et ainsi le document i est composé des mots N_iLes variables α et ß sont les hyperparamètres qui régissent la distribution des sujets dans les documents et la distribution des mots par sujet. Et finalement, Θ_i représente la distribution des sujets pour le document i.

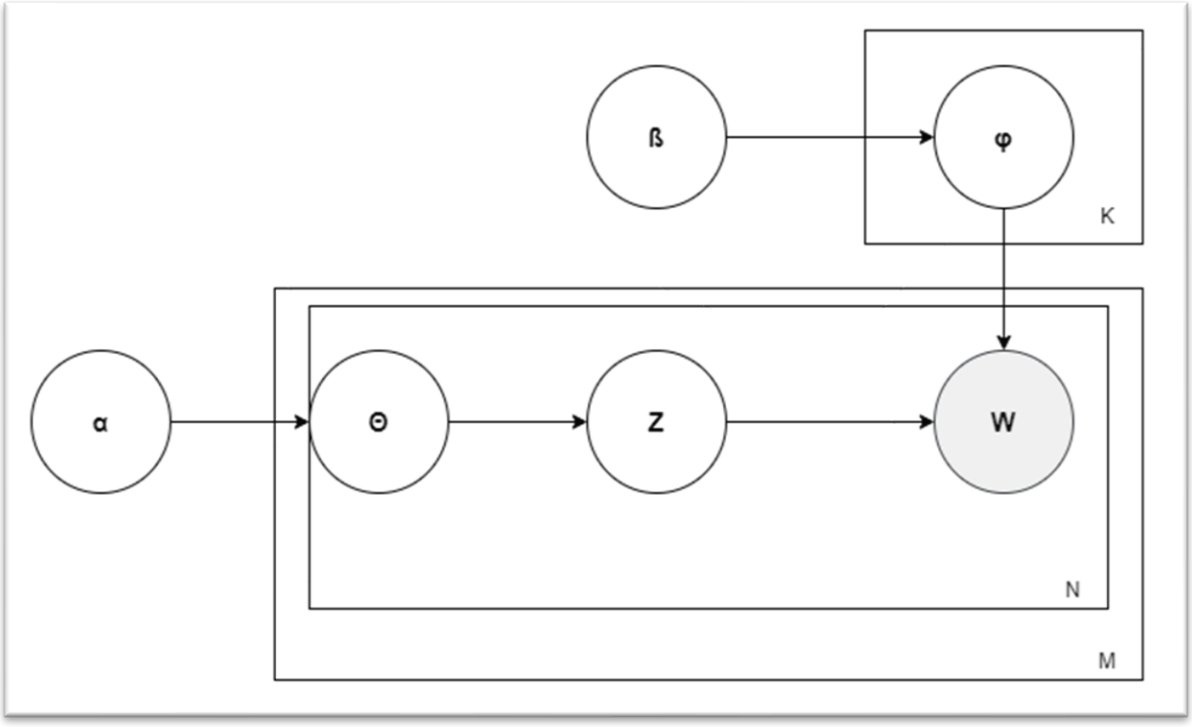


Figure 2 : représentation en assiette de la distribution de Dirichlet des sujets-mots pour le LDA (Saura, 2019)

Dans la deuxième figure, ci-dessus, nous proposons un diagramme qui représente le fonctionnement de la distribution des mots dans les sujets dans le cadre du LDA. On ajoute aux variables précédentes K, qui est l'ensemble défini par le nombre de sujets décidé en amont. Ainsi, φ_k est la distribution des mots pour le sujet k. Puis, z_{ij} est le sujet du j-ème mot du document i, et w_{ij} représente un mot spécifique. Sur ces schémas, la variable w est grise, car les mots sont les seules variables observées par le modèle, les autres variables sont inférées, elles sont latentes. L'utilisation de la loi de Dirichlet repose sur l'idée que la distribution des mots pour représenter un thème est biaisée. En effet, peu de mots ont une forte probabilité de définir un sujet (Blei et al., 2003). Via le paramètre α on peut donc fixer une distribution qui prend en compte la nature clairsemée de nos mots dans la représentation de nos sujets. On peut définir la fonction probabiliste de densité de la manière suivante, où ß représente une distribution des variables x :

$$f(X_1, \dots, X_k; \alpha_1, \dots, \alpha_k) = \frac{1}{ß(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_{i-1}}$$

Où $\sum_{i=1}^{k} x_i = 1$ et $x_i \in [0, 1]$ pour tout $i \in \{1, \dots, K\}$

De plus, dans la figure 2, K représente le nombre de sujets représentés par les vecteurs $\varphi_1, \ldots, \varphi_k$ de dimension V, qui incarne le nombre de mots du corpus. On peut voir φ et θ respectivement comme la matrice thèmes-mots et la matrice des documents-thèmes.

Si ce modèle est adapté pour la tâche d'extraction sémantique de nos textes, il est cependant affecté par des enjeux courants du NLP. En effet, la nature du langage induit des matrices clairsemées. La langue compte un grand nombre de variables impliquant qui ne se retrouvent pas toutes dans chacun des textes. De ce fait, on retrouve de nombreux mots absents de la représentation, laissant de nombreux points vides dans nos matrices. En lien avec cela on peut évoquer le fléau de la dimensionnalité. Dans le même ordre d'idées, le dictionnaire Larousse compte 63 800 mots et ce dernier n'inclut pas tous les mots issus du jargon professionnel, comme c'est le cas dans le droit. Chaque mot constitue une variable, ce qui peut amener à une dimension excessivement grande qui complique les calculs. Enfin, dans le langage, la majorité des mots ne sont pas porteurs de sens. On peut compter les mots grammaticaux, certains verbes ou encore des noms propres qui, en fonction du contexte, ne contribuent pas à comprendre le sens sémantique des textes. Pour pouvoir gérer ces facteurs, nous devons nettoyer nos textes afin de rendre le LDA plus performant. Plusieurs méthodes existent, nous détaillons ces étapes dans la partie qui discute de nos analyses.

III.3.b) La dimension juridique

Si l'aspect sémantique est une dimension importante pour regrouper les sujets, cette dernière ne prend pas forcément en compte la nature des arguments juridiques avancés par les parties pour défendre leurs cas. Effectivement, les plaidoiries peuvent évoquer plusieurs thèmes différents, mais s'appuyer sur des arguments similaires et, réciproquement, des sujets similaires peuvent utiliser des rhétoriques radicalement différentes. Par exemple, deux cas peuvent parler d'une fracture au genou, mais, sur le plan juridique, discuter de la répartition des indemnités ou de la reconnaissance d'une maladie professionnelle. Appareiller ces deux cas sur le simple fait sémantique n'est pas suffisant pour un professionnel qui cherche à élaborer avant tout un argumentaire cohérent auprès du tribunal. Notre réflexion s'est alors portée sur la nature de la *common law* et l'importance

des précédents dans la prise de décision des tribunaux. Comme expliqué précédemment, les tribunaux sont tenus les uns par les autres via les décisions passées dans le système de justice québécois. De ce fait, la grande majorité des jugements citent des décisions antérieures pour appuyer leurs décisions. On peut alors envisager les jugements comme un grand réseau interconnecté qui construit l'interprétation de la loi au Québec et au Canada. Outre cela, les articles de lois sont les bases des procédures juridiques pour les tribunaux. Tout comme les jurisprudences, les juges citent dans leurs décisions les lois et leurs articles qui étayent la nature de leurs décisions. Ainsi, on peut aussi intégrer ces derniers dans le réseau de textes qui composent la pyramide juridique canadienne. Ce qui rend notre tâche plus aisée est la standardisation globale des formes de citations qui sont compilées dans les textes. L'extraction des références est donc envisageable facilement et nous expliquerons cette dernière plus en détail dans la partie qui parle de nos résultats. Cette modélisation nous permet de créer une représentation de graphe orienté où nos nœuds sont les textes de notre corpus, les lois et les jurisprudences citées. La modélisation des liens entre les nœuds a été un enjeu pour ce travail. Pour ce faire, nous avons dû réfléchir à deux éléments principaux.

Le premier est la question de la modélisation des lois. Par exemple, la loi la plus fréquemment rencontrée est celle sur les accidents et les maladies professionnelles. Si les juges font référence aux articles de la loi, dans certains cas, ils font référence à l'ensemble du texte. Au début de notre travail, nous avons créé des liens distincts entre les textes, les chapitres de lois et les lois. Le problème que cette modélisation entraîne réside dans le calcul des mesures de distance et de similarité. Étant donné que les objets sont différents, cela veut dire qu'un texte qui cite un article de la loi n'aurait pas de points communs avec un texte qui cite la loi au complet. Voulant conserver un niveau de granularité fin, nous avons décidé de nous inspirer du travail précédent de Shounak et al. (Shounak et al., 2021). Dans cet article, les auteurs utilisent une représentation hiérarchique des lois. Ainsi, suivant cette

approche, nous avons décidé de modéliser les textes de loi de manière hiérarchique, comme le montre la figure 3 ci-dessous que nous avons produite.

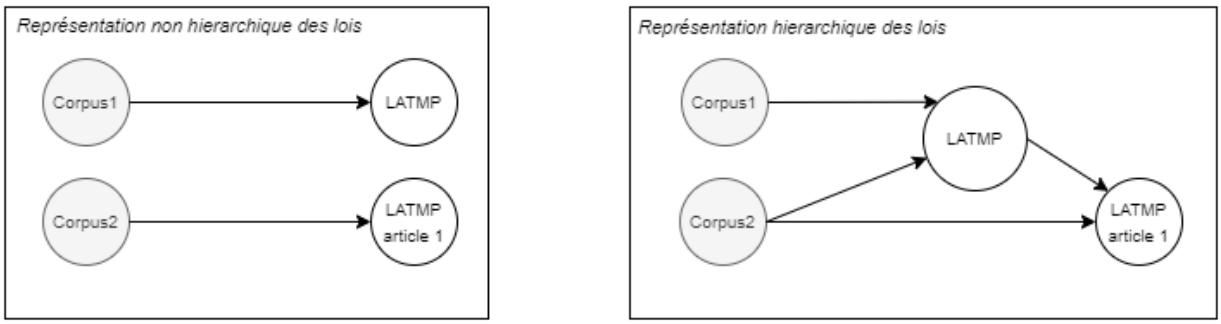


Figure 3 : représentation des lois dans le réseau de graphe

Dans cette figure, nous voyons deux exemples de représentation. Les nœuds « corpus » représentent les textes de notre corpus, tandis que les nœuds « LATMP » et « LATMP, article 1 » représentent la loi sur les accidents et les maladies professionnelles ainsi que son article 1. Au début, nous avions décidé d'opter pour la représentation non hiérarchique. Toutefois, nous avons jugé plus approprié de modifier nos données pour respecter la logique hiérarchique. En effet, les articles de lois sont connectés au chapitre de lois desquels ils sont issus. De plus, cela implique que les textes qui citent un article, par définition, citent la loi dont le chapitre est issu. Cette représentation a un impact sur les calculs de similarités et les distances entre les textes ainsi que sur la forme générale du graphe, notamment sur le nombre de communautés. Nous détaillerons ces éléments plus en détail dans la discussion de nos analyses.

Notre deuxième problématique résidait sur l'orientation du graph. Si nous avons considéré représenter notre graphe sans orientation des liens, il nous a semblé plus logique de respecter l'orientation des références. Plus précisément, cela veut dire que les textes du corpus citent des références, cela implique une orientation des liens en ce sens. Cette logique est importante, car nous ne pouvons pas estimer que les références jurisprudentielles, par définition antérieures aux jugements du corpus, s'appuient sur ces derniers. Donc, il serait faux de considérer que les citations ont une relation bilatérale, surtout si nous considérons élargir notre graphe sur plusieurs années dans le futur. En

addition, comme le montre la figure 3, nous respectons la logique hiérarchique des lois pour modéliser l'orientation des liens entre les chapitres de lois et leurs articles.

Une fois la modélisation effectuée, nous voulons obtenir des métriques pour mesurer la proximité de la logique entre deux textes. Deux mesures sont alors envisagées dans notre cas, la similarité de Jaccard et la distance BFS dans le graphe.

La distance de Jaccard est une mesure qui se base sur la similarité de Jaccard proposée par le botaniste Paul Jaccard, amène une méthode statistique pour comparer la diversité entre deux échantillons. Pour cela, on propose deux échantillons A et B respectivement décrits comme suit :

$$A = (a_1, a_2, \dots, a_n)$$

$$B = (b_1, b_2, \dots, b_n)$$

Où $a_i \in \{0,1\}$ pour tout $i \in \{1, \dots, n\}$

Ainsi, on peut définir la similarité de Jaccard comme le rapport entre la valeur absolue l'intersection des deux ensembles et la valeur absolue l'union de ces derniers, soit :

$$J(A,B) = \frac{|A \cap B|}{|A \cap B|}$$

Dans notre cas, les ensembles A et B représentent les différentes citations entre deux textes. La distance de Jaccard est alors la dissimilarité entre les deux ensembles qui consiste en la soustraction de 1 à la distance de Jaccard, soit :

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \,\Delta\, B|}{|A \cup B|}$$

Dans notre situation, nous procédons à une itération du calcul des distances et similarités de Jaccard entre les paires de nœuds de notre graphe. Par ce moyen, nous obtenons une matrice des distances entre chaque nœud nous permettant d'obtenir un indice de similarité et de distance entre les textes.

Outre la similarité des textes dans notre corpus, nous voulons aussi étudier la proximité entre nos textes dans le graphe des références. En effet, il est possible que des textes aient une proximité avec d'autres sans pour autant avoir un score de similarité élevé. Il est possible que des connexions de second plan puissent exister, ce qui revêt un intérêt dans notre cas. Afin de tester cette hypothèse, nous décidons d'utiliser le « *breadth-first search algorithm* » (BFS) afin de calculer la distance entre chaque vertices dans le graphe. Cette distance est en réalité le nombre d'arêtes minimum qui relie deux points du graphe. Pour cela, le BFS utilise un système de queue qui vérifie si le nœud a déjà été exploré au préalable avant de procéder au calcul de distances (Lee, 1961). Cette optimisation nous permet un gain d'efficacité essentiel dans le cas de grands ensembles comme le nôtre. Cependant, BFS fonctionne dans le cadre d'un graphe fini, c'est-à-dire que tous les nœuds sont connectés au graphe, ce qui, dans notre cas n'est pas forcément le cas. Nous avons donc utilisé des méthodes pour le calcul que nous décrirons dans nos analyses.

Ce cadre conceptuel nous permet d'étudier le droit du travail sous un angle sémantique et juridique. Nous tâchons de cumuler les deux dans nos analyses pour proposer une étude complète détaillant notre preuve de concept. Puis, nous essayons de cumuler ces approches dans nos résultats, pour voir les liens potentiels que ces dernières peuvent avoir.

III.4) Extraction des données de citations et constitution du graphe

En complément des analyses quantitatives de la sémantique, nous avons procédé à une étude des éléments qui constituent la base de notre réseau de graphes, les citations de jugements et de lois dans les textes de notre corpus. De manière à explorer le réseau, nous avons constitué le graphe en faisant une extraction des citations via un système d'expressions régulières (regex). Les regex sont des séquences de caractères qui forment un modèle à apparier dans le texte. L'avantage que comportent les textes de nature juridique repose sur la standardisation des citations que portent les processus d'édition des jugements. De cette manière, on peut retrouver l'ensemble des normes de citations dans des publications dédiées à cela (Luelles et Ringuette, 2019). Pour les lois, nous avons

remplacé tous les noms de loi cités par leurs numéros, comme : loi sur les assurances : RLRQ, c. A-32. Une fois cela fait, nous avons cherché un moyen pour extraire les articles reliés à ces textes de loi. Nous avons exploré plusieurs options pour effectuer cette tâche. Une des options estimées a été d'utiliser de grands modèles de langage (LLM) pour extraire ces informations. Après quelques essais de ChatGPT et Gemini, nous avons constaté que ces modèles omettent fréquemment des citations dans les résultats qu'ils nous offrent. Une autre piste que nous avons explorée a été l'entrainement d'un modèle supervisé pour extraire ces informations des textes. Or, cette méthode requiert une grande quantité de données textuelles annotées dans le but d'entrainer un modèle, ce dont nous ne disposons pas. C'est pour ces raisons que nous avons opté pour la méthode suivante. Nous avons tokenisé les phrases du corpus pour extraire les articles avec les lois qui y sont attachées. Puis nous avons filtré les phrases pour garder les mots « articles » et les chiffres après ces derniers. Finalement, nous avons manuellement vérifié et corrigé l'extraction pour nous assurer que ces dernières sont cohérentes. À partir de cela, nous obtenons un ensemble de données avec une colonne contenant le numéro de cas du texte de référence, sous la forme QCTAT123, et une colonne avec la référence de la loi extraite du texte et les articles qui y sont liés. Dans un premier temps, nous avons conservé cette structure de données qui respecte l'architecture présentée plus haut dans le premier diagramme de la figure 3, mais, comme expliqués dans notre sous-chapitre sur la dimension juridique, pour respecter la cohérence juridique de la hiérarchie des sources pour les calculs de similarité, nous avons opté pour la deuxième architecture présentée dans la figure 3. Ainsi, nous avons créé des liens entre la loi et leurs articles, puis entre les textes de référence et les lois de référence. Finalement, nous avons retiré les doublons de la liste des liens, car nous postulons que la quantité de références à la même loi ne témoigne pas automatiquement de l'importance de la loi pour l'argumentaire. En effet, nos cas étant relatifs à la santé et à la sécurité au travail, la loi sur les accidents du travail et la maladie professionnelle est souvent citée, car c'est le texte de référence pour ces situations. Pourtant, dans un même paragraphe, le juge peut la convoquer plusieurs fois sans parler spécifiquement d'un article pour appuyer son propos.

De ce fait, pour ne pas biaiser nos analyses ultérieures, nous décidons de ne considérer qu'une seule citation de chaque texte de loi différent pour chaque texte de référence.

En ce qui a trait aux citations de jugements, la tâche est plus aisée. Nous avons cherché les abréviations des tribunaux qui sont citées dans les numéros de dossier des jugements. Dans le même ordre d'idées que pour les lois, nous nous sommes appuyés sur le guide des références juridiques (Luelles et Ringuette, 2019) et nous avons cherché toutes les abréviations compilées dans cet ouvrage. Par ce biais, nous avons pu concentrer notre extraction sur les abréviations de jugements qui se trouvent dans notre corpus pour notre script d'extraction. Sur la base de cette liste, nous avons étudié les différents modèles des numéros de dossier de chaque abréviation pour élaborer des regex qui permettent d'appareiller les numéros de cas présents dans le corpus. Par exemple, la décision Précibois inc. et Brisebois est mentionnée sous la référence C.L.P. 111268-62A-9902. Cette norme de citation est la même pour tous les jugements de la Commission des lésions professionnelles, seuls les numéros changent. Consécutivement à cela, nous avons procédé à une vérification manuelle de l'extraction des citations de jurisprudences ainsi qu'à la correction de certains cas. En accord avec la décision de supprimer les doublons dans notre table de liens pour les lois, nous avons utilisé la même logique pour supprimer les doublons ici. Ensuite, nous avons concaténé les deux tableaux de données pour avoir une table de données des liens du graphe. Finalement, nous avons construit une table des nœuds uniques du graphe. Elle compile et spécifie les références des lois, des textes du corpus et les références des jugements.

L'ensemble des méthodes utilisées fonctionne pour le cadre de notre mémoire. Cela étant dit, nous soulignons que des améliorations sont possibles pour automatiser ce processus. Nous évoquons ces éléments dans notre chapitre sur les limites et les travaux futurs de ce mémoire. Une fois les données extraites, nous avons pu faire un graphe sur Gephi pour visualiser et regarder notre graphe parallèlement à nos analyses sur Python avec le paquet NetworkX adapté aux graphes.

IV) Analyses

Dans ce chapitre, nous expliquons en détail la nature de nos analyses, les problèmes face auxquels nous avons été confrontés et les solutions que nous avons apportées en réponse à ces enjeux. En accord avec notre démarche de recherche, nous décrirons l'exploration de données que nous avons faite, puis nous regardons les étapes d'optimisation et de sélection de variables pour notre LDA. Enfin, nous décrivons la construction de notre graphe et sa structure.

IV.1) Exploration des données

Pour mieux comprendre notre corpus, nous proposons une analyse exploratoire des données de ce dernier. En accord avec les tâches que nous souhaitons effectuer, nous regardons des données sémantiques et de structure de graphe. Ces données sont principalement descriptives, mais nous permettent de révéler des intuitions que nos analyses vérifieront. Outre cela, cela révèle les potentiels extremums à étudier dans le corpus ainsi que les enjeux à prendre en compte dans la mise en place de nos modèles.

IV.1.a) Exploration quantitative des données sémantiques

Comme précisé dans le chapitre sur la collecte de données, nous avons 5 403 textes au format .txt. Afin d'être rigoureux, nous avons regardé à l'aide d'un script Python s'il n'y avait pas de doublons dans nos données. Via cela, nous avons trouvé un texte identique à un autre que nous avons retiré, ce qui réduit notre corpus à un compte de 4 502 textes. Dans le but de mieux appréhender le contenu du corpus, nous procédons à une exploration de ce dernier via des analyses statistiques. Pour ce faire, nous avons regardé le nombre de caractères, le nombre de mots et le nombre de phrases présent dans les textes. Si, pour les caractères, Python propose des fonctionnalités natives pour les compter, les mots et les phrases requièrent une opération préalable, la tokenisation. La tokenisation consiste à diviser un texte en unités plus petites, appelées tokens, qui sont généralement des mots ou des phrases. Pour cela, nous utilisons la bibliothèque spacy, qui propose un tokenizer pour les mots et les phrases. Nous segmentons nos textes en mots et en phrases afin de réaliser

une analyse statistique de notre corpus. Les résultats de cette analyse sont synthétisés dans le tableau ci-dessous.

	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Écart-type
Caractères	3310	17194.5	25672.0	31215.414584	38568.5	517203	22570.102728
Mots	987	4380.0	6409.0	7770.632982	9554.0	126056	5468.132445
Phrases	43	194.0	288.0	339.094762	415.0	4745	229.858906

Tableau 1: Statistiques descriptives de l'exploration du corpus

La première remarque que nous pouvons formuler sur cette analyse concerne la taille des textes. En effet, nous avions inféré que les textes juridiques étaient relativement longs, car comportant des argumentations de la part des juges pour soutenir leurs décisions. On voit que notre corpus est composé de textes compris entre 987 et 126 056 mots, ce qui confirme cette intuition. Pour appuyer notre analyse, nous avons produit des histogrammes de la distribution des textes en fonction de leurs nombres de caractères, de mots et de phrases, disponibles dans les annexes 1 à 3. Ces annexes indiquent graphiquement une distribution asymétrique vers la droite de la taille de notre corpus. En d'autres termes, une grande majorité des textes ont moins de 10 000 mots, plus précisément 75 % de notre corpus comporte moins de 9554 mots. Par ailleurs, on peut aussi constater un groupe de jugements moins long que les autres ; on parle ici de moins de 1600 mots. Une partie de ces textes sont des révisions administratives par le tribunal qui confirme ou infirme une décision sans réellement développer d'argumentaires. La nature succincte de certaines décisions peut aussi s'expliquer par le renvoi des cas à d'autres instances par les juges. En effet, dans certains dossiers, les demandes ne relèvent pas de la compétence qu'attribue la loi au TAT. Dans cette situation, le juge justifie brièvement la nature du renvoi sans approfondir le cas. On peut aussi voir quelques jugements très longs, comme le montre notre maximum de 126 056 mots. Ces textes sont cependant rares, comme le montrent nos histogrammes. En général, il s'agit de cas complexes qui traitent de plusieurs sujets ou dossiers d'accidents dans un jugement. Le juge doit à ce moment passer en revue deux ou trois cas reliés les uns aux autres et élaborer un raisonnement pour chaque cas avant de présenter sa décision, ce qui allonge la taille du jugement. De cette première analyse, on

constate que la dimensionnalité de notre échantillon de données est grande. Puisque chaque mot unique est regroupé en tant que variable pour le modèle de LDA. Malgré le regroupement des tokens pour former les variables, on peut raisonnablement inférer que, sans manipulation des données, la quantité de variables reste trop vaste pour notre modèle. Cette grande dimensionnalité est une problématique récurrente dans les travaux de NLP et elle entraine une perte de signal statistique qui affecte les performances des modèles utilisés.

Afin d'approfondir la compréhension de notre corpus, nous avons procédé à une analyse des quarante mots les plus fréquents dans notre corpus, comme présenté dans les annexes 5 et 6. Nous proposons deux analyses, une sans aucune manipulation des données et une avec un nettoyage. La nécessité du nettoyage s'est avérée essentielle, car l'annexe 5 nous montre que les termes les plus fréquents sont des mots grammaticaux et des éléments de ponctuation. Cela souligne un autre élément caractéristique des données textuelles, l'asymétrie d'information. Si cette caractéristique n'est jamais nommée en tant que telle dans les publications qui traitent du NLP, elle est induite par la nature de la distribution des mots dans un texte. Dans les langues occidentales, la distribution suit fréquemment le postulat de George K. Zipf's (Moreno et al., 2019) qui nous indique que la distribution d'un mot dans un corpus est inversement proportionnelle à son rang soit :

$$Fréquence \approx \frac{1}{rang} \times constante$$

Si des conditions ont été ajoutées avec le temps sur cette loi, on constate que les textes ont tendance à suivre cette loi. Outre la distribution des mots dans le texte ou le corpus, on peut remarquer que, généralement, les mots de nature grammaticale sont les plus fréquents (le, la, les, ce, etc.). Or, ces éléments du langage ne sont pas porteurs de sens en dehors du contexte de la phrase dans laquelle nous pouvons les trouver. Par exemple, le mot « vélo » a un sens clair pour tous les locuteurs qui le connaissent. Il porte donc une information forte en comparaison du mot « cela », qui lui n'évoque rien de distinct, en fonction du contexte. Cette asymétrie de l'information doit nous amener à porter une attention particulière à la manière dont nous devons diminuer la dimensionnalité. Nous devons nous assurer de ne

pas retirer de variables fortement informatives pour nos modèles, même si elles sont moins fréquentes dans nos textes. L'annexe 6 nous montre l'effet des nettoyages sur l'informativité dégagée des mots les plus fréquents après un bref nettoyage de nos tokens. La réduction de taille a été effectuée en supprimant la ponctuation et les éléments de mise en page, tels que/n (pour un retour à la ligne) et/s (pour des espaces additionnels). Nous avons également utilisé le paquet NLTK, qui contient une liste de mots grammaticaux fréquemment enlevés du français, appelés *stop-words* (annexe 4). Si on conserve quelques mots grammaticaux, ces manipulations permettent déjà d'extraire plus d'information de notre corpus. On voit trois catégories de lexique qui se dégage :

- Les mots relatifs au travail : travailleur/se, employeur, emploi, etc.
- Les mots relatifs à la santé : docteur, lésion, médicale, etc.
- Les mots relatifs au jargon du tribunal : dossier, administratif, événement

On peut noter un mot comme gauche qui est relatif au côté de la lésion ou de la blessure, mais aussi droit qui est ambigu dans notre cas, car pouvant faire référence au droit comme juridique ou au côté droit. Nous avons proposé aussi une analyse des bigrammes et trigrammes qui sont la suite de deux ou trois mots les plus fréquents dans les textes. Nous proposons une modélisation des plus fréquents dans les annexes 7, 8, 9 et 10. Via cette exploration, ne nous remarquons que des termes comme l5 s1 sont présent fréquemment ensemble. Cet élément nous parait suspect et nous avons regardé de plus près les textes qui comportent ces bigrammes. Par cette vérification, nous avons constaté qu'il s'agissait des noms médicaux pour les vertèbres et les C1, C2, ..., C7 font référence aux cervicales. Ainsi, nous prenons en compte ces facteurs pour normaliser ces éléments, afin d'éviter la suppression de ces mots importants au moment de nos nettoyages. Concernant l'analyse des trigrammes, nous ne dégageons pas d'information majeure qui nécessite une attention particulière.

Nos analyses nous indiquent aussi que, si la dimensionnalité est élevée, les matrices sont clairsemées. En effet, plusieurs termes manqueront dans la plupart des textes de l'ensemble, ou ne seront présents que dans un seul d'entre eux. Cette caractéristique peut

complexifier l'interprétation et la visualisation des données ainsi que le fonctionnement de nos modèles.

En résumé, cette exploration nous a montré plusieurs éléments. Tout d'abord, nous avons énoncé trois enjeux du NLP auxquels nous sommes confrontés pour notre recherche :

- La grande dimensionnalité
- Les matrices clairsemées
- L'asymétrie de l'information

Ces enjeux sont des indications que plusieurs manipulations des données sont nécessaires afin d'optimiser le fonctionnement de nos modèles et l'interprétabilité des résultats. En effet, le LDA formant des vecteurs de mots, le bruit statistique risque de nuire à la lisibilité des sujets formés par l'algorithme ainsi que les performances et la robustesse de la modélisation de sujets. La réduction de dimension doit prendre en considération les caractéristiques particulières des données textuelles. Nous détaillons ces nettoyages dans notre partie sur les analyses du LDA.

IV.1.b) Explorations des données de graphe

Notre graphe est constitué de 856 lois, règlements et articles de lois et de règlements ainsi que 11 187 références jurisprudentielles qui sont liées à notre corpus de 5 403 textes de notre corpus. Cette différence notable entre les références jurisprudentielles et les lois s'explique par le fait que les juges citent souvent le ou les articles de lois au début du jugement ou de la décision, mais le nombre et la qualité des jurisprudences invoqués pour justifier l'interprétation de la loi par le tribunal sont un facteur important pour la robustesse du jugement face à de potentiels appels administratifs. En effet, comme expliqué dans notre chapitre sur le cas du droit du travail québécois, la logique de la cour est primordiale pour assurer un jugement juste et équitable. Ainsi, convoquer des décisions antérieures est un moyen fiable de s'assurer de cela, d'où la nature déséquilibrée entre les deux types de nœuds de notre graphe.

En complément, nous avons regardé des statistiques descriptives sur toutes les citations, les citations de jurisprudences et les citations de lois dans le corpus. Nous analysons la distribution des citations dans ce dernier présenté dans le tableau 2 et les annexes 11 à 13.

	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Écart-type
citations	0	5	8	12.0407	15	105	11.7563
jurisprudences	0	1	4	7.6651	10	91	10.1063
lois	0	2	4	4.3755	5	31	3.1823

Tableau 2 : Statistiques descriptives des citations dans le corpus

Notre analyse révèle que les textes contiennent entre 0 et 105 citations. En regardant de plus près nos données, on trouve 73 textes sans citations. Nous avons lu ces textes et, de cette analyse nous remarquons deux catégories de jugements, des corrections du tribunal et des avis médicaux. Les corrections du tribunal sont des jugements qui corrigent des éléments de formes de décisions antérieures ou qui obligent une partie à exécuter un ordre du tribunal. Ces derniers se concentrent plus sur la forme que sur le fond d'où la non-citation d'articles de lois ou de jurisprudences en leurs seins. Les avis médicaux sont des jugements qui se concentrent sur des contre-expertises qui vérifient la validité d'une évaluation médicale proposée par un premier médecin. De ce fait, ces jurisprudences ne s'appuient pas sur des avis juridiques, mais sur de la documentation médicale et des publications en médecine. Outre cela, on remarque que la moitié de notre corpus comporte entre 0 et 5 citations et les trois quarts de ce dernier en comportent moins de 15. En regardant les statistiques descriptives et notre histogramme, nous constatons que la distribution est décalée à gauche, ce qui est logique, car peu de textes ont tendance à citer un grand nombre de références. Cependant, si on se penche sur les textes compris entre le troisième quartile et le maximum, on constate que certains jugements ont une liste extensive de citations. En regardant quelques-uns de ces textes, notamment celui avec le plus de citations, on remarque qu'il s'agit de textes longs. Par exemple QCTAT2899 qui

contient 105 citations est un texte de 36 pages qui contient des débats sur certains arguments. On peut aussi souligner des paragraphes avec beaucoup de citations, ce qui peut être lié à des arguments litigieux amenant la cour a appuyé sa décision avec un grand nombre de références. Nous amenons ici l'hypothèse que les textes avec un grand nombre de citations sont des textes importants, car contenant des discussions nécessitant plus de citations à l'appui des arguments du juge. Si notre hypothèse est vérifiée, cela implique que ces textes ont une importance pour le droit du travail, car ils comportent des débats nécessitant l'appui de nombreuses sources. Nous tâchons de vérifier cela dans nos analyses et nos résultats. En complément à cela, nous avons aussi la distribution des textes en fonction du nombre de citations de jurisprudences et de lois séparément. On constate un déséquilibre entre les deux classes en regardant le troisième quartile dans les deux situations. Effectivement, les trois quarts des textes citent moins de 10 jurisprudences et moins de 5 lois. On voit aussi que le maximum de citations de précédents judiciaires est de 91 pour 31 dans le cadre des lois. En regardant graphiquement les distributions de l'annexe 9 et 10, on remarque que peu de texte du corpus ne cite aucune loi. La moyenne de citations des lois est aussi largement inférieure à la moyenne de citation des précédents. On peut expliquer cela du fait que la loi est l'élément de base pour l'introduction d'un cas au tribunal. En d'autres termes, le TAT utilise la loi comme socle de l'argumentation et généralement quelques articles de lois suffisent pour la conduite de la procédure, ce qui explique que moins de la moitié de notre corpus s'appuie sur moins de 4 lois. En revanche, pour comprendre comment appliquer la loi et l'interprété, les décisions antérieures sont essentielles comme expliqué dans notre chapitre sur le droit administratif du travail québécois. Ainsi, il semble naturel et logique qu'un plus grand nombre de jurisprudences soit convoqué pour appuyer un argument ou un autre. Dans ce sens, l'annexe 9 nous montre que la distribution des textes en fonction du nombre de jurisprudences qu'ils citent suit la tendance présentée pour l'ensemble des citations en annexe 8. Même si le nombre global de ces citations est plus grand que celui des lois, on voit tout de même qu'une grande majorité des textes ne citent qu'entre 0 et 5 jurisprudences. Dans certaines situations, la cour a probablement besoin de peu de références, car les cas présentés sont solidifiés par

la loi et les tribunaux ne nécessitant pas un grand nombre d'arguments à l'appui de la décision. Tout cela nous amène à évoquer un autre point, qui est la qualité des références. En effet, les cas qui citent peut peuvent s'appuyer sur des références communément reconnues comme de meilleures qualités se suffisant à elle-même. Cet élément implique que certaines jurisprudences sont des éléments de qualité qui reviennent dans des jugements traitants de cas similaires, donc qui ont une présence forte dans le corpus. Cette deuxième hypothèse complémentaire à la deuxième implique que les jugements avec peu de citations s'appuient sur des citations jurisprudentielles utilisées avec récurrences, ce qui témoigne de leurs qualités.

Notre analyse exploratoire des données de graphe nous révèle que le réseau des citations, bien que vaste, comporte des éléments clés à étudier en profondeur. Nous avons posé deux hypothèses ici :

- La grande quantité de citations que comporte un texte révèle l'importance de ce dernier pour le droit du travail québécois
- La faible quantité de citations que comporte un texte nous indique de la qualité de ces dernières.

La vérification de ces hypothèses est abordée dans nos résultats qui se concentrent sur les indices de similarité et les distances que notre corpus détient. Ces indices de similarités sont à analyser entre les textes du corpus, mais aussi entre les différentes références qu'il contient. La vérification de ces hypothèses nous permet de mettre de l'avant des informations importantes en outre de l'appareillement de cas si elles s'avèrent valides.

IV.2) Sélection du modèle de LDA

IV.2.a) Mesures de cohérence

Afin de mesurer la dimension sémantique de notre corpus, nous utilisons l'allocation latente de Dirichlet. Comme expliqué dans notre partie sur la sémantique dans notre chapitre sur la formalisation de la tâche, cet algorithme détermine une modélisation des sujets couverts par un texte en fonction du nombre de sujets k prédéfini par l'utilisateur. La question du choix du paramètre k se pose alors. De plus, nous avons évoqué trois grandes

problématiques qui affectent les analyses en NLP, soit le fléau de la dimensionnalité, les matrices clairsemées et finalement l'asymétrie de l'information. Ces enjeux affectent le LDA, car c'est un modèle probabiliste qui s'appuie sur les probabilités de cooccurrence des mots dans des documents. De ce fait, nous devons regarder plusieurs modèles et voir comment les nettoyages de nos données impactent les résultats de nos modèles. Pour cela nous décidons d'itérer des modèles de LDA pour $k \in \{1,2,3, \dots ,50\}$ et de regarder plusieurs scores de performances.

Le premier que l'on analyse est le score de cohérence Cv. Ce score propose de mesurer la cohérence des sujets modélisés en se basant sur la « *Normalized Pointwise Mutual Information* » (NPMI) (Syed et Spruit, 2017). La NPMI est la version normalisée de la « Pointwise Mutual Information » ou PMI, qui est une mesure qui permet de calculer la probabilité de deux mots d'apparaitre ensemble soit :

$$PMI(w_i, w_j) = log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

Où $P(w_i, w_j)$ est la probabilité d'apparition conjointe des mots w_i et w_j et $P(w_i)$ et $P(w_j)$ sont les probabilités indépendantes d'apparition pour les mots w_i et w_j.

La NPMI permet d'obtenir un résultat compris entre 1 et -1 qui nous permet de regarder les différentes associations entre les mots sans prendre en compte la fréquence d'apparition de ces derniers, soit :

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log P(w_i, w_j)}$$

Dans le cadre du LDA, le score de cohérence Cv utilise le NPMI combiné à la distance cosine avec la fenêtre glissante booléenne pour évaluer la proximité entre les vecteurs de mots (Röder et al., 2015). Avec cette méthode, on veut évaluer la cohérence entre les mots qui compose les sujets en fonction de leurs distances les uns par rapport aux autres. La technique de fenêtre glissante, ou « sliding boolean window », permet de segmenter les textes afin de vérifier la présence de certains mots dans cette dernière. Cette mesure de

cohérence est comprise entre -1 et 1, mais il est rare qu'elle soit négative. Afin de sélectionner un modèle, on cherche à maximiser le score Cv afin d'obtenir une meilleure distinction des sujets composés. Cependant, les chercheurs pointent plusieurs limites à ce score de cohérence. La première concerne la sensibilité au nombre de sujets k. En effet, Cv à tendance à favoriser les modèles avec un plus petit k, car ayant plus de probabilité de cooccurrence de mots. De plus, ce score a tendance à discriminer les sujets plus spécialisés au profit des sujets généraux avec des mots plus fréquemment utilisés à travers les sujets. Ainsi, il est important de souligner que la nature des données peut influencer le score de cohérence. Pour ces raisons, nous proposons d'y adjoindre une autre métrique de cohérence afin de nuancer les résultats que le Cv nous propose.

La deuxième métrique que nous prenons en compte est le score de cohérence Umass. Cette métrique s'appuie sur la cooccurrence des mots dans les documents du corpus de base afin de déterminer si les sujets composés sont cohérents et interprétables. Dans leurs articles de 2011, David Mimno et al. (David Mimno et al., 2011), propose la formule suivante pour définir Umass :

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{D(v_l^{(t)})}$$

Où t représente un sujet, $V^{(t)} = \left\{v_1^{(t)}, v_2^{(t)}, \dots, v_M^{(t)}\right\}$ est un ensemble M de mots qui compose un sujet. Si on regarde le contenu de la fraction, $D(v_m^{(t)})$ représente la fréquence du mot m dans le corpus de documents. Au numérateur on retrouve $D\left(v_m^{(t)}, v_l^{(t)}\right)$, qui représente la cooccurrence du mot « m » et du mot « l » dans le corpus de référence. Afin d'éviter d'obtenir un logarithme de 0, les auteurs ajoutent la constante 1. Ainsi, plus les mots cooccurrents dans le corpus, plus la valeur de Umass est proche de 0 et, réciproquement, moins ils cooccurrent, plus la valeur du score est négative. Donc, plus le score est faible et plus les sujets sont définis par des combinaisons de mots uniques qui les distinguent. Cette déduction est confirmée par l'article Farea et al (Farea et al., 2024), qui nous indique que, plus le score de cohérence est élevé, plus les sujets sont cohérents. L'avantage principal de

ce score réside dans la vitesse de son calcul en comparaison du Cv. En revanche, ce score prend en compte la fréquence de cooccurrence dans les documents. Contrairement au score Cv qui s'appuie sur une fenêtre, ce score peut passer à côté de liens contextuels entre les mots du corpus. Ainsi, on peut constater que la performance de ces scores est très relative à la nature des données présentées. C'est pour l'ensemble de ces raisons que nous les interprétons conjointement pour éviter le plus possible de générer des biais de sélection dans notre analyse du paramètre k.

Si les scores de cohérences Cv et CUmass nous permettent de nous aider à la sélection de notre modèle, il est important de souligner que maximiser ces scores n'implique pas forcément que les sujets formés par le LDA sont interprétables et lisibles. Ayant manifesté l'importance d'avoir des analyses lisibles pour les professionnels, nous devons introduire un élément de lisibilité pour le choix de notre modèle. L'utilisation des deux scores a pour objectif d'affiner l'échantillon k à vérifier pour l'entrainement du LDA. Mais, la sélection finale du modèle s'appuie sur la lecture des sujets composés et l'interprétation que nous pouvons en faire. Cependant, le nettoyage de nos données est une partie essentielle dans le prétraitement du corpus. Or, ces scores ont une sensibilité à la taille et à la nature du vocabulaire. Nous utilisons cette propriété pour voir si nos nettoyages ont un effet sur le modèle ainsi que pour comparer les différents nettoyages de données les uns aux autres. En résumé, les éléments quantitatifs et qualitatifs doivent être pris en considération lors du choix de notre modèle. Cela permet d'assurer que nos sujets ainsi que la robustesse du modèle correspondent à nos besoins.

La dernière métrique que nous utilisons pour évaluer notre LDA est la perplexité. Cette mesure est fondée sur la fonction de vraisemblance « likelihood » des documents pour le modèle de LDA à k sujets. En d'autres termes, cette mesure nous permet de voir à quel point ce modèle décrit précisément le corpus de document sur lequel il est utilisé (Blei et al., 2003). La bibliothèque *Gensim* sur python propose une version logarithmique de la mesure, car elle rend les résultats plus interprétables du fait de la diminution de l'échelle amenée par la fonction logarithme. Dans l'article « *Latente Dirichlet Allocation* », les chercheurs définissent la perplexité par la formule suivante :

$$Perplexité(D) = exp\left(-\frac{\sum_{d=1}^{D}\sum_{w=1}^{N_d}\log P(w_{d,w}|\theta_d, \Phi)}{\sum_{d=1}^{D} N_d}\right)$$

Où D correspond à l'ensemble des documents du corpus, N_d est le nombre de mots dans le corpus et $P(w_{d,w}|\theta_{d,\Phi})$ est la probabilité d'apparition du mot w en prenant en compte la distribution sujets-mot Φ de ce dernier ainsi que de la distribution sujet-document θ_d. De cela, on peut déduire le logarithme de la perplexité comme :

$$\log Perplexité(D) = -\frac{\sum_{d=1}^{D}\sum_{w=1}^{N_d}\log P(w_{d,w}|\theta_d, \Phi)}{\sum_{d=1}^{D} N_d}$$

Dans ce cas, nous cherchons un score de perplexité faible qui nous indique un modèle que ce dernier est capable de s'ajuster aux données de notre corpus. Néanmoins, cela ne veut pas dire que le modèle avec la meilleure perplexité n'est pas victime de surapprentissage. Les nettoyages peuvent faire augmenter le score de perplexité à cause du retrait de contexte que peuvent amener ces derniers. Dans notre cas, nous regardons cette métrique pour voir comment les différentes itérations de LDA s'adaptent à nos données.

Ces trois métriques nous permettent donc de mesurer la cohérence des sujets ainsi que la résilience des modèles sur notre corpus en fonction du paramètre k. cependant, on doit souligner la sensibilité de ces dernières à la variation de k. De ce fait, nous ne pouvons pas simplement choisir la valeur qui maximise nos résultats dans ces trois comme une méthode pour choisir le nombre de sujets. De ce fait, la méthode du coude de Cattell semble la plus adaptée pour affiner et définir le nombre optimal de k de nos modèles. Cette méthode consiste à détecter visuellement le ralentissement de la croissance ou de la décroissance des scores, ici en corrélation avec l'augmentation de K. Une fois notre échantillon de k à tester affiné, nous regardons de plus près manuellement les sujets formés par ces différents modèles de LDA. L'interprétabilité est au cœur de nos préoccupations. C'est pourquoi nous choisissons les vecteurs de mots composés par le LDA en fonction de

notre capacité à les nommer et à les comprendre. Finalement, nous utilisons ces métriques pour voir si nos nettoyages ont un effet concret sur les performances du modèle. Cela nous aidant à ne pas trop diminuer la dimensionnalité de ce dernier tout en étudiant les différents scénarios.

IV.2.b) Nettoyages et sélection du nombre de sujets

Plutôt, nous avons évoqué l'importance de la diminution de dimensionnalité de nos données pour rendre le LDA plus efficace. À ces fins, nous observons plusieurs étapes de nettoyages dans l'objectif de regarder l'impact que ces étapes ont sur les mesures de performances ainsi que sur la lisibilité des sujets formés. Par conséquent, nous considérons quatre modèles à évaluer.

Pour établir une base de comparaison, nous avons d'abord examiné un modèle que nous avons baptisé « brut ». Ce modèle n'a subi aucun nettoyage et contient de ce fait un grand nombre de variables. L'objectif de l'évaluation des performances de ce dernier est de nous donner une base pour voir si l'effet de nos prétraitements de nos données sur les résultats obtenus, nous le considérons comme notre expérience témoin.

Notre deuxième jeu de donnée est simplement normalisé. La normalisation de nos données consiste en la mise aux normes des différentes orthographes comme « Montreal » en « Montréal ». Une autre normalisation qui a été faite repose sur des termes médicaux. Au cours de notre exploration de données, nous avons eu l'occasion de lire les jugements fréquemment et avec l'analyse des bigrammes, nous avons remarqué l'apparition de symboles comme « l4-l5 » ou « C5-C6 ». À la lecture de notre corpus, nous avons remarqué que ces termes étaient des numéros de vertèbres ou de disques cervicaux. Nous avons décidé de les remplacer par « vertebreI, vertebreII, ..., vertebreVII » et « cervicaleI, cervicaleII, ..., cervicaleVII » pour les raisons que nous avons évoquées dans notre exploration de données. À cela, nous pouvons ajouter la suppression des numéros de dossiers présents dans les textes ainsi que la suppression des mots avec moins de trois lettres. Cet ensemble de nettoyages est réalisé en utilisant plusieurs modèles de regex qui nous permet de cibler des suites de caractères particuliers, ce qui permet de les remplacer ou de les supprimer.

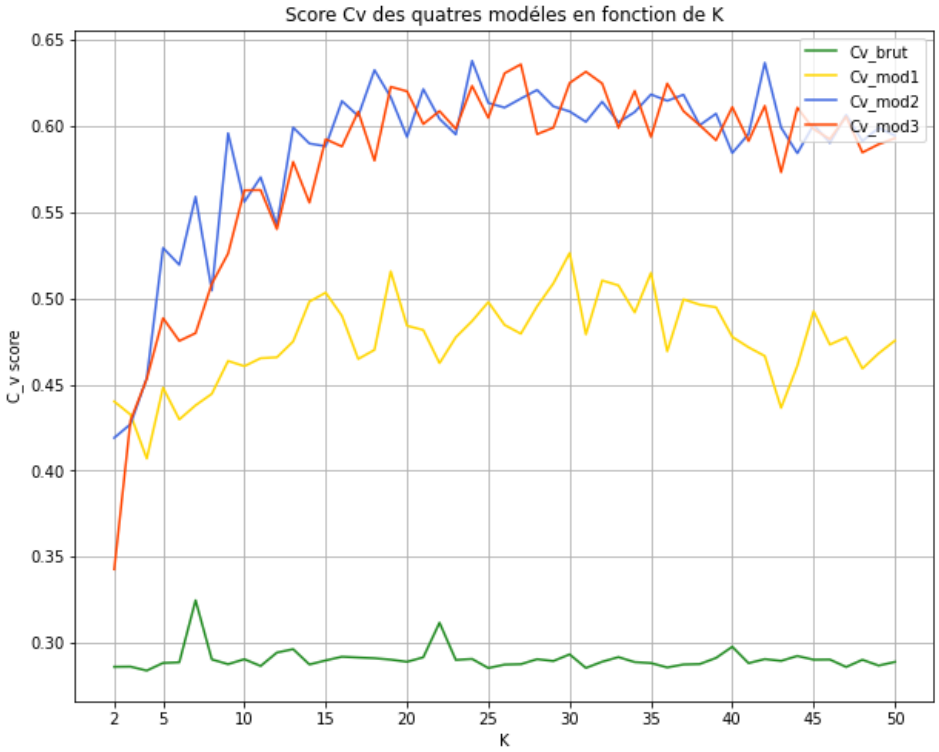
Notre troisième modèle reprend la normalisation et les nettoyages amenés par la normalisation des données, mais nous y ajoutons d'autres étapes pour diminuer la dimensionnalité. Nous commençons par lemmatiser le texte. La lemmatisation est un procédé qui permet de regrouper plusieurs mots de même forme dans leurs racines. Ce processus est commun en NLP et nous pouvons facilement l'implémenter à l'aide de la bibliothèque spacy pour python. À la suite de la lemmatisation, nous retirons une liste de stopwords, que nous avons déterminé avec une méthode heuristique. En effet, en regardant plusieurs résultats de LDA, nous avons vu que certains mots rendaient complexe la constitution de sujets distincts. Le cas le plus flagrant pour nous concerne le mot « exposition ». En effet, sans le retrait de ce dernier, nous avions régulièrement des sujets composés des mots : «exposition», «pulmonaire», «auditifs», «artrose», «amiant», «surdité». Nous voyons dans nos annexes 15 et 16 que, malgré ces nettoyages avec les stopwords, la dissociation des sujets peut s'avérer complexe. Cependant, ce nettoyage permet tout de même de rendre les sujets plus lisibles, en particulier dans le cas des maladies. En ce qui concerne la confusion entre les maladies pulmonaires et auditives, notre intuition repose sur le fait que l'examen des cas de maladies professionnelles reliées à la surdité, aux maladies pulmonaires et à l'arthrose se base sur une exposition à certains risques. Ce qui est intéressant ici est l'idée que le même vocabulaire est utilisé par le tribunal pour parler de pathologies à priori différentes. Cela peut nous indiquer que l'examen logique effectué par le juge est similaire pour ces diverses affections. Cependant, nous avons estimé dans ce mémoire qu'il était préférable de conserver des sujets distincts en fonction des pathologies et nous avons retiré le mot du vocabulaire. Cette liste contient aussi des noms de tribunaux ou des abréviations qui n'apportent pas d'information importante à notre étude. Finalement, nous avons fait un nettoyage par fréquence pour réduire un peu plus la dimensionnalité. Ce nettoyage retire les mots présents moins de 100 fois dans le corpus et plus de 20 000 fois dans ce dernier. Pour effectuer cela, on regarde la distribution des mots dans le corpus en comptant la fréquence d'apparition des tokens. La communauté du NLP semble s'accorder sur le fait que le nettoyage par fréquence n'obéit qu'à une règle du pouce. Dans notre cas, pour choisir nos intervalles de coupes, nous avons utilisé un mélange de

logique et de tâtonnements. Nous avons commencé par retirer 2% des tokens les plus fréquents et 5% des moins fréquents. Puis, nous avons augmenté progressivement le nettoyage en fonction des résultats que nous obtenions dans nos itérations de LDA.

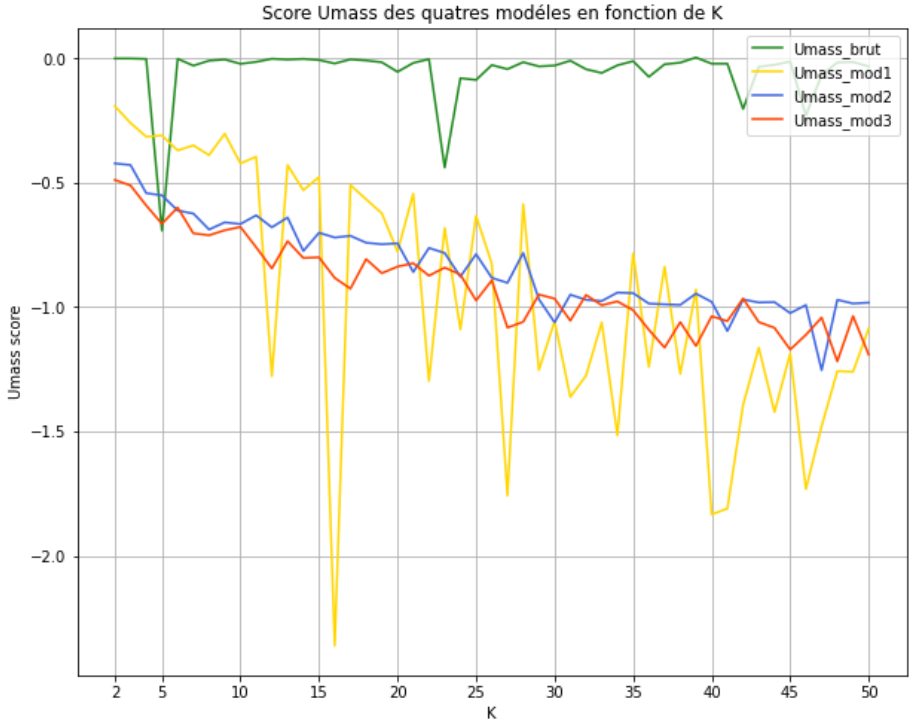
Finalement, notre dernier modèle ajoute un nettoyage en utilisant le « Part-of-speech » (POS tagging). Le *POS tagging*, permet d'identifier la nature et la fonction d'un mot dans une phrase en d'autres termes, cette méthode, nous indique si le mot est un verbe, un nom, un adjectif, etc. Grâce à spacy, nous pouvons mettre en place du *POS tagging*, ce qui nous permet ultimement de supprimer tous les mots qui ne sont pas des verbes ou des noms dans notre cas. Nous avons décidé de ne conserver que les verbes et les noms, car les verbes d'action ont une importance dans l'exécution d'un jugement. Les noms sont en général les éléments qui portent le plus de sens quand ils sont pris indépendamment du reste de la phrase. Après ce filtrage, nous faisons une lemmatisation, un retrait des *stopwords* et un nettoyage par fréquences.

Pour synthétiser nos étapes de nettoyages, nous avons produit un diagramme en annexe 14 qui résume les étapes de nettoyages de chacun de nos modèles.

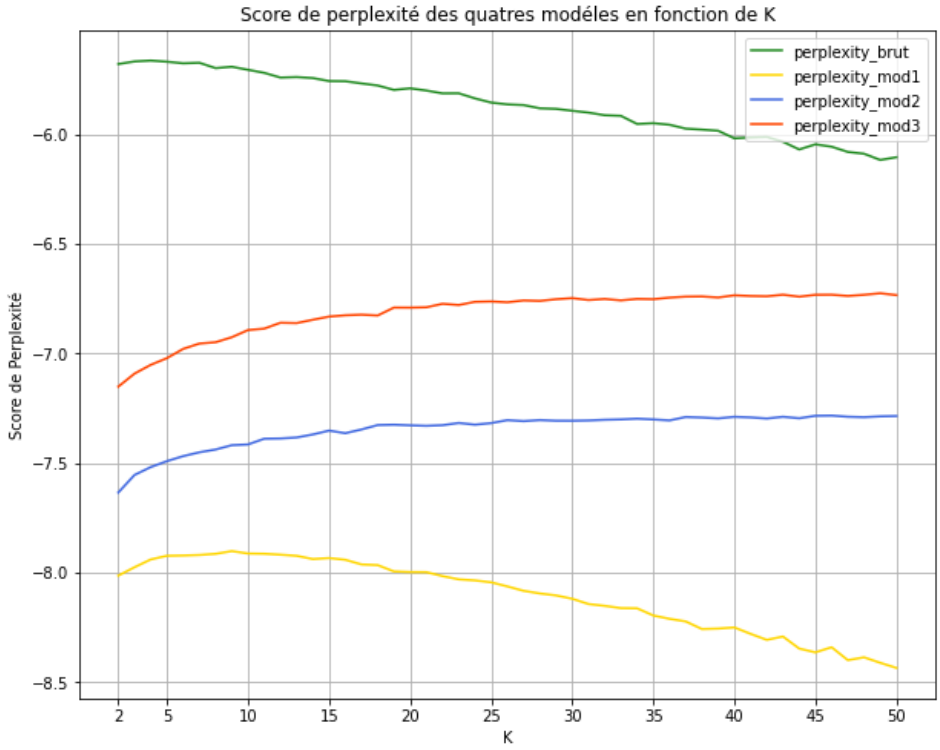
Une fois ces nettoyages effectués, nous avons procédé aux calculs de cohérence et de perplexité pour $k \in \{2,3, \dots ,50\}$. Nous commençons nos calculs à partir de deux sujets, car c'est la valeur minimale pour effectuer un regroupement. Concernant la valeur maximum de 50, nous avons estimé qu'au-dessus de ce nombre de sujets l'interprétation du modèle serait trop complexe outre le fait qu'il est peu probable que le modèle optimal soit passé par cette marque. En utilisant la méthode évoquée dans notre sous-chapitre sur les mesures de cohérences, nous avons produit trois graphiques qui mettent en parallèle les performances des quatre modèles ci-dessous dans les graphiques 1, 2 et 3.



Graphique 1 : score de cohérence Cv en fonction de k pour les quatre modèles



Graphique 2 : score de cohérence Umass en fonction de k pour les quatre modèles



Graphique 3 : score de perplexité en fonction de k pour les quatre modèles

Ces trois graphiques nous révèlent plusieurs informations utiles quant à la sélection de modèles. Pour commencer notre analyse, nous regardons notre modèle brut, qui est notre expérience témoin dans cette situation. Nos calculs nous révèlent que, de tous les modèles, il a la pire cohérence Cv et Umass ainsi que la pire perplexité. Cela est dû à sa complexité, car il possède une multitude de dimensions, ce qui rend difficile la conception de thèmes uniques et distincts. De plus, la présence de nombreuses matrices clairsemées impacte négativement la perplexité du modèle. En réalité, nous ne cherchons pas à optimiser le paramètre k dans ce cas. En effet, cette modélisation de sujets produirait probablement un ensemble de vecteurs de mots non interprétables, car composée de mots grammaticaux et d'artefacts de mise en page.

Nous regardons ensuite le modèle qui a subi la normalisation des données. On voit une nette amélioration de toutes les métriques. Ce qui nous intéresse plus particulièrement est le score Umass qui, s'il est nettement meilleur, comporte beaucoup d'optimums locaux.

Par exemple, pour k = 13 et k = 15, on observe des minimums locaux très prononcés. Cela nous indique que la cohérence des sujets composés avec ces données est très sensible à la variation de k, rendant difficile l'optimisation du modèle en se basant sur le score Umass. En revanche, quand on regarde le score Cv, la courbe semble atteindre un plateau à partir de 15 sujets. Quant à la perplexité, la courbe se stabilise entre 15 et 18 sujets. Ce qui peut paraitre surprenant est que des modèles avec plus de nettoyage ont un score plus élevé indiquant une moins bonne perplexité. On peut expliquer cela par l'impact des étapes de lemmatisation et de nettoyage par POS qui enlève du contexte à notre vocabulaire rendant le LDA moins résilient aux corpus.

Cela nous amène à analyser les modèles 2 et 3 qui ont été plus amplement nettoyés. Bien que le score CV soit relativement similaire pour les deux modèles, le score Umass semble indiquer une légère amélioration de la cohérence pour le modèle qui n'a pas subi de nettoyage POS. Quand on regarde la perplexité pour ces deux situations, le modèle 3 ressort plus fiable que le deuxième. Pour le modèle 1, les courbes semblent atteindre un plateau pour $k \in [15, 25]$. En ce qui concerne les modèles 2 et 3, les pentes des courbes semblent généralement s'atténuer dans l'intervalle $k \in [18, 25]$. Afin de sélectionner la segmentation de sujet la plus cohérente, nous avons décidé d'analyser les résultats de LDA pour des valeurs de k dans un intervalle un peu plus large, soit pour le modèle 1, $k \in \{12, \dots, 27\}$ et pour le modèle 2 et 3, $k \in \{15, \dots, 27\}$En élargissant le champ de notre recherche par essais et erreurs, nous souhaitons nous assurer de ne pas manquer une modélisation finale plus claire et plus facile à interpréter, malgré les différentes métriques de sélection que nous avons utilisées.

Cette méthode nous a amenés à choisir un modèle avec 15 sujets pour le premier et 19 pour les modèles 2 et 3. Nous pouvons voir les résultats de ces différents LDA en annexe 15 et 16 ainsi que dans le tableau suivant.

Modèle 3						
Numéro du sujet	**Nom du Sujet**	**Mot 1**	**Mot 2**	**Mot 3**	**Mot 4**	**Mot 5**
0	**genou**	0.096*"genou"	0.035*"interne"	0.034*"déchirure"	0.019*"ménisqu"	0.017*"méniscal"
1	**prestation de handicape**	0.037*"déficience"	0.032*"coût"	0.019*"partage"	0.019*"conséquence"	0.018*"handicap"
2	**Frais de soins**	0.035*"frais"	0.032*"aide"	0.031*"assistance"	0.026*"soin"	0.025*"coût"
3	**Représentation**	0.009*"dernier"	0.008*"syndicat"	0.007*"situation"	0.007*"personne"	0.006*"rencontre"
4	**Condition Psychologique**	0.025*"trouble"	0.014*"psychologique"	0.013*"adaptation"	0.012*"symptôme"	0.010*"dépression"
5	**Blessure aux pieds**	0.041*"fracture"	0.032*"pied"	0.029*"chevill"	0.017*"syndrome"	0.014*"douloureux"
6	**Indemnité de remplacement de revenu**	0.053*"indemnité"	0.050*"revenir"	0.033*"remplacemer"	0.019*"capacité"	0.015*"assistance"
7	**Capacité de retour**	0.026*"convenable"	0.021*"réadaptation"	0.014*"capacité"	0.012*"tâche"	0.010*"physique"
8	**Partage de prestation et responsabilité**	0.038*"coût"	0.028*"prestation"	0.021*"imputation"	0.016*"activité"	0.014*"tiers"
9	**Rechute**	0.075*"aggravation"	0.071*"rechute"	0.050*"récidiv"	0.033*"initial"	0.021*"récidive"
10	**Blessure main/bras**	0.039*"poignet"	0.037*"main"	0.028*"coude"	0.020*"épicondylite"	0.019*"doigt"
11	**Entorse cervicale**	0.024*"cervical"	0.018*"entors"	0.013*"normal"	0.012*"flexion"	0.010*"mouvement"
12	**Maladie respiratoire**	0.027*"pulmonaire"	0.021*"amiant"	0.011*"maladie"	0.010*"comité"	0.010*"cancer"
13	**Causalité de la blessure**	0.026*"blessur"	0.020*"présomption"	0.010*"cause"	0.010*"occasion"	0.009*"existence"
14	**Délai de contestation**	0.019*"raisonnable"	0.016*"délai"	0.016*"déler"	0.016*"contestation"	0.012*"question"
15	**Déchirure de la coiffe des rotateurs**	0.048*"déchirure"	0.025*"coiff"	0.023*"rotateur"	0.022*"tendon"	0.012*"biceps"
16	**Blessure au dos**	0.054*"discal"	0.032*"sacrum"	0.031*"vertebrev"	0.025*"hernie"	0.021*"herni"
17	**Blessure épaule et bras**	0.032*"tendinite"	0.023*"supérieur"	0.018*"mouvement"	0.017*"épaule"	0.016*"bras"
18	**Maladie de l'audition**	0.024*"bruit"	0.015*"surdité"	0.012*"tâche"	0.011*"auditif"	0.011*"risque"

Tableau 3 : Tableau résumé des vecteurs de mots pour le modèle 3

Notre analyse révèle que le modèle 3 à 19 sujets est le plus lisible et interprétable. En effet, chaque vecteur de mot est tout de même très lisible et relatif à une forme de pathologie ou un sujet relié au cadre du droit. Le modèle 1 est rempli de mots difficiles à interpréter comme « elle » ou « date », par exemple. Malgré cela, on peut tout de même constater de cohérences pour certains vecteurs formés, comme pour le sujet 2, qui est relatif aux blessures au dos. Cela étant dit, la majorité des regroupements sont ininterprétables et contiennent trop de bruit statistique pour les rendre cohérents par notre lecture. Le modèle 2, lui, ressemble beaucoup à notre modèle final. Cependant, comme

nous l'avons évoqué précédemment, nonobstant le retrait des *stopwords*, nous avons des vecteurs comme le 18 (annexe 16) qui sont confus à interpréter. En effet, comme nous l'avons souligné plutôt, le regroupement de notions relatives à des maladies pulmonaires et à des pathologies affectant l'audition rend la lecture de ce dernier compliqué. On peut aussi souligner le sujet 7 dans la même annexe qui semble être relatif au mouvement, mais qui ne nous indique pas vraiment un thème clair et identifiable à lui attribuer. Les nettoyages ont cependant porté leurs fruits, car on voit se distinguer des sujets portant sur les enjeux psychologiques, par exemple, ou encore des sujets spécifiques relatifs aux déchirures de la coiffe des rotateurs. Finalement, lorsque l'on regarde le dernier modèle présenté dans le Tableau 3, on remarque que la diminution de dimensionnalité a bien fonctionné nous donnant des sujets distincts. Néanmoins, certains de ces vecteurs nous ont demandé de plus amples recherches pour comprendre leurs contenus. Dans cette catégorie, on retrouve le sujet 2 relatif à la représentation. Les mots qui le composent sont compliqués à relier les uns aux autres pour un novice dans le domaine du droit du travail. Ainsi, pour mieux comprendre ce regroupement, nous avons décidé de lire les 10 jugements qui avaient le plus de proximité avec ce sujet. À la lecture de ces derniers, nous avons remarqué que ces jugements traitaient d'enjeux de représentation par le syndicat. Indéniablement, ces textes abordent souvent la dernière rencontre avec le syndicat, le statut de salarié de l'employeur, ainsi que le contenu des réunions avec le syndicat. De plus, dans tous ces jugements, le syndicat a joué le rôle de représentant. Parfois, la représentation du syndicat est contestée, ce qui amène le tribunal à examiner la dynamique des relations entre l'employé et son représentant. C'est pourquoi le mot « temps » est utilisé. Pour d'autres sujets, nous avons effectué quelques recherches pour savoir si les mots pouvaient correspondre à des symptômes précis, comme pour le sujet 15 qui aborde la déchirure de la coiffe des rotateurs. Finalement, le sujet 6 et 7 est très proche sémantiquement l'un de l'autre, mais la composante financière du sujet 6 ressort avec le mot indemnité ou encore montant ou prestation. En complément à cela, nous avons poursuivi nos recherches en analysant les composantes essentielles relatives aux indemnités de remplacement et nous avons trouvé sur le site internet de la CNESST (CNESST, 2024) que l'examen pour évaluer la date de début

ou de fin de cette indemnité comportait de nombreux points sémantiques communs avec ce vecteur de mot. Cette étape de notre projet souligne l'importance de la compréhension et de la maitrise des éléments du droit pour sélectionner les modèles cohérents. Dans ce travail, nous avons complété nos analyses quantitatives avec des recherches superficielles pour nous assurer de la cohérence de notre modèle, mais l'avis d'un expert juridique serait un atout pour permettre de mieux comprendre nos sujets et d'approfondir nos analyses. Afin de finaliser notre analyse, nous avons calculé les scores de performances de ce modèle. Nous obtenons un score de cohérence Umass de -2,68, Cv de 0,61 et un score de perplexité de -8.67. Ces scores peuvent être modifiés, car le modèle de LDA reste stochastique, ce qui introduit de l'aléa avec l'utilisation de l'algorithme. De ce fait, la perplexité de cette modélisation de sujet est améliorée là où la cohérence Umass est légèrement dépréciée. Le score Cv est relativement stable avec nos itérations précédentes. Nous soulignons ici que nous avons fait plusieurs modélisations avec les mêmes paramètres qui nous donnaient des résultats très proches, mais nous avons tout de même choisi le résultat le plus cohérent à notre lecture. Finalement, nous n'avons pas ajusté les autres hyperparamètres que sont α et ß. Le LDA postule une distribution des sujets à priori et le paramètre α sert à déterminer cette dernière. Plus on suppose qu'un texte aborde plusieurs sujets, plus on cherche à augmenter ce dernier paramètre. Dans notre cas, nous avons gardé la valeur par défaut de ce paramètre, soit α = 1. Ce choix est motivé par les prémices qui consistent à voir les jugements comme des textes avec un sujet central qui est la question de droit adressé au tribunal, puis des sujets périphériques se greffent autour de cela pour alimenter la réflexion et la preuve. Le paramètre ß régit la distribution à priori des mots dans les thèmes composés. Plus le dernier est élevé, plus on postule que les sujets contiennent une grande variété de mots. À l'inverse, moins il est élevé et plus nos sujets sont concentrés autour de quelques mots. Nous avons aussi choisi de garder un ß par défaut qui donne la même probabilité à chaque mot d'appartenir à un sujet. Nous avons fait ce choix, car nous n'avions pas d'idée prédéterminée de la distribution des mots dans les sujets. L'ajustement de ces paramètres est un élément à considérer pour d'autres travaux, mais les résultats obtenus sont satisfaisants pour la suite de ce travail.

IV.3) Analyse de la structure du réseau de graphe

Dans notre chapitre sur la formalisation de la tâche concernant la dimension juridique ainsi que dans le sous-chapitre sur la collecte des données de graphes, nous avons expliqué des éléments d'extraction de données. Une fois nos tables de données constituées, nous procédons à la modélisation des citations en utilisant un mélange entre Gephi et la bibliothèque NetworkX de python. Ce qui nous intéresse avant de procéder à l'étude de nos résultats est la structure du graphe et plus particulièrement les potentiels regroupements de nœuds les uns autour des autres. Pour analyser cela, nous nous appuyons sur le calcul de modularité que comporte Gephi et qui s'appuie sur la méthode de Louvain que nous décrivons ici. Puis, nous examinons de plus près la modularité et les regroupements formés dans le graphe pour le cas distinct des citations de jurisprudence, des lois et, finalement, pour l'ensemble du graphe. Ensuite, nous proposons de regarder la structure du graphe. Pour cela, nous divisons le graphe en deux, le réseau des jurisprudences et le réseau des lois. À la suite de l'analyse des différentes métriques pour ces réseaux, nous regardons le réseau au complet pour mieux comprendre sa structure avant de présenter les résultats dans le chapitre suivant.

IV.3.a) Mesures d'évaluation du graphe

Afin d'analyser notre graphe, nous nous appuyons sur des éléments graphiques, mais nous regardons aussi des éléments statistiques. Dans cette partie, nous expliquons les principales métriques que nous utilisons pour analyser le graphe.

Pour commencer, nous regardons les éléments relatifs aux degrés des nœuds. Le degré est défini par le nombre de liens qu'un nœud possède avec les autres. Il existe deux types de degrés, sortant et entrant en fonction de la destination de la connexion. Une connexion partant du nœud est considérée comme sortante là où celle pointant vers le nœud est entrante. L'intérêt de l'étude du degré repose dans l'analyse de la distribution des connexions des nœuds dans le graphe. Dans notre cas les textes sont souvent les points d'origines des arêtes, donc nous nous attendons à voir un degré entrant plus fort dans le

graphe. Plus généralement, cette analyse nous permet de vérifier les éléments soulevés dans notre exploration de données.

La deuxième étude que nous regardons concerne les composantes connexes. Les composantes connexes sont des éléments du graphe qui sont plus ou moins complexes et qui ne sont reliés à aucun sous-graphe plus grand (Clark, 1991). En termes plus triviaux, on parle ici de parties flottantes du graphe qui ne sont pas reliées aux autres. Pour ces calculs, nous nous appuyons sur les outils proposés par Gephi, qui utilisent l'algorithme « *depth-first search* » (DFS). La DFS est une méthode par itération qui permet de chercher les nœuds voisins qui ressemblent au BFS évoqué dans notre chapitre sur la dimension juridique. Les principales différences entre ces deux approches sont surtout liées aux méthodes d'optimisation du calcul. Cependant, la méthode proposée par Gephi s'avère efficace dans notre cas. Par ce biais, on obtient le nombre de composantes connexes que notre graphe comporte ainsi que leurs degrés de connexion c'est-à-dire la taille qu'elles ont et l'importance des connexions qu'elles comportent. Cela pourrait nous montrer des portions du graphe isolées qui présenteraient certaines particularités. Cela peut révéler des jugements avec une nature spécifique que nous devons alors expliquer. Nous devons tout de même rappeler qu'un certain nombre de textes ne possèdent pas de citations, ce qui les amène à former des composantes isolées que nous traitons de manière particulière.

Dans un graphe des sous-ensembles de nœuds plus densément regroupés peuvent former ce que l'on appelle une communauté. Ces dernières nous indiquent généralement une forte interconnexion entre les nœuds de cette portion du graphe. Dans notre cas, nous souhaitons regarder les communautés dans le graphe de plus proches afin de voir si des jugements ou des lois ont un lien plus important que les autres dans notre corpus. Dans le but de mesurer cela, nous calculons la modularité dans le graphe pour le diviser en différents groupes nous permettant d'isoler les communautés de nœuds. Comme évoquer plutôt, la modularité est calculée via la méthode de Louvain par Gephi. Cette méthode est basée sur les travaux de Vincent Blondel et al. Qui proposent une approche pour calculer la modularité plus efficacement dans de larges réseaux de graphes (Blondel et al., 2008). La modularité peut être définie par l'équation suivante :

$$Q = \frac{1}{2m}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

Où, Q est la modularité, k_i et k_j sont la somme des poids des arêtes liées aux nœuds i et j, m est le nombre d'arêtes du le graphe, N est le nombre total de nœuds dans le graphe, c_i et c_j sont les communautés auxquels appartiennent le nœud i et j. Finalement, δ représente la fonction delta de Kronecker et, dans le cas de la modularité si $\delta(c_i, c_j) = 1$ alors c_i, c_j appartiennent au même regroupement, en revanche si cette variable prend la valeur zéro alors c_i, c_j Se trouvent dans deux ensembles distincts. La modularité est une échelle comprise entre -1 et 1 qui permet de mesurer la possibilité de diviser le réseau en communautés. Dans la pratique, il est généralement commun que la modularité soit positive. Ainsi, une modularité proche de 0 indique que le réseau ne peut pas se diviser en structures communautaires cohérentes. À l'inverse, une modularité proche de 1 nous indique une forte cohérence dans les structures de communauté du réseau. Si cette métrique est pratique pour déterminer les ensembles sous-jacents à un réseau, l'enjeu réside dans les méthodes de calculs pour obtenir ce résultat dans des graphes de grandes tailles. C'est à cette problématique que l'algorithme de Louvain propose une réponse en deux temps (Clark, 1991). Dans la première partie de l'algorithme, les auteurs postulent que chaque nœud correspond à une communauté et que, par ce fait, le graphe est composé d'autant de communautés que de nœuds qui le composent. À la suite de cela, on évalue le gain ou la perte de modularité du graphe en regroupant les nœuds voisins en communauté. Dans le cas où ce regroupement occasionne un gain, alors on regroupe les nœuds ensemble. On répète le processus pour chacun des nœuds du réseau jusqu'à ce qu'aucune amélioration ne soit possible. Puis, dans la deuxième phase de l'algorithme, on reconstruit un graphe en prenant pour nœud les communautés de nœuds formés dans la première partie. C'est grâce à la synergie de ces deux méthodes que nous pouvons assurer un calcul optimal dans des réseaux larges.

IV.3.b) Le réseau des citations de jurisprudences

La première analyse que nous menons est relative au réseau des citations de jurisprudence. Nous proposons la modélisation suivante de ce réseau faite sur Gephi, où les points bleus représentent les jurisprudences et les points rouges nos textes du corpus.

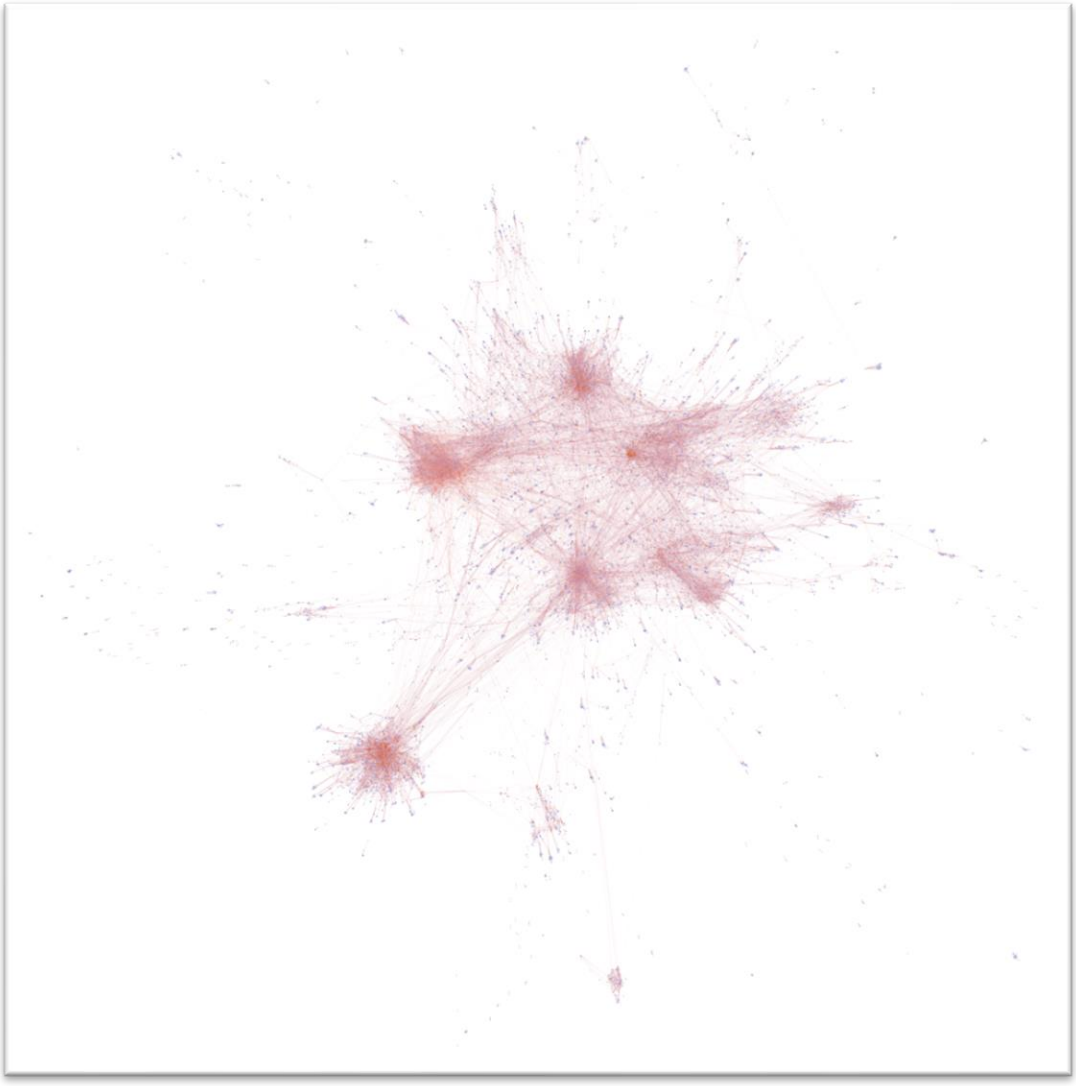
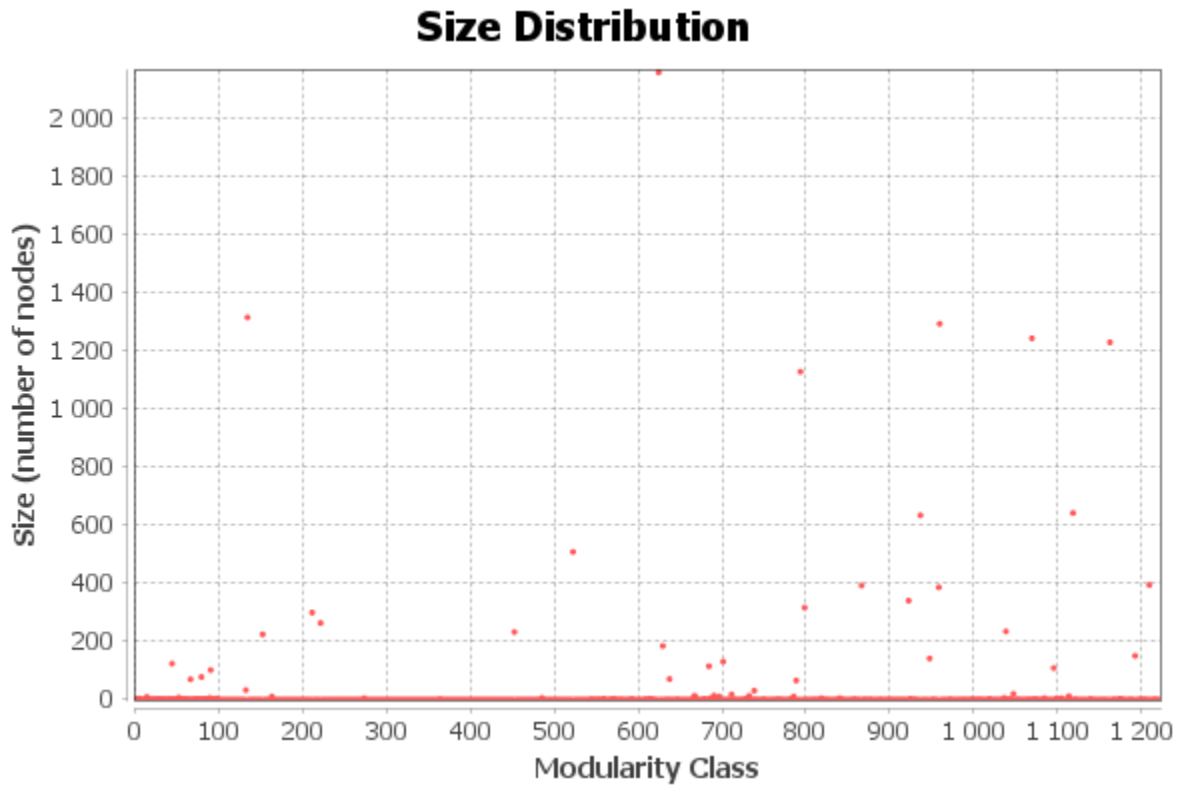


Figure 4 : réseau de graphe des textes de références et des citations de jurisprudence

L'analyse visuelle de ce graphe nous indique des zones de densité plus élevées, qui semble former des centres d'attraction et des communautés. Nous commençons par regarder la distribution des degrés dans le graphe. Le graphe comporte en moyenne un degré de 2.49 par nœuds. Ce résultat est amoindri du fait que les vertex représentant la jurisprudence n'ont que des connexions rentrantes dans notre modélisation. Dans l'annexe 17, 18 et 19,

on peut voir la distribution des degrés dans le graphe. On constate que certains nœuds ont un degré faible, ce qui nous indique des connexions faibles avec le reste du graphe voir potentiellement une composante connexe. En regardant nos données, nous constatons que certains textes n'ont pas de connexions à de la jurisprudence, ce qui explique ces faibles degrés. En regardant de plus près les données, nous réalisons que 953 textes de notre corpus ne contiennent pas de référence à la jurisprudence, ce qui représente un peu plus de 17% du corpus. Quand nous regardons notre annexe 12, ce chiffre est cohérent avec la distribution des citations dans le corpus que nous avons proposé dans notre exploration de données. En regardant de plus près la littérature légale sur le sujet de la jurisprudence, on peut expliquer cela par plusieurs facteurs. Charles Breton-Demeule (Breton, 2021) explique que la citation de la jurisprudence dépend de la nature du jugement et que, dans des cas par souci de simplification préfère éviter le jargon juridique et les références à la jurisprudence. De plus, il ajoute que les procédures de conciliation mises en place par les tribunaux se concentrent généralement sur les solutions pour résoudre le conflit et accordent donc moins d'importance aux arguments de droit. Finalement, le TAT peut aussi s'appuyer sur la loi sans citer des jugements antérieurs dans des cas où l'état du droit est solidifié par des décisions s'imposant comme des références évidentes pour les parties. Pour vérifier ces hypothèses, nous avons décidé de lire un échantillon de 50 textes n'ayant pas de connexion au réseau de citation. Nous avons rapidement constaté que la majeure partie de ces derniers relèvent de la mitigation de conflits ou de la défense d'avis médicaux en accord avec ce que nous avons décelé dans notre exploration de données. Ces jugements sont alors axés sur des points spécifiques de loi ou des comptes rendus d'experts qui s'appuient sur de la documentation médicale. Par le fait même, on peut comprendre que nous obtenions 1184 composantes qui comportent ces textes sans liens jurisprudentiels. Pour comprendre la densité du graphe, nous analysons en parallèle la modularité qui est de 0,828, nous indiquant une structure de communauté forte et densément connectée. Nos calculs nous indiquent que le graphe est composé de 1 222 communautés nous indiquant le nombre de graphes non connectés. Encore une fois, on peut considérer que les textes qui

ne s'appuient pas sur la jurisprudence représentent une majorité de ces communautés, comme nous l'indique le graphique suivant.



Graphique 4 : Modularité dans le graphe des jurisprudences

Ce graphique de modularité représente la distribution des communautés en fonction du nombre de nœuds qui les composent. Notre intuition que des communautés avec peu de nœuds sont présentes est ici validée lorsque l'on regarde en bas du graphique. Ce qui nous intéresse plus particulièrement est l'ensemble des communautés qui comptent beaucoup de nœuds et qui sont donc très denses. En mêlant cette analyse avec l'étude visuelle du réseau offerte dans la figure 4, nous constatons que des portions du graphe sont plus densément peuplées et, par ce fait, méritent une attention plus particulière de notre part. L'étude approfondie de ces communautés est proposée dans notre partie sur les résultats, ici nous nous concentrons principalement sur la structure du graphe.

En résumé, nous avons vu que le réseau des citations de jurisprudences est composé de nombreuses composantes et communautés. La modularité proche de 1 de ce dernier

nous indique une densité de connexion des communautés reflétant la complexité de ce graphe. À la lumière de cette analyse, nous pouvons aussi détecter la présence de nombreux jugements ne faisant que peu ou pas du tout référence à la jurisprudence. Cette particularité peut s'expliquer par la nature des jugements qui ne mettent pas forcément au centre de leurs raisonnements des précédents, mais plutôt des éléments concrets de lois ou médicaux.

IV.3.c) Le réseau des citations de lois

Pour poursuivre notre enquête sur le réseau de graphe, nous décidons de regarder la structure des données pour la citation de la loi dans les textes. Avec un filtre, nous avons procédé à la modélisation de ce dernier sur Gephi, que nous proposons dans la figure 5 ci-dessous.

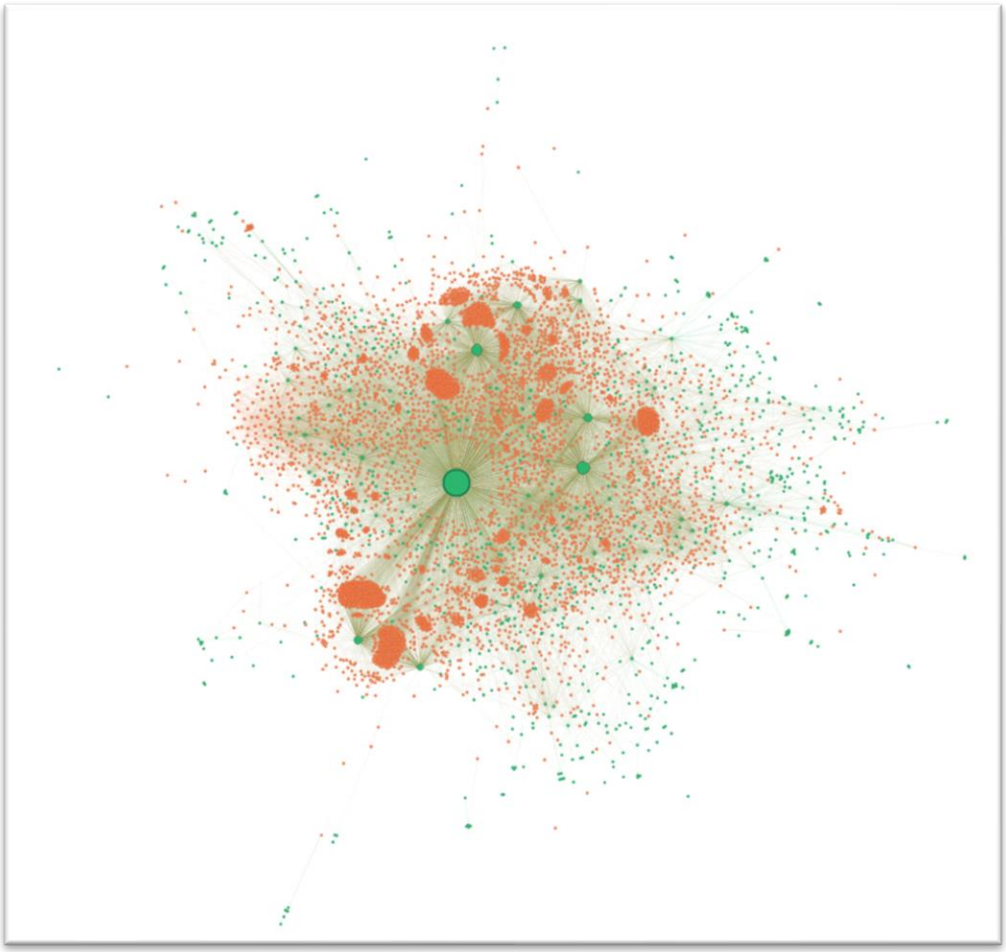
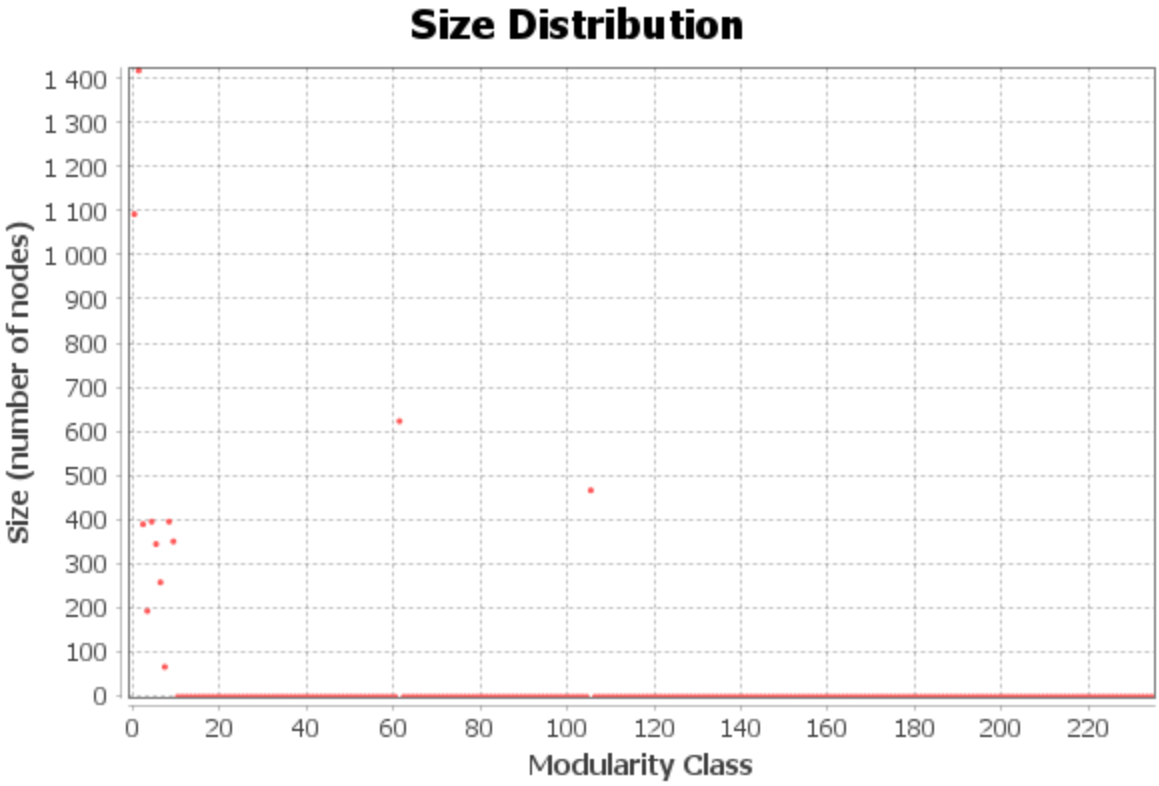


Figure 5 : réseau de graphe des textes de références et des citations de lois

Dans cette figure, les lois sont représentées par les points verts et les lois en orange. Ce dernier semble nous indiquer que certaines lois ont une grande polarité et sont citées fréquemment. Ce constat est cohérent avec la nature du droit du travail concernant les accidents du travail, qui s'appuie sur la loi sur les accidents et les maladies professionnelles et la loi instituant le tribunal administratif du travail, qui à elles deux abordent une grande portion des sujets susceptibles d'être discutés au TAT. De plus, nous avons opté pour la modélisation de la présenté dans la figure 3 pour de considérations pratiques. Cette forme de modélisation ajoute du poids aux chapitres de lois parents des articles cités. Dans le but de mieux comprendre la structure de notre graphe, comme c'est le cas pour les jurisprudences, nous commençons par examiner les statistiques relatives aux degrés dans le graphe.

Nous proposons les graphiques présents dans les annexes 20, 21 et 22 qui représentent la distribution des degrés entrant, sortant et cumulés dans le réseau de graphe des lois. On remarque que la distribution des degrés pour chaque nœud est décroissante. Il y a beaucoup de nœuds qui n'ont pas beaucoup d'arêtes et de rares nœuds qui ont un fort degré, ce qui confirme nos intuitions décrites dans notre exploration de données. Il est crucial de mettre en évidence que notre approche axée sur la modélisation des données révèle un fait intéressant : le degré entrant est généralement associé à une citation, tandis que le degré sortant est lié à un texte du corpus qui fait référence à cette citation. Le degré sortant est particulièrement concentré autour de 0. Cela s'explique par le nombre généralement faible de lois abordées par le tribunal, qui ne traite que d'un ou deux enjeux de droits à la fois en général. Le degré d'entrée, quant à lui, est plus large, ce qui suggère que plusieurs textes législatifs contiennent un grand nombre de références. Nous, regardons ensuite le nombre de nœuds avec un degré nul et nous trouvons 223 textes sans connexion. Cela explique le nombre de composantes faiblement connectées qui est de 224. Une de cette composante cite un texte de loi qui n'est pas relié au graphe là où le reste est des textes qui ne s'appuient pas sur la loi. La raison pour laquelle ces textes ne citent pas la loi est probablement la même que dans le cas des jurisprudences. À des fins de simplification du jugement, certains textes d'avis ou de médiation ne citent pas la loi, mais se basent sur les

faits. Cela nous amène donc à regarder la modularité dans le graphe pour comprendre comment ce dernier se structure. On trouve une modularité de 0,440, ce qui nous indique une connexion des communautés plus faibles que dans le cas des jurisprudences cependant acceptables. On dénombre 235 communautés dont une grande partie est constituée des 224 composantes faibles nous laissant avec 11 communautés qui ont des schémas de liens identifiables. Le graphique suivant nous donne la répartition des communautés en fonction de leurs tailles.



Graphique 5 : Distribution des communautés de nœuds dans le graphe des citations de lois

On peut voir ici le réseau de communauté se détacher en opposition aux textes sans citation regroupés sur la ligne du bas. On voit des maximums se détacher qui sont probablement les lois que nous avons évoquées plus haut. Le graphe relatif aux citations de lois à une structure plus concentré autour de quelques points de centralité. Nous devons inclure cela dans nos analyses pour comprendre les résultats de notre mémoire. En cohérence avec la figure 5 et les analyses quantitatives, nous pouvons déjà déceler des noyaux de centralités

plus distincts en opposition aux jurisprudences qui ont un réseau de citation plus dense et complexe à analyser.

IV.3.d) Le réseau de citation complet

Finalement, nous analysons la structure du graphe complet dans cette dernière partie. Si, dans notre travail, nous détaillons la structure du graphe en sous-graphes de jurisprudences et de lois, nous avons commencé par construire le graphe complet, puis nous avons filtré les nœuds de lois et de jurisprudence afin d'analyser les graphes séparément. Effectivement, pour créer notre graphe dans l'application Gephi, nous avons dû réviser les données de notre extraction. En repassant dans les textes sans référence que le premier graphe nous donnait, nous avons remarqué que quelques références n'avaient pas été extraites. Parmi ces références, on trouvait d'anciennes classifications de tribunaux comme T.T., qui sont l'ancien nom du TAT et que nous n'avions pas dans notre guide des références juridiques. Ces aménagements concernent 30 textes du corpus et ont été intégrés dans toutes nos analyses antérieures. Cela nous amène au graphe complet présenté dans la figure 6. Cela nous laisse avec un graphe bipartite qui comporte 17 447 nœuds avec 65 866 arrêtes.

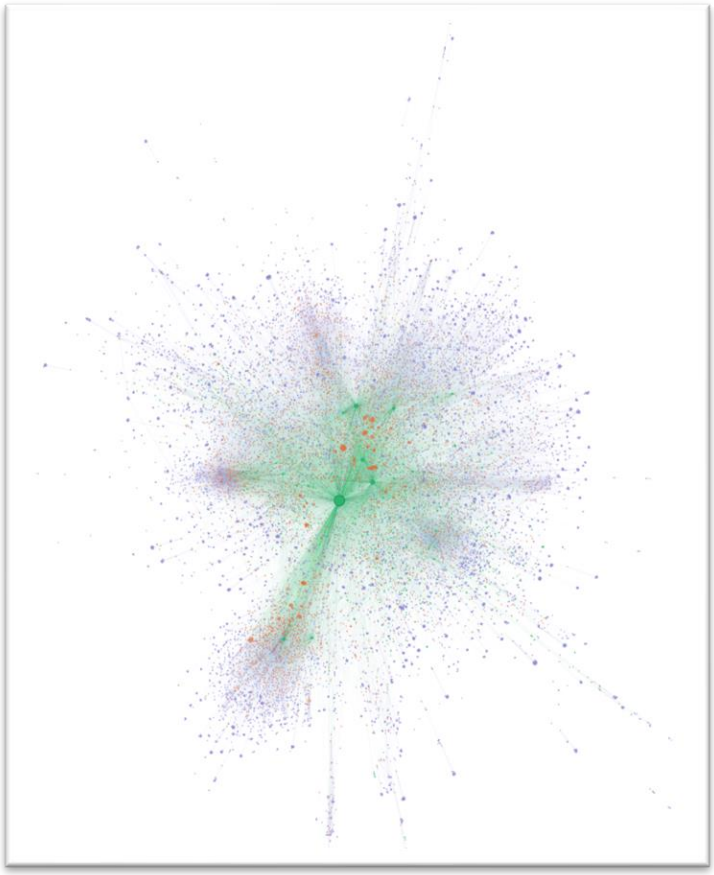
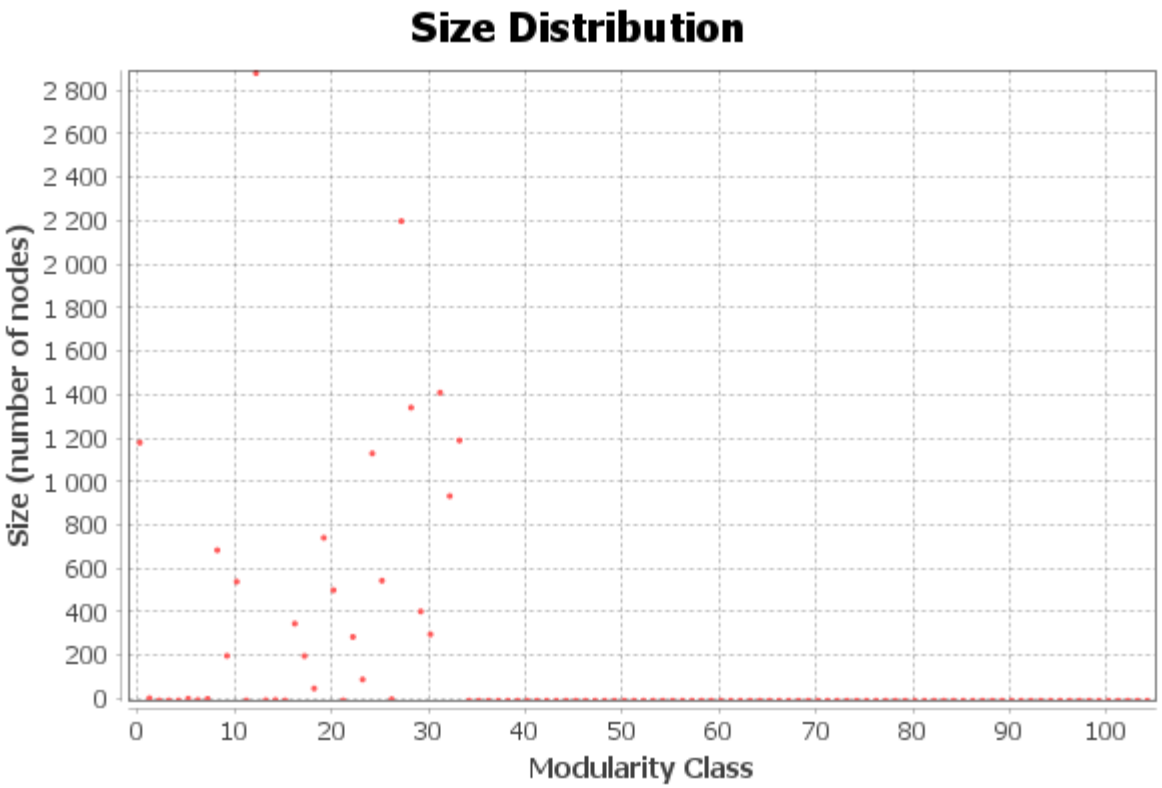


Figure 6 : Réseaux de graphe de l'ensemble des citations du corpus

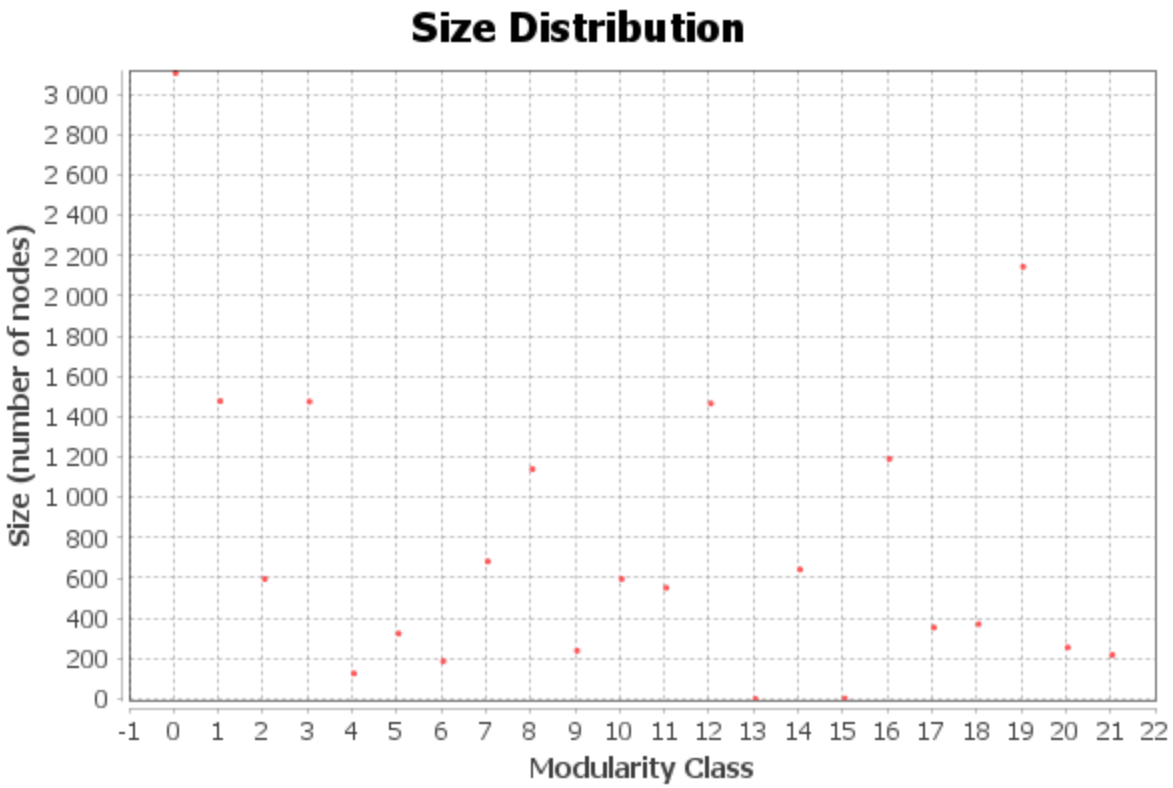
Dans la figure 6, les nœuds orange représentent les textes du corpus, les nœuds verts les textes de loi et les bleus les références de jurisprudence. La modélisation du graph est dense et complexe à analyser par le simple biais visuel. Outre cela, ce réseau de graphe étant l'agrégation des deux réseaux précédents, on peut voir que des tendances analysées se confirment graphiquement. Dans ce sens, un nœud de loi ressort comme central et plus important que les autres, comme dans le cadre de notre réseau de lois. On observe aussi des concentrations de nœuds dans l'espace qui s'orientent autour du nœud que nous avons relevé. Dans les annexes 23 à 25, on présente les graphiques de distribution des degrés. Dans le même ordre d'idées que précédemment décrit, on remarque une forte concentration des degrés de connexion centrés entre 0 et 150 avec un degré moyen de 3,775 en accord avec notre exploration de données. Concernant les composantes, on dénombre 79 composantes faiblement connectées. Au cours de notre exploration de données, nous

avons relevé 73 textes qui ne comportaient pas de citations. Cela représente 1,35% du corpus qui, pour les raisons explicitées dans nos analyses ultérieures, ne s'appuie pas sur le réseau juridique québécois traditionnel. Ainsi, 6 composantes avec des connexions ne sont pas reliées au réseau. Pour mieux comprendre ces dernières, nous avons pris le temps de les regarder de plus près dans Gephi. Ces composantes sont toutes composées de deux nœuds, un texte de référence du corpus et un texte de citation jurisprudentiel attaché. Aucune de ces composantes ne comprend de lois, ce qui s'explique par la nature très resserrée du corpus de lois généralement utilisé dans des jugements. La jurisprudence, en revanche, est large et comprend tous les jugements valides parus depuis le début de la création de la *common law*. Cependant, on peut réduire les champs de ces citations, car il est communément admis que les jugements plus récents ont plus d'impact dans la plaidoirie. En considérant ces différents éléments, on constate une plus grande profondeur de la jurisprudence, et certaines de ces références peuvent être uniques. Dans notre situation, c'est moins de 0,15% de notre corpus qui s'appuie sur des précédents uniques à ce corpus. On peut donc minimiser l'impact de ces composantes dans le graph et se concentrer sur la composante la plus forte. Ensuite, nous regardons les communautés qui composent le graphe. La modularité est de 0,648, ce qui est une valeur comprise entre celle des citations de jurisprudences et celle des citations de lois. Ce chiffre nous indique une structure de communautés nous indiquant que ce graphe est divisible en ensembles de réseaux densément connectés. On peut donc présumer que le réseau est organisé autour de ces communautés et que nous pouvons hiérarchiser les textes du corpus en fonction des lois et de jurisprudences citées. De plus, le graphe est composé de 105 communautés qui comprennent les 76 composantes faiblement connectées, le graphique 6 représente la distribution des communautés en fonction de leurs tailles.



Graphique 6 : Distribution des communautés de nœuds dans le graphe complet

Le graphique ci-dessus nous montre les communautés qui comprennent les composantes faiblement connectées qui sont les communautés proches de 0 nœud. Aussi, on compte une vingtaine de communautés qui concentre un grand nombre de nœuds avec une communauté au-dessus de 2 800 nœuds qui est probablement celui qui comprend la loi la plus importante du graphe. Le graphe complet étant au cœur de nos analyses, nous avons décidé d'éliminer les composantes faiblement connectées. En enlevant les 79 composantes faibles, nous obtenons un graphe avec une composante. Puis nous analysons la modularité de ce dernier, qui est de 0,651, constituant une légère amélioration de cette dernière. Nous regardons ensuite la distribution des communautés avec après la suppression de ces composantes faibles dans le graphique 7.



Graphique 7 : Distribution des communautés de nœuds dans le graphe complet sans composantes faibles

On distingue alors 21 communautés qui sont celles qui attirent notre attention. Dans nos résultats, nous interprétons ces communautés et nous tâchons de voir si celles-ci ont un lien avec l'analyse sémantique que nous avons faite plutôt.

En conclusion, nos analyses du graphe nous montrent une structure dense avec plusieurs communautés. Le retrait des textes ne comprenant que peu de citations nous confirme une structure de graphe avec 21 communautés qui sont des regroupements autour de certaines lois et jurisprudences, nous indiquant qu'une structure juridique est probablement décelable derrière ces chiffres. De plus, nous devons souligner que ce chiffre est proche de nos 19 sujets dégagés par le LDA, renforçant notre intuition que le lien sémantique et juridique peut être décelé par la combinaison de l'analyse de graphe et la modélisation de sujet. Nous discutons alors nos résultats avec cet élément en tête afin d'orienter la structure du reste de ce mémoire.

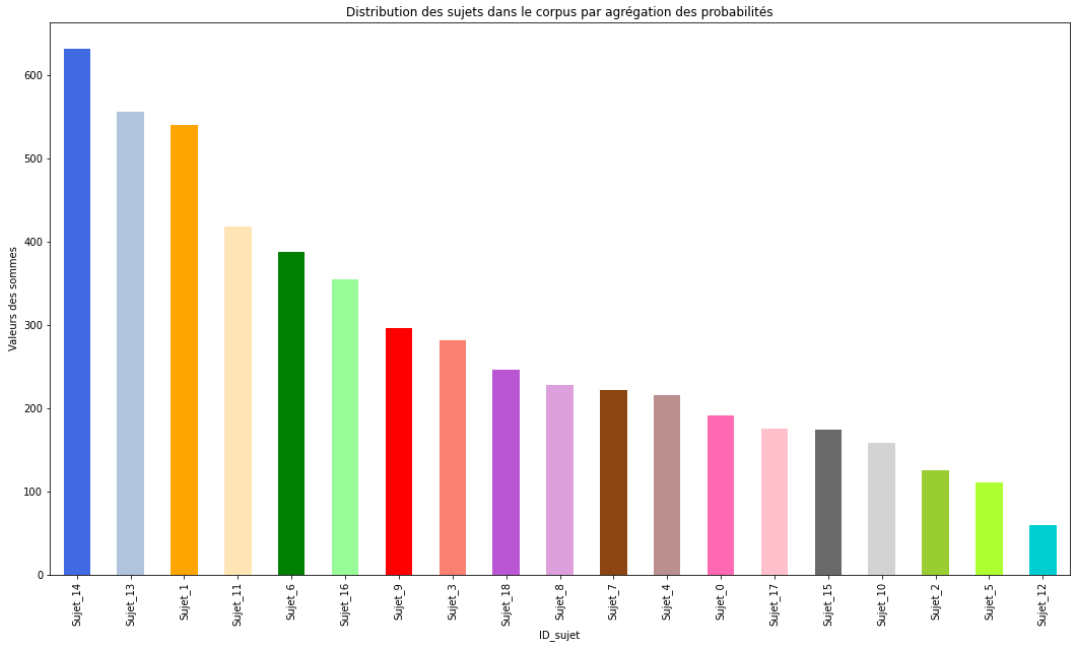
V) Résultats

Jusqu'à présent, nous avons présenté nos données et détaillé les analyses que nous avons sélectionnées pour aborder notre problématique. Dans cette partie, nous procédons à l'étude des résultats obtenus. Pour cela, nous commençons par présenter les résultats de la modélisation de sujet. Dans cette partie, nous discutons de la représentation des sujets dans le corpus, la distribution de ces derniers et les cooccurrences de sujets. En effet, le LDA permet de modéliser plusieurs sujets dans un même texte, nous regardons alors les cooccurrences de sujets afin de voir si nous pouvons dégager des tendances dans la modélisation. Dans un second temps, nous regardons les résultats du graphe. Dans cette section, nous regardons plus en détail les communautés et les nœuds prépondérants du réseau en adéquation avec nos analyses. Nous regardons aussi les mesures de similarité et de distances dans le graphe pour modéliser la similarité entre les textes du corpus. Finalement, nous regardons si des liens potentiels existent entre la modélisation de graphe et le LDA. Cette approche mixte nous permet de déterminer si une corrélation entre la structure sémantique et la structure juridique existe.

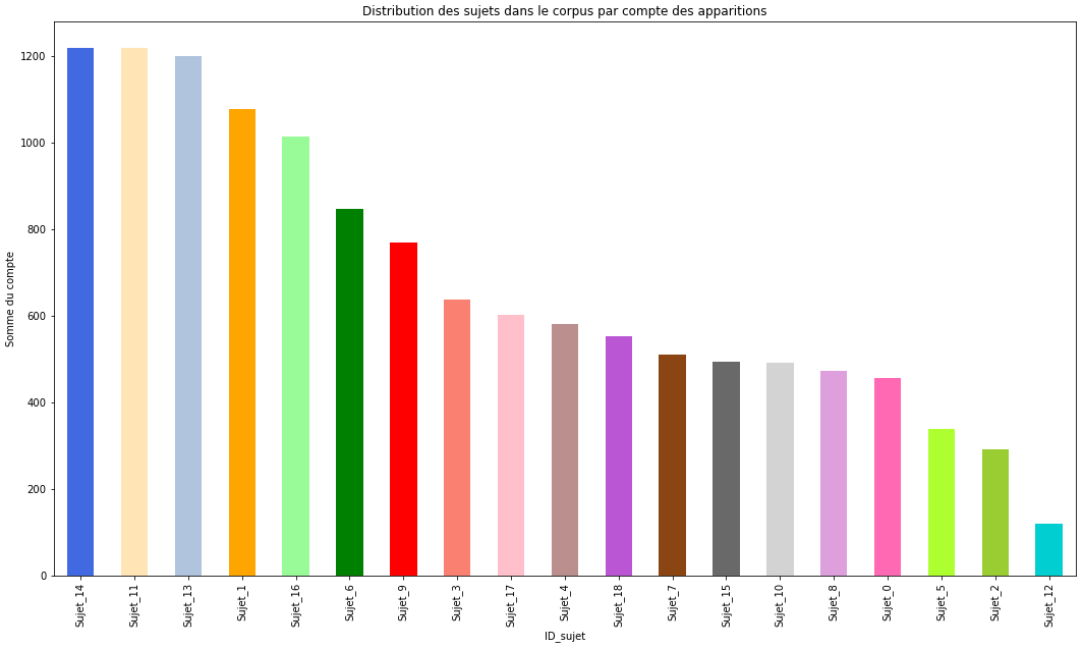
V.1) Modélisation de sujets

Dans nos analyses, nous avons discuté de la sélection du modèle de LDA et particulièrement du nombre de sujets sélectionné pour composer notre analyse sémantique. De cette étude, nous en avons déduit 19 sujets résumés dans le tableau 3. L'objet de l'étude des résultats réside dans la compréhension de la distribution de ces derniers dans le corpus. Le LDA nous donne une distribution sujet/documents qui est interprétable en probabilité d'appartenance d'un document à un sujet. Comme spécifié dans nos analyses, nous avons examiné les 10 textes les plus représentatifs de chaque sujet afin de confirmer que notre modélisation était cohérente. De cette analyse manuelle, nous avons constaté que les sujets ont une cohérence avec les textes qui les représente le plus. À la suite de cela, nous regardons comment les sujets se répartissent dans le corpus. Pour cela, nous avons regardé la somme des probabilités dans le corpus dans le graphique 8.

Puis nous avons regardé le compte d'apparition des sujets dans le corpus dans le graphique 9.



Graphique 8 : Distribution des sujets dans le corpus par agrégation des probabilités



Graphique 9 : Distribution des sujets dans le corpus en fonction de leurs comptes

Pour lire ces graphiques, il est important de consulter le tableau 3 des résultats qui décrit les sujets dans nos analyses. Le sujet le plus fréquent tant en probabilités qu'en quantités

est celui relatif aux délais de contestations. Ce résultat est cohérent, car la discussion des délais de réclamation est un enjeu fréquemment abordé pour attester de la validité de la demande. À l'opposé, on retrouve le sujet sur les maladies respiratoires dans les sujets les moins représentés dans le corpus. Cela demeure cohérent, car cette thématique aborde une pathologie spécifique et qui ne concerne pas une majorité d'industries et de métiers. En effet, on peut raisonnablement inférer que cette pathologie affecte des métiers en lien avec l'amiante ou des environnements composant une exposition forte à des agents aériens pathogènes. Plus généralement, quand on regarde les 5 sujets les plus fréquents et probables dans le corpus, on peut remarquer quelques éléments. On remarque que certains de ces sujets, comme le 11 et le 13, respectivement relatifs aux entorses cervicales et au lien de causalité de la blessure n'ont pas la même position entre le compte et la probabilité. On peut souligner que les cas relatifs aux entorses cervicales sont plus présents dans le corpus, mais les textes qui traitent de ce sujet ont une probabilité d'appartenir à ce sujet globalement plus fort. Cela peut s'expliquer par le fait que les mots qui composent le vecteur du sujet sur la blessure cervicale comportent des mots que l'on peut retrouver dans plusieurs autres vecteurs, comme « entors », par exemple. Cela influence le compte, car ce sujet peut être présent de manière diffuse dans le corpus. Plus généralement, on retrouve ces 6 sujets dans des ordres différents cependant haut dans les deux graphes. Cela s'explique par la nature de ces sujets, que l'on explicite ci-dessous:

- Sujet 1 : prestation de handicap
- Sujet 6 : indemnités de remplacement
- Sujet 11 : entorse cervicale
- Sujet 13 : lien de causalité
- Sujet 14 : délais de contestation
- Sujet 16 : blessures au dos

L'ensemble de ces sujets constitue des pathologies plus communes ou des éléments fréquemment débattus devant le tribunal. Pour ces raisons, il est logique que notre corpus parle plus des blessures aux dos qui peuvent affecter une grande variété de métiers que des maladies respiratoires plus relatives à certains types d'activités. De même, les délais de

contestations, les prestations de handicap ou les indemnités de remplacement sont souvent les éléments qui sont en jeux pour les demandeurs. En effet, le but de nombreuses actions face au TAT vise au dédommagement d'un préjudice subi. Il est donc naturel que cet élément constitue une grande portion des sujets abordés par notre corpus. Pour donner suite à cela, nous regardons les trois sujets les moins fréquents :

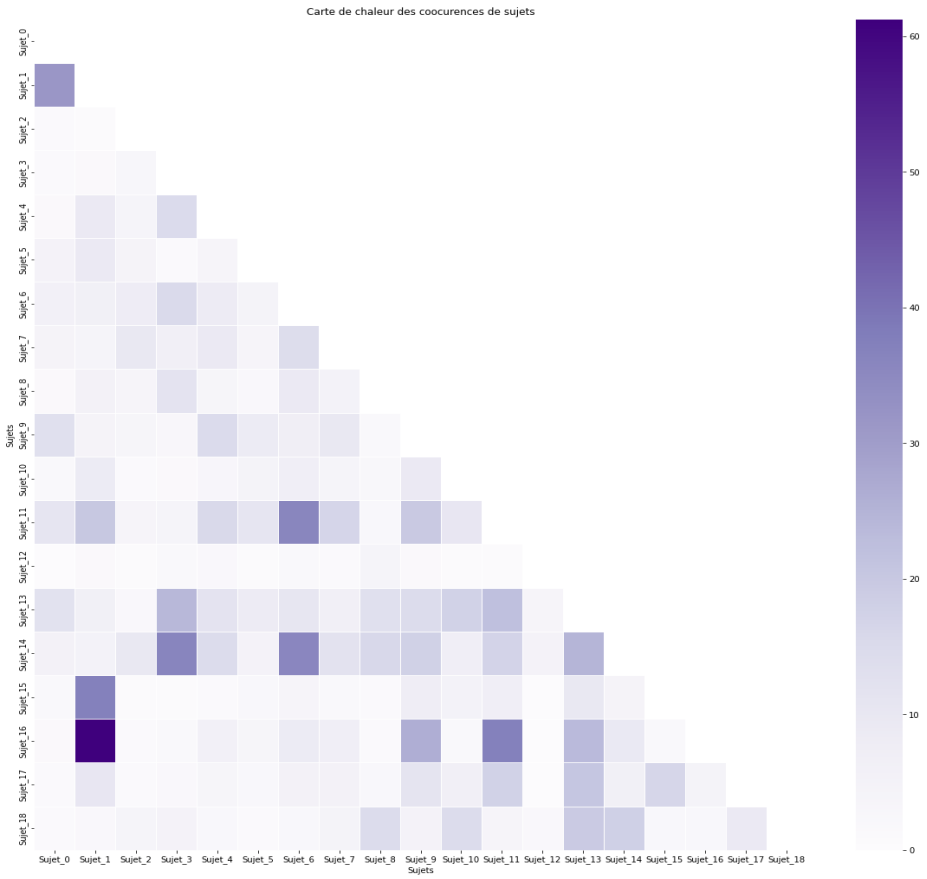
- Sujet 2 : frais de soins
- Sujet 5 : blessures au pieds
- Sujet 12 : maladies respiratoires

Les sujets 5 et 12 traitent de pathologies très spécifiques qui, comme on l'a vu plutôt, ne se retrouvent pas dans les décisions concernant cette niche particulière des accidents du travail. La portion sur les frais de soins est plus surprenante. En effet, nous nous serions attendus à ce que ce sujet soit plus fortement représenté dans le corpus, car constituant un élément crucial dans les démarches au tribunal. En regardant les 10 textes les plus représentatifs du sujet, nous avons une piste de réflexion sur la faible représentation de ce sujet. Dans ces jugements, nous retrouvons systématiquement des travailleurs qui viennent demander un remboursement après une intervention médicale ou une installation de prothèses auditive, par exemple. L'aspect rétroactif de ces demandes en fait des cas plus rares, car il est plus difficile de montrer a posteriori le lien entre la maladie et le travail après intervention médicale. De plus, ces cas traitent souvent de travailleurs agissant plusieurs années après la fin de l'emploi, rendant la démarche encore plus complexe. En revanche, il est plus simple de faire reconnaitre la pathologie par le tribunal en vue de faire prendre en charge via les prestations de handicap ou de remplacement les soins et les préjudices. Bien que nos connaissances juridiques soient limitées sur ce sujet, nous vous proposons cette explication. Il serait cependant judicieux de creuser davantage ce thème avec un spécialiste du domaine. Pour finir notre analyse sur la distribution des sujets dans le corpus, nous regardons de plus près les sujets restants. Si les sujets 9 et 3 relatifs à la rechute et à la représentation sont plutôt souvent représentés, car, encore une fois, ils touchent à des enjeux centraux du tribunal, nous sommes surpris par le changement de position entre les probabilités agrégées et le compte d'autres sujets. Le premier qui mobilise notre intérêt est

le sujet 17 relatif aux blessures aux bras et épaules. En effet, on retrouve ce sujet dans de nombreux textes. Cependant, l'agrégation des probabilités montre que ce dernier n'est pas aussi central dans l'appartenance des textes qui le représentent. Dans ce cas, le sujet est très large, car de nombreux sujets touchent à des éléments relatifs aux bras et épaules. En revanche, peu de textes se concentrent fondamentalement sur une déchirure des coiffes des rotateurs, par exemple, qui est une pathologie précisément reliée aux épaules. Ainsi, comme dans le cas du sujet 6 relatif aux entorses cervicales, ce dernier est diffus dans le corpus, ce qui explique le compte élevé, mais des probabilités plus faibles. Un autre sujet surprenant est le 18 relatif aux maladies auditives. Ce dernier est fort dans le classement probabiliste, mais plus faible dans le compte. Ce cas peut s'expliquer par la présence d'un corpus de mot très précis qui rend uniques les jugements qui y sont rattachés et par le fait les rend fortement reliés au sujet. Ainsi, le faible nombre de textes dans lequel il est présent ont des probabilités fortes d'y appartenir, rendant le sujet fortement représenté en probabilités, mais moins en ce qui concerne le nombre de sujets susceptibles d'y appartenir. On peut aussi souligner le sujet 4 relatif à la condition psychologique, qui est tout de même le 10e sujet le plus présent dans le corpus et le 12e le plus probables. Cela nous indique que les enjeux psychologiques ont tout de même une importance dans les débats du TAT en 2017. Nous aurions pensé voir ce dernier plus bas, nous attendant à ce que les pathologies physiques soient plus représentées dans le corpus. Dans le même ordre d'idées, nous aurions pensé que le sujet 8 relatif au partage des prestations aurait été plus haut, car constituant un élément important des enjeux du tribunal. Il se classe dixième en somme de probabilité, mais quinzième en importance, ce qui signifie que le sujet est crucial pour les textes qui s'y intéressent, mais qu'il est relativement négligé par la plupart des écrits. Nous nous serions plutôt attendus à voir l'inverse dans ce cas, pensant que le partage des prestations était un élément récurant des arbitrages du tribunal. Il semble qu'en réalité, en déterminant les responsabilités et les liens de causalité, les éléments de partages soient déterminés assez naturellement par le tribunal. Les éléments relatifs aux partages sont alors plus rares et précis, ce qui explique son importance moins décisive que nous l'avions imaginé. Les autres sujets sont répartis de manière relativement cohérente dans le corpus

à nos yeux et ne montrent pas d'écarts majeurs entre les deux graphes. Les différences observées peuvent s'expliquer par la nature des sujets. En effet, ils peuvent être très larges, ce qui les rend présents dans de nombreux textes. Cela explique un rang plus élevé dans le compte que pour la somme des probabilités. Pour continuer d'exposer nos résultats, nous proposons de regarder la cooccurrence des sujets afin de déceler de potentielles convergences entre certains d'entre eux.

Afin de modéliser les cooccurrences de sujets dans le corpus, nous avons produit une matrice sujets-sujets se basant sur les probabilités d'apparition des sujets dans chaque texte du corpus. Cette matrice est composée de 361 cases au total, soit 19 colonnes par 19 lignes qui représentent nos sujets. Afin de rendre ces résultats plus lisibles, nous proposons une carte des chaleurs dans le graphique 10. Cette matrice étant symétrique, nous avons décidé de modéliser que le bas de la diagonale de celle-ci.



Graphique 10 : carte des chaleurs de cooccurrence des sujets

Dans cette carte de chaleurs, plus les sujets cooccurrents et plus les tuiles qui les représentent sont foncés. Nous constatons que les sujets 16 et 1 apparaissent très fréquemment ensemble. Il s'agit des thèmes relatifs aux prestations de handicap et aux lésions au dos. Dans le but de comprendre ce fait, nous avons regardé plusieurs documents de la CNESST. Dans son guide des orientations en imputations de 2023 (CNESST, 2022), nous avons trouvé que les blessures au dos étaient une des lésions les plus fréquemment évoquées dans le cadre de l'imputation de prestation de handicap. Le même document mentionne également les entorses cervicales dans les lésions communément sujettes à des prestations d'invalidité, ce qui explique la cooccurrence relativement élevée du sujet 11 et 1. Il en est de même pour la déchirure de la coiffe des rotateurs, qui est illustrée par le sujet 15 et des lésions aux genoux incarnés par le sujet 0, qui sont aussi détaillés dans ce document. Regardons de plus près la ligne du sujet 16, qui présente plusieurs autres points de cooccurrence élevés. Outre le sujet 1, on note le sujet 9 relatif aux rechutes, le sujet 11 relatif aux entorses cervicales et le 13 relatif à la causalité de la blessure. Il est logique de penser que les entorses cervicales peuvent entraîner des lésions dorsales à long terme. Le dos étant impliqué dans divers mouvements, en particulier ceux de portage, les rechutes sont alors probables. Finalement, afin d'obtenir une compensation, le travailleur doit prouver le lien entre la tâche en emploi et la blessure subie. Dans le cas du dos, de nombreuses activités extra-professionnelles peuvent intervenir dans une blessure, on peut donc déduire que l'examen du juge repose sur la capacité de déterminer la causalité entre un événement au travail et la blessure subie par le demandeur. Plus généralement, on remarque que le sujet 13 est cooccurrent avec de nombreux autres thèmes. De même sorte que le sujet 14 relatif aux délais de contestations, il s'agit d'enjeux de procédure qui sont très fréquemment abordés dans le sujet, comme le montrent les graphique 8 et 9. On peut constater que ces derniers sont aussi très cooccurrents l'un avec l'autre, nous indiquant l'importance de ces procédures dans la logique du tribunal. Une autre cooccurrence élevée concerne le sujet 6 et 11. Cela nous indique que les entorses cervicales et les enjeux d'indemnités de remplacement sont fréquemment évoqués simultanément dans plusieurs jugements de notre corpus. Nous n'avons pas trouvé d'explications satisfaisantes pour

justifier ce point dans notre corpus. Cela peut être lié à un biais de nos données, en effet, peut-être qu'un 2017 de nombreuses réclamations de remplacement de revenus ont été faites dans le cas d'entorse cervicale. Il est aussi possible qu'un lien juridique plus complexe soit à l'œuvre qui amène une cooccurrence de ces sujets qu'un spécialiste pourrait possiblement expliquer. Finalement, nous voulons regarder deux sujets de plus près qui portent des surprises. Le sujet 4 relatif à la condition psychologique apparait relativement fréquemment avec la thématique des entorses cervicales. S'il peut y avoir un aspect psychologique à une blessure, nous sommes surpris que le cas des entorses cervicales soit plus important que les autres. Une explication possible est que les entorses cervicales surviennent régulièrement à l'occasion d'accidents de la route qui peuvent traumatiser la victime de ce dernier. L'autre sujet qui comprend des surprises est celui relatif aux maladies auditives. S'il a une forte cooccurrence avec les thèmes relatifs aux délais, aux liens de causalité et à la rechute, nous avons aussi une cooccurrence avec les blessures aux mains et aux épaules. Plusieurs pistes s'ouvrent à nous pour expliquer cela. La première est technique et concerne le nettoyage de nos données. En effet, nous avons souligné que le mot « exposition » compliquait la distinction entre les sujets, tels que ceux liés à la surdité et aux maladies respiratoires. Peut-être alors que des éléments de vocabulaires approchent les sujets relatifs aux blessures aux bras et à la main avec le sujet de la surdité. Une autre explication repose dans la compréhension de la surdité et de la détection de cette dernière. Concernant la surdité, de nombreuses personnes affectées par cette pathologie communiquent via le langage signé donc avec les mains. Cet élément peut avoir un impact dans la cooccurrence des deux sujets, car le mot « main » peut alors apparaitre dans le vocabulaire relatif au langage signé. Relativement à la détection de la surdité, en regardant le manuel de Merck (Hamiter, 2024), nous avons pu comprendre que, pour obtenir un diagnostic de surdité, les professionnels font passer un test aux deux oreilles avec différentes fréquences d'ondes en demandant de lever la main ou le bras au patient. Les sources médicales étant souvent convoqué et cité dans nos textes, ces éléments peuvent avoir un impact dans la cooccurrence des deux sujets. Cette hypothèse est celle que nous

favorisons, car nous avons constaté qu'il était fréquent de la part du tribunal d'utiliser des évaluations médicales pour appuyer la preuve.

Pour conclure sur les résultats de nos analyses sémantiques, nous avons constaté que la modélisation de sujet était efficace pour modéliser les textes de notre corpus. Il est intéressant de voir que les classifications proposées par la CNESST sont présentes dans cette modélisation. De plus, on peut voir que deux grandes familles de sujets sont formées. La première est relative aux éléments de procédures comme le sujet relatif aux délais, aux prestations de handicap ou de remplacement de revenus, et d'autres sont relatifs aux pathologies, comme les maladies auditives, les lésions au dos ou les entorses cervicales. L'étude de l'élément sémantique nous démontre que, par ce biais nous pouvons efficacement et facilement comprendre les sujets abordés par un corpus juridique. Si ce mémoire se concentre sur la tâche d'appareillements de cas, cette modélisation porte de nombreux éléments d'intelligence qui peuvent possiblement améliorer notre compréhension du fonctionnement du TAT.

V.2) Modélisation de réseau de graphe

Dans nos analyses nous avons démontré que le réseau de graphe complet possédait une structure de communauté intéressante avec une modularité de 0,648. En outre, nous avons constaté que des nœuds semblent polariser une grande quantité de connexion dans le graph. Nous avons également discuté de la décomposition du graphe en examinant les sous-graphes des lois et des jurisprudences, ce qui nous a permis de comprendre visuellement la constitution de notre réseau. Le calcul de modularité nous a permis de déterminer des communautés qui peuvent être étudiées pour comprendre leurs structures. Nous proposons de partager les résultats de ces analyses dans un premier temps, ce qui nous permet aussi de mettre de l'avant les nœuds les plus importants du graphe. Puis, nous avons calculé deux mesures de distances, celle de Jaccard et la BFS. Dans un deuxième temps, nous regardons les résultats de ces calculs au sein du corpus afin de déterminer si nous pouvons déceler des proximités dans notre corpus pouvant nous servir à appareiller des jugements les uns avec les autres.

Nous commençons donc par étudier les regroupements de communautés du réseau que nous avons identifiés. Comme évoqué dans nos analyses, on retrouve 21 communautés dans notre graphe une fois les éléments indépendants isolés. Par ce biais, on peut alors identifier 22 communautés, car nous considérons que les 22 textes non connectés sont alors à regrouper sur cette caractéristique. Pour analyser ces regroupements, nous avons utilisé une approche similaire à celle des sujets formés par le LDA. Nous avons examiné dans chaque communauté les lois, les références jurisprudentielles et les textes de nos corpus avec les plus hauts degrés. Dans un souci de constance, nous avons consulté entre 15 et 20 documents par communautés afin de déterminer les enjeux que ces dernières portent. Nous résumons nos résultats dans le tableau 4 ci-dessous.

ID de communauté	NB de noeuds	Thème de la communauté
0	1687	Coûts de matériel pour aménagement du domicile
1	431	Détermination de l'allocation pour l'aide à domicile
2	1750	Preuve, présomption, pour les accidents survenus sur le lieu de travail
3	347	Partage de prestations et imputation des coûts
4	2102	Définition du handicap
5	670	Indemnité de remplacement : détermination des montants
6	33	Validité des montants de frais de déplacement
7	1513	Détermination de la capacité du retour en emploi
8	1439	Délais de la prolongation
9	1624	Droit de la personne, égalité, convenabilité de l'emploi
10	263	Détermination du caractère soudain imprévu occasionnant une lésion
11	114	Question sur l'unité syndicale et la légitimité de représentation
12	616	Relation employeur employé
13	594	définition de la condition psychologique, harcelement
14	827	Preuve de l'aggravation de la maladie par le travail
15	229	Expertise et évaluation médicale
16	335	Responsabilité de la nature accidentelle
17	219	Apparition de nouveau diagnostic
18	1294	preuve de la rechute, récidive, aggravation
19	1171	Procédure d'appel et de révision
20	103	Collaboration des parties notamment en terme de rééducation
21	81	Composantes non connectées

Tableau 4 : tableau résumé des communautés de nœuds dans le réseau

En regardant ce tableau résumé, on peut voir que notre intuition de mesurer les éléments de procédure juridique à l'aide des citations se confirme. En effet, on retrouve dans les intitulés des groupes uniquement des éléments afférents à de la procédure, mais pas d'éléments afférents à des termes médicaux, contrairement aux sujets composés par le LDA. Pour notre analyse, nous regardons la communauté 21, qui est composée de textes sans connexion ou avec de faibles connexions c'est-à-dire deux nœuds. Nous avons expliqué plus haut que ces textes constituent des cas à part qui s'inscrivent dans la volonté de clarification du TAT ou qui ne traitent que d'éléments d'experts. Ensuite, nous regardons la communauté 0, qui est relative à plusieurs interprétations de procédures reliées à la LATMP. C'est la troisième plus grande communauté avec 1 687 nœuds. C'est aussi la plus complexe à analyser et à détailler. Cependant, la plus grosse partie de notre réseau est constituée de la LATMP, et on y trouve exclusivement des articles qui y sont liés. En regardant la liste des références, on retrouve la jurisprudence *société de transport de la communauté urbaine de Montréal C. Racine* qui parle de la manière de déterminer la responsabilité dans une maladie professionnelle si plusieurs employeurs sont impliqués. On peut aussi souligner, *Perfect-Bois inc.* qui évoque la présomption de la maladie et la place de la preuve médicale dans la détermination des faits. Les articles les plus évoqués de la LATMP sont le 204, 145, 328 et 142. Ces articles parlent tant des exigences que le TAT peut soumettre en termes d'évaluation à un travailleur que des mesures d'imputation des coûts. En regardant les textes de notre corpus, on réalise que le lien qui unit ces textes est le débat du fardeau des coûts du matériel installé à domicile pour assister une victime de maladie professionnelle. À ces fins, déterminer la lésion et les capacités d'action du tribunal dans le cadre de la réadaptation et de la réhabilitation devient logique. Des 20 textes du corpus que nous avons consultés, 18 portaient sur les frais relatifs à l'installation de lits mécaniques ou d'appareils auditifs et autres matériels d'assistance. Finalement, on retrouve la jurisprudence Mathieu et *Désourdy-Durenceau ent. Inc.* qui adresse la conception de réadaptation sociale en lien avec le remboursement de l'achat d'un appareil optique, qui est une des références les plus citées de la communauté avec un degré de 9. Ensuite, nous regardons le regroupement 1 qui est relatif aux éléments de détermination de l'allocation

d'aide à domicile contenant 431 nœuds. Ici, on retrouve l'article 165 de la LATMP qui établit le montant de la prestation d'aide à domicile auquel est éligible un travailleur atteint par une lésion. On retrouve aussi le règlement sur les normes et les barèmes de l'aide personnelle à domicile, qui édicte le cadre spécifique de l'application de l'article 165 de la LATMP. À cela s'ajoute le cas *P.E. Boisvert auto ltée.* qui propose une interprétation de l'article 165 de la LATMP. L'étude des textes du corpus qui compose cette communauté nous révèle que ceux-ci mettent l'emphase sur le montant, la légitimité et la nécessité d'aide à domicile d'une victime de lésion professionnelle. La communauté 2 est la deuxième plus grande et comporte 1 750 nœuds. Elle a trait aux éléments de preuve, présomption pour les accidents survenus en milieux de travail. Ici, on retrouve l'article 28 de la LATMP, qui est l'article qui postule qu'une blessure arrivant sur le lieu de travail est frappée d'une présomption de lésion professionnelle. Cela implique que le fardeau de la preuve repose sur l'employeur, c'est à lui de prouver que ce n'est pas une lésion professionnelle. La jurisprudence la plus citée est *Chaput c. Société de transport de la Communauté urbaine de Montréal* qui parle de la norme de santé sur laquelle la cour doit se baser pour juger la condition physique de la personne en cause. Outre cela, l'article 224 de la LATMP, avec un niveau de 517, traite de la valeur du diagnostic pour la prise de décision du tribunal. En effet, les textes du corpus qui composent ce sujet débattent de preuves et des éléments d'évaluation de la validité de la lésion. La 3e communauté regroupe des nœuds qui sont relatifs au partage des prestations et à l'imputation des coûts. On retrouve ici l'article 31 de la LATMP relatif à la définition de la lésion professionnelle ainsi que les articles 327 et 73 qui abordent l'imputation des coûts aux employeurs et la détermination du revenu brut d'un travailleur en vue de fixer le montant de l'indemnité de remplacement de ce dernier. On retrouve plusieurs textes du corpus et jurisprudences qui abordent la contestation de l'imputation des coûts reliés à ces enjeux. La quatrième communauté est la plus grande et est composée de 2 102 nœuds. Cela s'explique par la thématique que balaient les documents qui la composent. Ici, on retrouve des jugements comme *Municipalité Petite-Rivière-St-François et C.S.S.T.* qui définissent ce qu'est un travailleur déjà handicapé avant lésion et le rôle que joue cet élément. On retrouve aussi l'article 329 de la LATMP qui intègre cette notion de handicap

dans la loi. Les textes du corpus que nous avons regardés comportent cette notion de handicap de manière centrale. La 5e communauté relative à la détermination des montants pour l'indemnité de remplacement se caractérise par le nœud représentant l'article 67 de la LATMP relatif au détail de la détermination du revenu brut du travailleur à considérer pour cette indemnité. Ce qui la différencie de la communauté 3 est les références à la loi sur les impôts, notamment les articles 1019.4 et 42.11 relatifs aux pourboires. Par ce fait, on peut déduire que ce regroupement est fortement relié à la détermination du revenu en vue de fixer le montant spécifiquement de cette indemnité. La sixième communauté elle est la plus petite. Elle regroupe les documents parlant de la validité des indemnisations des montants pour les déplacements. Ici on parle de taxis ou de véhicules utilisés pour pallier le manque de mobilité qu'entraine une lésion. On retrouve que trois textes du corpus, mais de nombreuses références pour appuyer ces cas. Le 7e groupe est plus imposant, car composé de 1 513 nœuds. Il regroupe les textes qui parlent de la capacité de retour en emploi. Dans ce cas, la difficulté repose sur la grande diversité de lois et de références qui sont convoqués. Mais on peut retrouver les articles 44, 46 et 48 de la LATMP qui portent tous une notion de retour en emploi. Cela, cumulé avec les jurisprudences et la lecture du corpus, nous a fait comprendre que ces éléments font toujours référence à la capacité de l'employeur à retourner en emploi à des fins de maintien ou de suppression des montants qu'il touche. Le 8e regroupement est relatif aux délais de prolongation et de contestation, ici on relève l'article 15 de la LITAT qui donne le cadre des prolongations de délais de contestations pour le TAT. On peut aussi souligner les articles 358, 359 et suivant de la LATMP qui sont concentrés sur les délais de contestation. Finalement, la jurisprudence *Hôpital Maisonneuve-Rosemont* est la plus sollicitée dans cette communauté et établit de manière pratique ce qu'est un motif raisonnable de prolongation. La neuvième communauté est la troisième plus grosse avec 1 624 nœuds. Elle parle des droits de la personne de l'égalité en milieu de travail pour une personne affectée par une lésion et du caractère convenable de l'emploi. Dans ce dernier, on peut lire l'article 47.2 du Code du travail, qui évoque la bonne foi, la négligence et l'arbitraire dans le milieu de travail. On peut y adjoindre l'article 32 de la LATMP qui encadre le comportement de l'employeur vis-à-vis de la

personne blessée, notamment en ce qui concerne le congédiement ou le changement de poste du travailleur. On retrouve aussi l'article 2858 du Code civil qui parle des droits et de devoir dans le respect des droits de la personne en vue de l'obtention de preuves. De plus, on peut souligner les références à des jugements de la Cour suprême, comme le [1984] 1 RCS 509 qui touche à la digne représentation du travailleur par son syndicat. Plus généralement, de nombreuses références sont issues de la Cour suprême ou de la cour supérieure, qui ont généralement l'expertise sur les enjeux de droit à la personne. La communauté 10 comporte 263 nœuds et est relative à la détermination du caractère soudain d'un événement occasionnant une lésion. La référence la plus citée est Plomberie et *chauffage Plombec inc. et Deslongchamps* qui clarifie la notion d'accident à l'occasion du travail. Aucune loi ne se retrouve dans ce regroupement, mais de nombreux textes du corpus, dont les 15 plus connectés, portent tous cette composante en leurs seins. La 11e communauté est relative à la légitimité de la représentation syndicale. Ici on retrouve seulement 114 nœuds et 9 textes du corpus. Cette communauté se différencie de la 9e par le fait qu'on ne parle pas de la notion de représentation juste ici, mais plutôt de savoir si l'employé peut être représenté par le syndicat, s'il appartient à la bonne unité de travail. La communauté 12 est relative au lien employeur employé. On retrouve dans la référence de nombreux articles du Code civil du Québec qui traite du contrat. De plus, on retrouve des jugements comme Les *Rôtisseries St-Hubert* qui parle de la conciliation travailleur employeur. Les textes du corpus reliés ici sont majoritairement relatifs à la relation de conciliation et quelques-uns traitent du statut de travailleur autonome. Le 13e regroupement est relatif à la définition du cadre pour les maladies professionnelles de type psychologique. On retrouve le jugement *Welch et Groupe pharmaceutique Bristol Myers* qui met en place le cadre d'interprétation d'événements constituant un caractère traumatique psychique. Globalement, on retrouve beaucoup de jurisprudence liée à la preuve du harcèlement moral dans le but de se qualifier comme victime de maladie professionnelle. On ne retrouve pas de lois spécifiques dans cette communauté, mais les textes du corpus qui y sont reliés portent toute la notion de maladie psychologique ou de traumatisme psychique. La 14e communauté est reliée à l'aggravation d'une maladie liée à l'activité

professionnelle. L'interprétation de ce regroupement a été plus complexe, car les références majoritaires sont reliées à l'audition. Cependant, les textes du corpus de cette communauté ne parlent pas tous de problèmes à l'oreille. Cependant, ces citations sont convoquées pour établir si l'aggravation d'une condition peut être reliée à un travail spécifique. Ensuite, le 15e regroupement affaires aux cadres juridiques de l'évaluation médicale. On retrouve des lois comme la loi sur les laboratoires médicaux, article 1, qui définit des termes médicaux aux yeux de la loi, mais on retrouve aussi de nombreuses jurisprudences traitant du cadre dans lequel l'évaluation est admissible par le tribunal. La communauté 16 est relative aux responsabilités de la nature accidentelle ayant entrainé une lésion. Ici, on retrouve des jugements qui portent sur la responsabilité. Par exemple, le jugement *Maçonnerie Francis Tousignant inc*. l'employeur met en cause la responsabilité du travailleur qui a omis de mettre un cran de sécurité sur un levier hydraulique. La communauté 17 traite des éléments relatifs à de nouveaux diagnostics. Ici, les documents traitent de l'arrivée d'un nouveau fait au moyen d'un diagnostic et par le fait même de la révision d'une décision antérieure. La communauté 18 est composée de 1 294 nœuds, ce qui la classe dans les plus grandes. Par le fait, elle traite de la rechute, récidive et aggravation. Ici, les textes traitent du cadre juridique qui permet de déterminer une rechute, une récidive ou une aggravation, en s'appuyant sur l'analyse des événements possiblement connectés à cette situation. Juste derrière sur le plan de la taille, on retrouve la communauté 19, qui est relative aux procédures d'appels et de révision. L'article 15 de la LITAT est prépondérant avec un degré de 231. Cet article est relatif au temps du délai d'appels pour un demandeur. On relève aussi l'article 49 et 51 tous deux traitant des éléments de procédure d'appel. On retrouve aussi l'article 429.56 de la LATMP relatif à la révocation d'une décision. Le jugement *Franchellini et Sousa* avec un degré de 180, est centré sur l'interprétation d'un appel en vice de fond. On retrouve cet élément dans les 17 textes que nous avons analysés pour cette catégorie. Finalement, la dernière communauté, la 20e, est la deuxième plus petite avec 103 nœuds. On y retrouve 16 textes du corpus, que nous avons lus, qui traitent de la collaboration des parties en matière de rééducation. Ici, le sujet est la bonne foi et l'application de la part du travailleur ou de l'employeur du programme de réadaptation en vue de réintégrer l'emploi.

Afin de proposer une modélisation plus lisible, nous avons réduit le graph en n'affichant que les nœuds avec un degré supérieur à 70. Nous proposons dans la figure 7 la modélisation de ce dernier. La légende propose une couleur par classe de modularité du réseau ainsi que la part que représente cette classe dans les données en pourcentage. Cette figure nous confirme les éléments décrits plus haut et nous permet d'explorer les nœuds les plus importants dans le réseau de graphe ainsi que d'identifier les principales communautés visuellement.

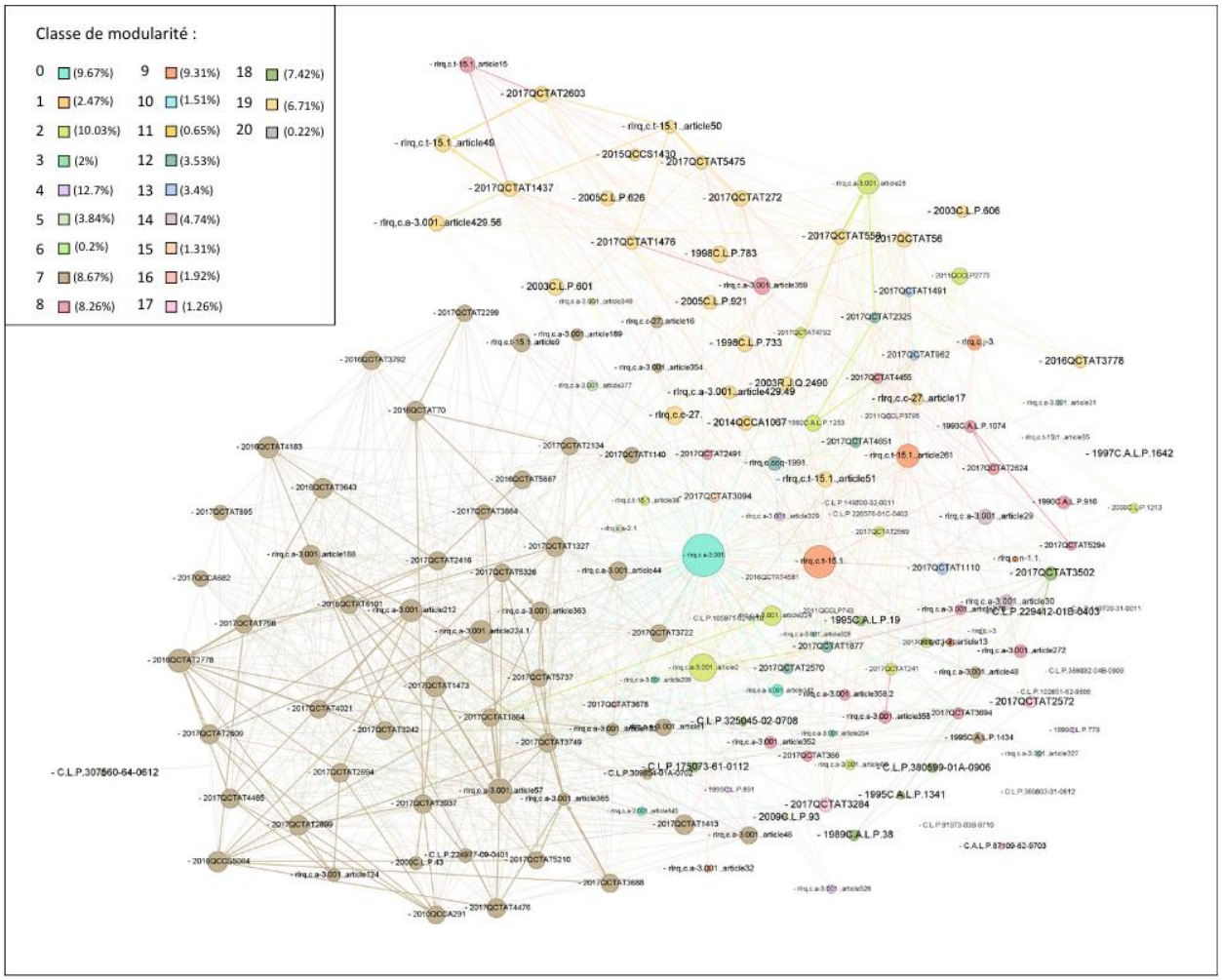


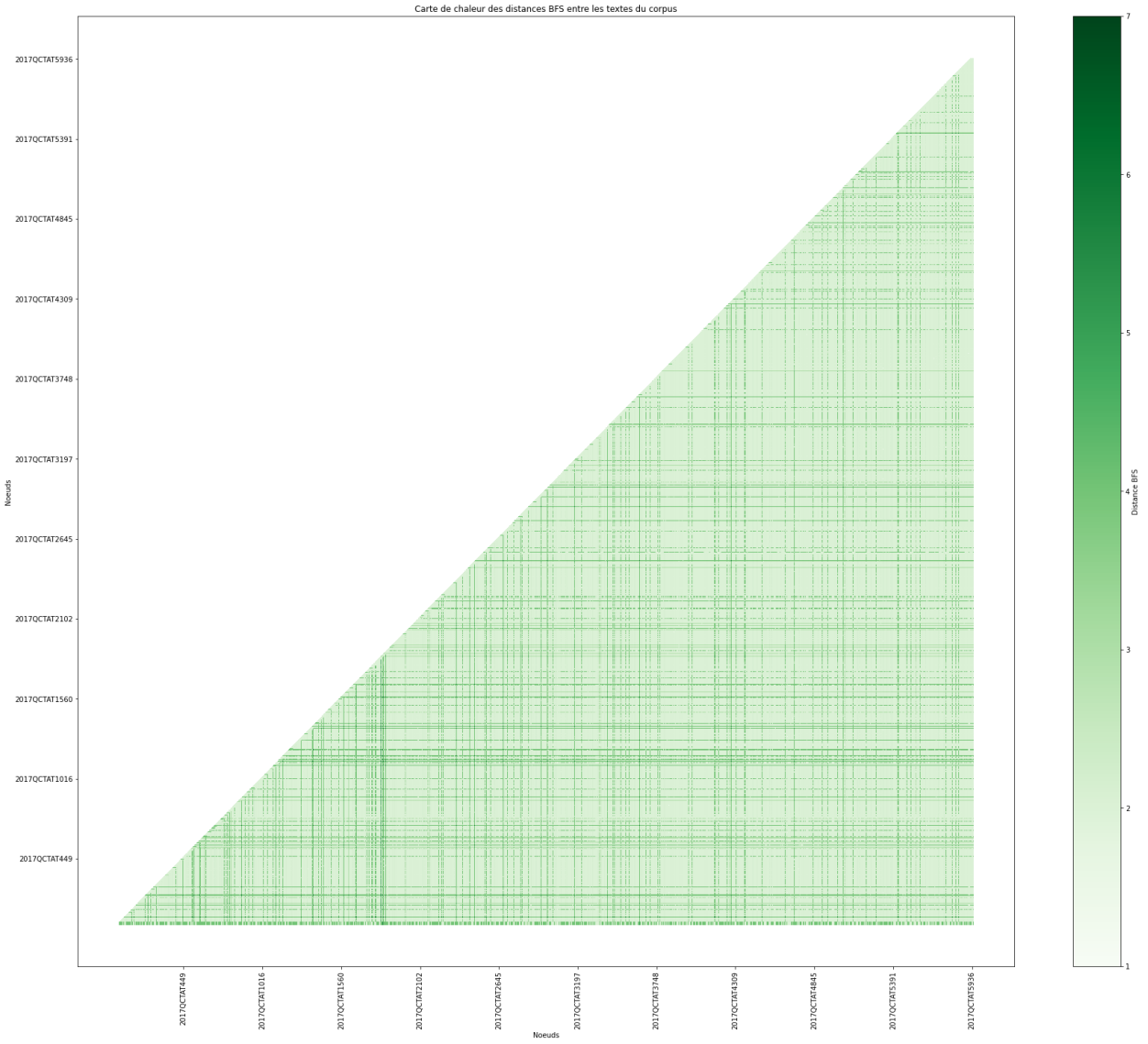
Figure 7 : Graph réduit sans les nœuds avec un degré inférieur à 70

L'analyse extensive des communautés que nous venons de proposer est essentielle, car elle nous révèle que l'analyse de ces dernières reposes sur la comparaison des différents documents qui les composent. De plus, on peut voir que, certaines fois les lois ne sont pas

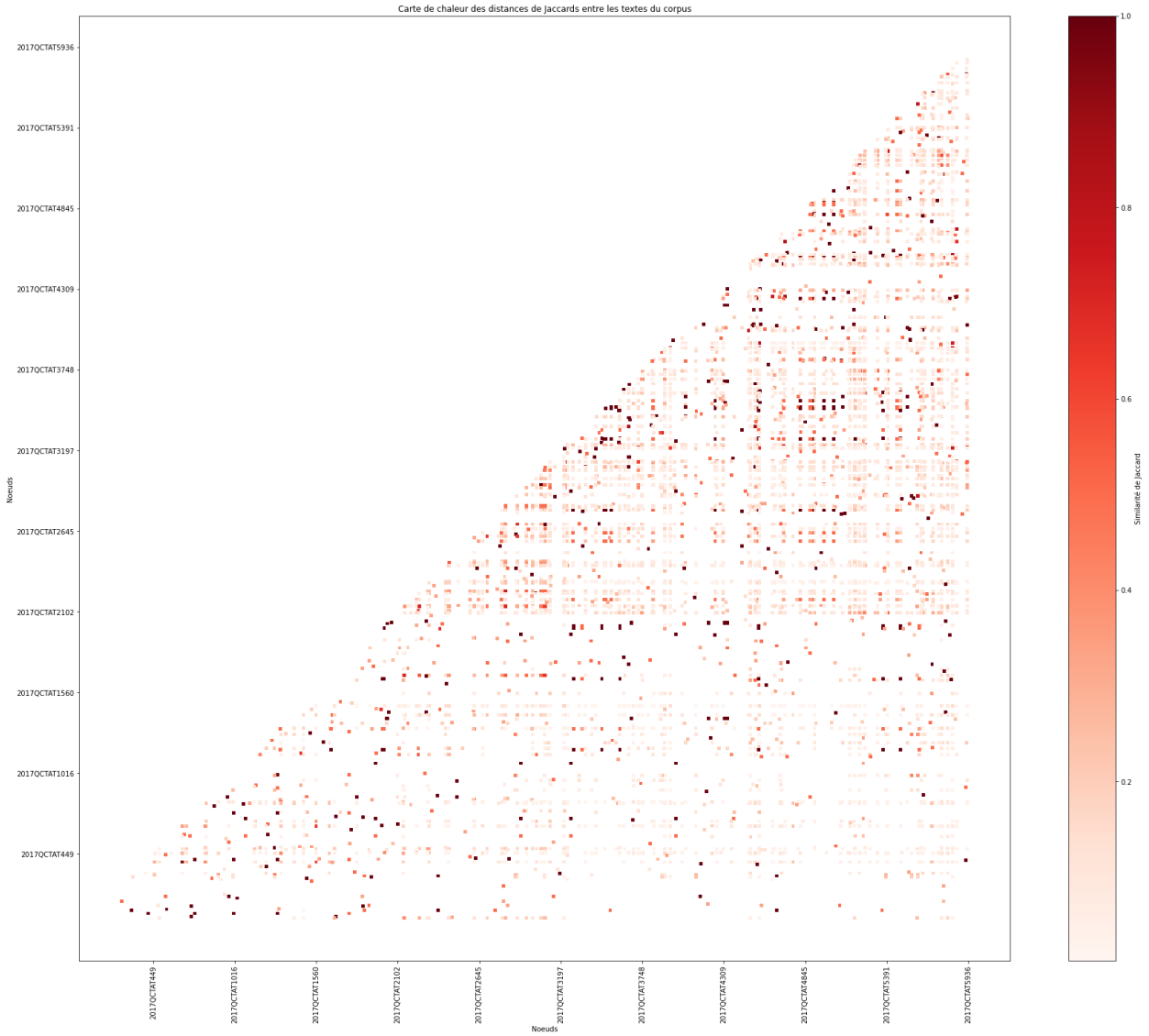
connectées directement aux communautés, mais plutôt des jurisprudences. Cela nous révèle que certains sujets juridiques s'appuient sur la subtilité de l'interprétation du cadre juridique et non sur la simple application des règles de droit. Finalement, on constate que le réseau nous indique des éléments juridiques que le LDA n'a pas soulignés. Nous indiquons donc que l'analyse de ces éléments a été faite à l'aide de nos connaissances du droit et de notre compréhension du corpus, qui nous permet tout de même de révéler une intelligence de nos données. À des fins de précisions, nous aimerions solliciter l'expertise d'un avocat ou d'un juge pour préciser l'analyse de nos résultats. Outre cela, nous avons concentré nos évaluations sur les nœuds les plus importants de chaque regroupement pour des raisons de moyens et de temps. Or, nous pourrions regarder un échantillon plus dispersé du regroupement afin de déterminer si des sous-catégories peuvent composer ces derniers.

Afin de compléter notre partie sur la formalisation de la tâche juridique, nous avons proposé une étude de la similarité à l'aide de la distance et la similarité de Jaccard et la distance BFS. La taille du graphe étant imposante, nous regardons les textes du corpus entre eux pour regarder la proximité au sein du corpus. Dans ce but, nous proposons de regarder la carte des chaleurs pour la similarité de Jaccard et pour la distance BFS. Puis, dans chaque cas, nous regardons quelques éléments de statistiques agrégées relatives à ces deux métriques.

Nous présentons les deux cartes de chaleurs dans les graphiques 11 et 12 ci-dessous. Il est important de noter que nos outils de visualisation se basent sur la similarité et non la distance de Jaccard, car cette dernière est plus facilement interprétable. En réalité, la distance de Jaccard se lit inversement à la similarité, plus elle est grande et moins les textes sont proches. Par le fait, les deux calculs sont des interprétations similaires de la même métrique. De plus, les deux cartes de chaleurs sont des outils pour avoir une vision globale du corpus. Cependant, la grande quantité de variables composant les matrices, soit 5 403 par 5 403, rend les matrices complexes à interpréter.



Graphique 11 : Carte de chaleurs des distances BFS dans le corpus



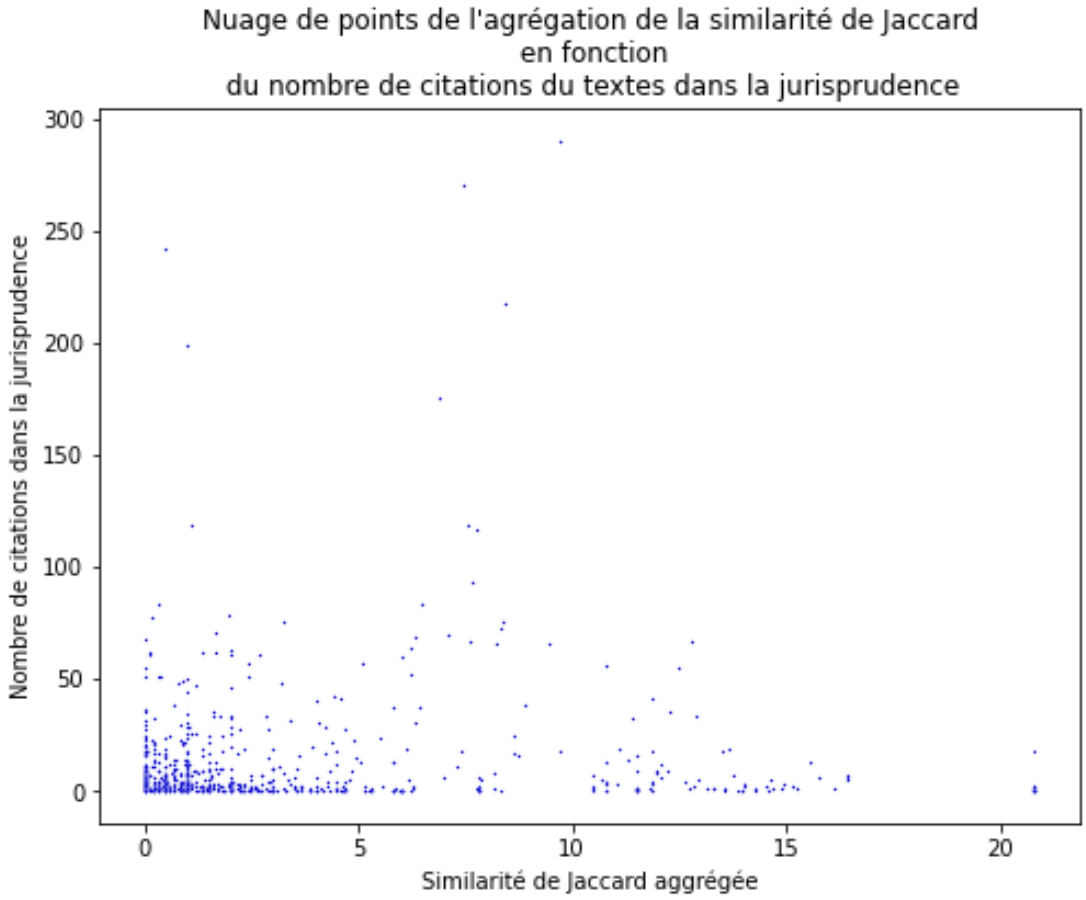
Graphique 12 : Carte de chaleurs des distances de Jaccard pour les textes du corpus

Pour la carte des chaleurs de la distance entre les nœuds, la teinte foncée témoigne d'un écartement des nœuds les uns avec les autres. À l'inverse, les teintes claires révèlent une similarité éloignée, tandis que les teintes sombres révèlent une similarité plus grande. Finalement, les deux graphiques étant symétriques, nous proposons de montrer une seule partie de la diagonale. Plusieurs remarques émergent à la vue de ces deux graphiques. On souligne la nature clairsemée de la matrice pour la similarité de Jaccard. Cela implique que de nombreux textes ne sont pas similaires les uns aux autres. Cela s'explique par la grande dimensionnalité des références et le nombre de citations que comportent les textes du corpus. En effet, nous avions calculé une moyenne d'un peu plus de 12 citations par textes

et nous avons compté 12 043 nœuds de références de lois et jurisprudences. Ainsi, la grande dimensionnalité explique le grand nombre de points nuls de la matrice. Malgré cela, on semble discerné des lignes plus foncées dans la carte que les autres. On peut donc voir que certains textes semblent avoir une forme de centralité dans leurs similarités. En regardant le graphique 11, on remarque que des lignes émergent dans les distances entre les textes. Ces lignes foncées reflètent des textes déconnectés du graphe et/ou faiblement connectés, comme notre analyse du graphe le révèle. Ici, on souligne une proximité globale des nœuds liés à la centralité de la LATMP dans le graphe. En effet, nous avons montré qu'une grande partie des textes s'appuient sur cette loi entrainant une forte centralité de cette dernière dans le corpus. Cela nous amène au constat que la distance ne nous permet pas forcément de déterminer plus d'éléments que celle que l'analyse des communautés nous apporte. Ce que cette carte révèle, c'est une modélisation des communautés du graphe, calculées par Gephi avec l'algorithme DFS, qui calcule également les distances entre les nœuds. Pour la suite de la présentation de nos résultats, nous regardons de plus près les statistiques relatives à la similarité de Jaccard qui portent plus d'information pour l'étude de notre corpus. Pour approfondir nos analyses, nous regardons des éléments relatifs à l'importance de la similarité de Jaccard pour déterminer l'importance du jugement dans la jurisprudence. Pour cela, nous avons cherché sur CANLII, un site de référence juridique, le nombre de citations des textes de notre corpus dans l'ensemble de leurs bases de données jusqu'à aujourd'hui. Devant renseigner à la main ces informations, nous avons pris un échantillon de 1 392, soit un échantillon de 25% du corpus texte du corpus auquel nous avons ajouté cette dimension. Puis nous comparons cela à la somme de la similarité de chaque texte de la matrice que nous avons créée en retirant 1, car chaque texte a une similarité de 1 avec eux même. Ici, nous voulons voir si une forte similarité de Jaccard agrégée dans le corpus est liée au nombre de fois où ce texte est cité dans la jurisprudence actuelle. Pour vérifier, cela, nous propose une corrélation linéaire avec en variable cible le nombre de citations dans la jurisprudence et en variable explicative, soit :

$$Y = a + ß_1 X_1 + \in_i$$

Ici Y est l'agrégation des scores de Jaccard, a est l'ordonnée à l'origine, $ß_1$ est le coefficient de corrélation de la variable X_1 qui est le compte du nombre de citations dans la jurisprudence et $\in_i$ représente les résidus. On commence par montrer le nuage de point de ces deux variables dans le graphique 13.



Graphique 13 : Nuage de points de l'agrégation de la similarité de Jaccard en fonction du nombre de citations du texte dans la jurisprudence

Ce nuage de points nous indique plusieurs choses concernant nos données. La première est le déséquilibre de l'échantillon qui comporte de nombreux points avec une similarité agrégée de 0, soit 698 points de l'échantillon, donc un peu plus de 50% de ce dernier. Cela est logique, car nous avons vu que le corpus comprenait de nombreux textes avec un score de Jaccard faible. Nous avons l'impression d'avoir identifié une corrélation entre l'augmentation du nombre de citations et l'augmentation du score de Jaccard. Cette tendance est difficile à détecter à l'œil. Nous regardons alors les éléments de corrélation de

ces deux variables. Notre corrélation nous donne les résultats que nous présentons dans la figure 7 ci-dessous.

OLS Regression Results

Dep. Variable:	citations	R-squared:	0.030
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	43.07
Date:	Wed, 26 Mar 2025	Prob (F-statistic):	7.44e-11
Time:	15:06:21	Log-Likelihood:	-6137.1
No. Observations:	1392	AIC:	1.228e+04
Df Residuals:	1390	BIC:	1.229e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.8303	0.595	8.118	0.000	3.663	5.998
aggreg_jacc	0.9609	0.146	6.563	0.000	0.674	1.248

Omnibus:	1899.647	Durbin-Watson:	1.492
Prob(Omnibus):	0.000	Jarque-Bera (JB):	388387.921
Skew:	7.604	Prob(JB):	0.00
Kurtosis:	83.406	Cond. No.	4.59

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 8 : tableau résumé des statistiques de la corrélation Citations/agrégation de similarité de Jaccard

Si on regarde la statistique F de 43,07 et la probabilité de la statistique F de 7,44 e-11, on peut en déduire que la variance que le modèle explique est plus grande que celle qu'il n'explique pas. On peut donc déduire que ce dernier est globalement significatif. En revanche, le r-carré est de 0,030, ce qui veut dire que seulement 3% de la variation du nombre de citations peut s'expliquer par l'effet de la similarité de Jaccard dans le corpus. Les coefficients de corrélation pour l'ordonnée à l'origine et l'agrégation de Jaccard sont tous deux significatifs, en témoigne leurs valeurs P et leurs erreurs standards faibles. Ce modèle nous indique que, pour une similarité de Jaccard agrégée de 0, on retrouve en moyenne 4,83 références dans le corpus complet du droit canadien. Puis on peut dire qu'une augmentation d'une unité de la similarité agrégée de Jaccard entraine une augmentation de 0,9609 le nombre de citations. Cependant, le score de Durbin-Watson de 1,492 nous indique une possible autocorrélation positive potentielle des erreurs du modèle. Le faible r carré du modèle et le score de Durbin-Watson ne nous permettent pas de retenir ce modèle comme modélisation complète pour expliquer la variabilité du nombre de citations d'un texte dans le droit canadien. Or, on peut tout de même voir une tendance nous indiquant qu'une similarité de Jaccard agrégée forte d'un texte du corpus de 2017 avec les autres

augmente sa probabilité d'être une référence dans le droit futur. Cette approche si imparfaite nous donne une base pour détecter de potentiels jugements importants d'une année en se basant sur sa similarité avec les autres de la même année. Nous pourrions considérer un modèle prédictif plus précis dans le futur afin d'identifier les jugements majeurs d'une année qui pourrait faciliter le travail qualitatif des analystes juridiques.

En conclusion, nous avons montré que des communautés dans le réseau existent et qu'elles correspondent à des éléments de procédure en lien avec les références de lois et de jugement sur lesquelles elles s'appuient. De plus, si la distance dans le graphe ne permet pas d'apporter plus d'information pour nos analyses, la similarité de Jaccard nous permet d'appareiller les textes du corpus les uns avec les autres. Nous avons aussi proposé de voir s'il existait un lien entre un texte avec une forte similarité dans le corpus et une potentielle importance dans le nombre de citations actuelles de ce dernier. Si ce modèle ne permet que de voir une tendance, car il n'explique que 3% de la variabilité du nombre de citations, cela nous donne une base de réflexion pour élaborer des modèles plus précis pour identifier les jugements centraux d'une année de jugement en intégrant d'autres variables au modèle. L'approche de graphe nous indique que des informations sous-jacentes peuvent être détectées autrement que via la simple analyse sémantique nous permettant de compléter cette dernière. Cela nous amène à regarder si un lien entre les deux approches existe qui pourrait nous amener de nouveaux éléments d'analyse.

V.3) Approche mixte

Jusqu'à présent, nous avons proposé deux approches distinctes pour modéliser notre corpus de textes. Nous pourrions nous appuyer sur les résultats du LDA pour ensuite les localiser dans le graphe et trouver les lois et les jurisprudences utiles pour un cas similaire sémantiquement au nôtre. Or, dans cette partie de notre travail, nous voulons voir si ces deux dimensions peuvent converger. En effet, si nous avons pu voir que les mots peuvent décrire plus précisément certaines maladies et certains éléments de procédures, notre modèle de graphe est principalement axé sur des concepts juridiques. Cependant, il semble évident que le sens des mots contribue à définir des enjeux de procédure et que, par

le fait un lien entre le LDA et la modélisation de réseau de graphe est probable. Dans le but de vérifier cela, nous utilisons la régression logistique multinomiale qui nous permet de voir si, en nous appuyant sur la probabilité d'appartenir à un sujet, il est possible de prédire la communauté à laquelle appartient un texte du corpus.

Cet objectif en tête, nous proposons d'analyser plusieurs modèles de régression qui s'appuient sur un échantillon d'entrainement de 70% de nos données ainsi que sur un échantillon de test et de validation respectivement composés de 15% des données. Ces échantillons sont composés aléatoirement en fixant l'aléa à 42 pour assurer le caractère reproductible de nos résultats. Outre les éléments de prédiction du modèle, nous regardons aussi un modèle de régression multinomiale statistique afin de regarder les valeurs P des coefficients, le pseudo r carré et les indicateurs de vraisemblance. Dans ce sens, nous proposons trois schémas de données différents. Un premier est constitué des données du LDA auxquels est adjoint le numéro de communauté d'appartenance du texte de référence sans aucune autre modification. Ce modèle nous sert de référence de comparaison pour voir si des améliorations sont apportées à ceux subséquents. Ensuite, nous proposons un modèle qui comporte des agrégations de classe de composante basée sur les erreurs de prédiction du premier ainsi que sur leurs proximités déduites par la classification proposée dans le tableau 4 de la partie sur la modélisation de graphe. Finalement, nous regardons un dernier modèle qui ajoute à ces modifications un processus de sélection de variables basé sur la colinéarité de certaines d'entre elles.

Le premier modèle prend les 21 communautés comme variable cible et les 19 sujets comme variables explicatives. Nous pouvons résumer ce modèle à l'aide de l'équation suivante :

$$P(y = C_k|X) = \frac{\exp\left(ß_{k_0} + ß_{k_1 x_1} + ß_{k_2 x_2} + \cdots + ß_{k_{19} x_{19}}\right)}{\sum_{j=1}^{21} \exp\left(ß_{j_0} + ß_{j_1 x_1} + ß_{j_2 x_2} + \cdots + ß_{j_{19} x_{19}}\right)}$$

Ou y peut appartenir aux 21 classes de modalités, $ß_{kj}$ sont les coefficients de corrélation pour la classe C_k et la variable x_j. Pour ce modèle, nous avons pris les données du LDA,

auxquelles nous avons ajouté le numéro de communauté d'appartenance du texte de notre corpus dans le graphe, sans aucune modification. Nous ne proposons pas de montrer les résultats des coefficients de ces corrélations détaillés pour les 21 classes et 19 variables, mais les figures 8 et 9 nous montrent les performances du modèle en termes de prédiction et de viabilité de la régression. Ces deux figures nous montrent que 50% de la variabilité de la classe de modularité dans le graphe est expliqué par la probabilité d'appartenance à un sujet. De plus, la figure 8 nous indique qu'on arrive à prédire correctement 58% de notre échantillon de test et 55% avec notre échantillon de validation. L'écart entre les deux chiffres peut s'expliquer par le déséquilibre des classes dans l'échantillon qui rend sensible la régression à certaines classes de modularité. Nous notons tout de même que le modèle fonctionne à moindre mesure, car, si nous faisions des prévisions aléatoirement, nous aurions 1/21 chance de choisir la bonne catégorie, soit une probabilité de 4,76% de bonne prédiction. À présent, si l'on regarde le modèle statistique présenté dans la figure 9, nous constatons que le LLR p-value à une valeur nulle. Cela nous indique que nous pouvons rejeter l'hypothèse nulle qui postule qu'il n'existe aucun lien entre la modélisation de sujets et l'appartenance à une classe de modularité. Malgré cela, nous avons regardé les valeurs P pour chacun des coefficients et nous n'en trouvons aucune inférieure à 0,05 comme montré en annexe 29. Cela ne nous indique qu'aucun de ces coefficients ne sont significatifs individuellement pour ce modèle.

Rapport de classification pour l'échantillon de test

Précision_Tst:0.5881627620221949

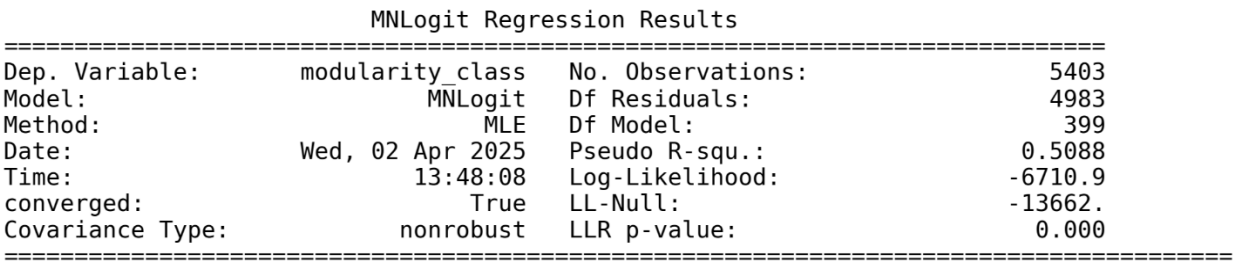
	precision	recall	f1-score	support
0	0.29	0.35	0.32	93
1	0.00	0.00	0.00	13
2	0.66	0.75	0.70	123
3	0.00	0.00	0.00	8
4	0.86	0.97	0.91	130
5	1.00	0.07	0.12	15
7	0.45	0.85	0.58	72
8	0.52	0.72	0.60	82
9	0.36	0.21	0.27	61
10	0.00	0.00	0.00	8
11	0.00	0.00	0.00	1
12	0.00	0.00	0.00	17
13	0.50	0.08	0.14	12
14	0.75	0.75	0.75	28
15	0.00	0.00	0.00	3
16	0.75	0.67	0.71	9
17	0.00	0.00	0.00	8
18	0.74	0.89	0.81	72
19	0.00	0.00	0.00	39
20	0.00	0.00	0.00	3
21	0.00	0.00	0.00	14
accuracy			0.59	811
macro avg	0.33	0.30	0.28	811
weighted avg	0.52	0.59	0.53	811

Rapport de classification pour l'échantillon de validation

Précision_Val:0.554320987654321

	precision	recall	f1-score	support
0	0.32	0.36	0.34	111
1	0.00	0.00	0.00	18
2	0.65	0.81	0.72	104
3	0.00	0.00	0.00	12
4	0.81	0.95	0.88	119
5	0.00	0.00	0.00	12
6	0.00	0.00	0.00	1
7	0.46	0.79	0.58	82
8	0.51	0.80	0.62	71
9	0.22	0.14	0.17	74
10	0.00	0.00	0.00	6
11	0.00	0.00	0.00	2
12	0.00	0.00	0.00	13
13	0.00	0.00	0.00	18
14	0.83	0.68	0.75	28
15	0.00	0.00	0.00	6
16	0.77	0.77	0.77	13
17	0.00	0.00	0.00	7
18	0.64	0.85	0.73	60
19	0.00	0.00	0.00	40
20	0.00	0.00	0.00	4
21	0.00	0.00	0.00	9
accuracy			0.55	810
macro avg	0.24	0.28	0.25	810
weighted avg	0.45	0.55	0.49	810

Figure 9: performance de prédiction du modèle de régression brut



```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:       modularity_class   No. Observations:                 5403
Model:                         MNLogit   Df Residuals:                     4983
Method:                            MLE   Df Model:                          399
Date:                Wed, 02 Apr 2025   Pseudo R-squ.:                  0.5088
Time:                        13:48:08   Log-Likelihood:                -6710.9
converged:                       True   LL-Null:                       -13662.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
```

Figure 10: statistique d'évaluation du modèle brut

Ces constats nous révèlent une possible multicolinéarité entre les variables prédictives. Par ailleurs, les classes qui composent notre variable Y sont très déséquilibrées et peuvent être difficilement séparables les unes des autres dans certains cas. Ainsi, dans le but de mieux comprendre et d'améliorer le modèle, nous analysons plus en détail les différentes erreurs de classifications pour chacune des classes. En nous appuyant sur la figure 8, nous réalisons que plusieurs classes ont des niveaux de précision nuls. On peut particulièrement isoler les classes de modularité 1, 3, 5, 10, 11, 12,15, 19, 20 et 21. Pour la majorité de ces classes, on peut noter la petitesse de leurs tailles, rendant difficile l'apprentissage pour l'algorithme. On peut isoler la classe 21, qui est composée de textes sans lien au graphe et qui, par le fait, constitue une classe de modularité à part. Face à cela, nous décidons de regarder s'il est possible de regrouper des communautés les unes avec les autres. En analysant de plus près le tableau 4 et en mettant en parallèle les erreurs de prédiction de chacune de ces classes, nous arrivons à plusieurs conclusions que nous détaillons. Les classes 1, 3, 6 et 20 ont des thématiques relatives aux coûts engendrés par une maladie professionnelle. Que cela soit lié à l'aide à domicile, au partage de coûts et à la collaboration des parties, nous décidons de toutes les regrouper dans la classe 0 qui a trait au coût du matériel pour l'aménagement. Pour chacune de ces classes, le modèle a prédit des 0 entre 60% et 100% du temps. Ensuite, nous constatons que les classes 5, 15 et 4 attraient toutes à des thématiques proches du retour en emplois. En effet, ici, nous avons des thèmes relatifs à la définition du handicap, l'expertise médicale, la fixation du montant de remplacement de revenu et la détermination de la capacité de retour en emploi. Nous avons choisi de regrouper ces thèmes, car ils sont suffisamment proches les uns des autres, et nous remarquons une fois de plus une concentration d'erreurs de prévision autour de la

valeur 7. Puis, nous décidons de regrouper la classe 11, 12, 13 qui ont une dominance de prédictions en classe 9 et qui globalement parlent d'enjeux reliés à des enjeux de droits de la personne, comme la condition psychologique, la relation employeur/employé ou finalement les notions relatives à la nature convenable de l'emploi. Outre cela, nous regroupons les classes 16, 17 et 10 avec la 2, car toutes relatives à des éléments spécifiques de preuves qui convergent. Finalement, nous mettons la classe 19 relative à la procédure d'appel et de révision avec la classe 8, qui traite des délais de prolongations. En effet, 80% des prédictions de la classe 19 étaient 8 et on peut expliquer cela par le fait que les droits d'appel sont souvent reliés à la temporalité de la demande pour déterminer sa validité. Après avoir apporté ces modifications, nous avons procédé à plusieurs essais et nous avons constaté que les nouvelles classes 0 et 7 pouvaient être agrégées, car des modèles intermédiaires nous indiquent que ces deux classes se distinguent peu.Cela s'explique par le fait que les thématiques abordées dans toutes ces catégories agrégées sont en réalité la marche suivie par les tribunaux globalement dans les cas d'accident et de maladie professionnelle. De fait, la cour doit valider la présence d'un handicap afin de débattre des montants à allouer pour pallier le préjudice encouru par la victime. En complément, il est important de souligner que, fréquemment pour évaluer la validité d'une prestation de remplacement de revenu, la cour procède à un examen de la capacité de retour en emploi du demandeur. C'est alors à partir de cette évaluation que les montants d'indemnités vont être fixé par le juge administratif. Nous proposons de synthétiser nos regroupements de catégories dans le tableau 5.

ID de communauté	Nb de textes	Communauté fusionnées	Thème de la communauté
0	2469	0, 1, 3, 4, 5, 6, 7, 15, 20	Détermination du handicap et des prestations associées
1	883	2, 10, 16, 17	Éléments de preuves de l'accident du travail
2	679	8, 19	Procédure d'appel et délais
3	618	9, 11, 12, 13	Enjeux de droit de la personne
4	215	14	Aggravation de la maladie par le travail
5	466	18	preuve de la rechute, récidive, aggravation
6	73	21	Composantes non connectées

Table 5 : Tableau récapitulatif de la fusion des communautés

Comme nous le montre ce tableau, nous avons diminué notre nombre de communautés de 21 à 7. Si la classe 0 est très grande et compte de nombreux textes du corpus, globalement, on peut constater que les classes sont relativement équilibrées.

À la suite de cette reclassification, nous réentraînons notre régression multinomiale et nous regardons les performances présentées dans les figures 10 et 11 ci-dessous.

Rapport de classification pour l'échantillon de Test

Précision_Val:0.7459926017262639

	precision	recall	f1-score	support
0	0.71	0.93	0.80	337
1	0.80	0.69	0.74	148
2	0.80	0.74	0.77	121
3	0.65	0.26	0.38	91
4	0.76	0.68	0.72	28
5	0.83	0.82	0.83	72
6	0.00	0.00	0.00	14
accuracy			0.75	811
macro avg	0.65	0.59	0.60	811
weighted avg	0.73	0.75	0.72	811

Rapport de classification pour l'échantillon de Validation

Précision_Val:0.745679012345679

	precision	recall	f1-score	support
0	0.73	0.92	0.81	365
1	0.77	0.67	0.72	130
2	0.81	0.79	0.80	111
3	0.70	0.29	0.41	107
4	0.86	0.68	0.76	28
5	0.69	0.73	0.71	60
6	0.00	0.00	0.00	9
accuracy			0.75	810
macro avg	0.65	0.58	0.60	810
weighted avg	0.74	0.75	0.72	810

Figure 11 : performance de prédiction du modèle de régression avec agrégation des classes de modularité

MNLogit Regression Results

Dep. Variable:	modularity_class	No. Observations:	5403
Model:	MNLogit	Df Residuals:	5283
Method:	MLE	Df Model:	114
Date:	Thu, 03 Apr 2025	Pseudo R-squ.:	0.5249
Time:	13:51:57	Log-Likelihood:	-4005.7
converged:	True	LL-Null:	-8430.6
Covariance Type:	nonrobust	LLR p-value:	0.000

Figure 12 : statistique d'évaluation du modèle avec agrégation des classes de modularité

Ces deux figures nous montrent une amélioration des performances avec une amélioration substantielle de la précision qui passe de 55% à 74%. De plus, on voit dans la figure 11 une amélioration du pseudo R carré, qui est de 52% ainsi que du log-likelihood et du LL-null. En essence, on peut dire que ce modèle est plus performant que le premier nous indiquant que notre agrégation fonctionne. Si on regarde plus en détail le recul et le F1 score, on constate que la classe 3 et 6 est celle avec les moins bonnes performances. Concernant la classe 6, il s'agit des textes isolés, donc nous pouvons postuler que ces textes sont en réalité proches de certaines classes, mais le manque de citations les isole du graphe expliquant les faibles performances. Concernant la classe 3, elle est relative aux droits de

la personne, ce sujet est en réalité transversal à de nombreux sujets. En effet, on peut, par exemple, voire beaucoup de sujets qui sont relatifs à une blessure et qui abordent des enjeux de souffrance psychologique au travail liés à cette blessure. Cela explique la difficulté d'isoler cette communauté via la modélisation de sujet expliquant le 71% de faux négatifs détectés par le modèle. Quand nous regardons les prédictions, nous voyons que la majeure partie des faux négatifs sont classifiés dans la catégorie 0, ce qui renforce notre intuition. Malgré cela, nous ne décidons pas d'agréger ces deux classes, car nous ne voulons pas perdre trop d'information et nous pensons que le débalancement qu'engendrerait cette démarche biaiserait trop les résultats du modèle. Finalement, nous regardons les valeurs P des coefficients distincts en annexe 30. Une fois de plus, nous constatons que toutes sont au-dessus de la valeur acceptable de 5%. Par le fait, nous pensons que la multicolinéarité des variables est probablement la raison derrière cette faible significativité des coefficients. Nous tâchons alors de regarder un modèle avec une sélection de variables pour améliorer cela.

Afin de sélectionner les variables explicatives, nous décidons d'éliminer les sujets avec une variance inférieure ou égale à 0.01. En effet, des variables avec une faible variance ont de plus fortes chances de contribuer à la colinéarité entre les variables. De plus, on peut estimer que ces dernières on peut de chance de contribuer à la prédiction du modèle. Ce filtre élimine quatre sujets :

- Sujet 5 : blessures aux pieds
- Sujet 10 : blessures aux mains
- Sujet 12 : maladies respiratoires
- Sujet 17 : Blessures aux épaules et bras

Ces sujets ont comme particularité d'être dans l'ensemble parmi les moins présents dans le corpus, comme en témoignent les graphiques 8 et 9. À la suite du retrait de ces variables, nous proposons notre dernière régression multinomiale. La figure 12 nous donne les performances de prédiction de ce modèle. On constate une légère dépréciation de la

précision pour les données de test qui peut être lié à l'aléatoire. Cependant, globalement, les performances restent correctes et proches du modèle précédent.

Rapport de classification pour l'échantillon de test

Précision_tst:0.7398273736128237

	precision	recall	f1-score	support
0	0.70	0.93	0.80	337
1	0.80	0.68	0.73	148
2	0.80	0.73	0.76	121
3	0.65	0.26	0.38	91
4	0.71	0.54	0.61	28
5	0.83	0.83	0.83	72
6	0.00	0.00	0.00	14
accuracy			0.74	811
macro avg	0.64	0.57	0.59	811
weighted avg	0.73	0.74	0.72	811

Rapport de classification pour l'échantillon de Validation

Précision_Val:0.7469135802469136

	precision	recall	f1-score	support
0	0.73	0.92	0.81	365
1	0.78	0.67	0.72	130
2	0.82	0.79	0.81	111
3	0.69	0.29	0.41	107
4	0.85	0.61	0.71	28
5	0.70	0.75	0.73	60
6	0.00	0.00	0.00	9
accuracy			0.75	810
macro avg	0.65	0.58	0.60	810
weighted avg	0.74	0.75	0.73	810

Figure 13 : statistique d'évaluation du modèle avec agrégation des classes de modularité et retrait de variables explicatives

La figure 13 ci-dessous nous donne les statistiques du modèle de régression. La diminution du nombre de variables explique la diminution du pseudo r carré et du LL-null ainsi que de la LLR p-value. Cependant, une fois de plus, ces pertes ne sont pas si importantes en comparaison du modèle précédent.

MNLogit Regression Results

Dep. Variable:	modularity_class	No. Observations:	5403
Model:	MNLogit	Df Residuals:	5307
Method:	MLE	Df Model:	90
Date:	Thu, 03 Apr 2025	Pseudo R-squ.:	0.5215
Time:	15:26:58	Log-Likelihood:	-4034.2
converged:	True	LL-Null:	-8430.6
Covariance Type:	nonrobust	LLR p-value:	0.000

Figure 14 : statistique d'évaluation du modèle avec agrégation des classes de modularité et retrait de variables explicatives

La principale différence se situe dans l'analyse des coefficients de corrélation et de leurs significativités. Au regard de l'annexe 31, nous constatons que plusieurs coefficients ont une valeur P significative. Nous ne nous attardons pas à analyser l'ensemble des coefficients et leurs implications, car ce qui nous intéresse est de montrer le lien entre la modélisation de sujet et de graphe. Néanmoins, nous pouvons souligner que certains coefficients nous apportent de l'information supplémentaire. Le premier est le coefficient du X15, qui est le sujet 18 relatif aux maladies de l'audition qui, en comparaison de la classe de modularité de référence 0, est importante et significative pour isoler les classes 2 et 4.

Ces classes sont relatives aux procédures d'appels et aux délais ainsi qu'à l'aggravation de la maladie par le travail. Cela nous indique que les maladies de l'audition ont significativement plus de chance d'être au sein d'enjeux d'appels ou d'aggravation en comparaison des enjeux de détermination du handicap et des prestations qui y sont reliées. Nous soulignons cet élément, car il révèle une information que nous n'aurions pas intuitivement déduite, soulignant ainsi la valeur ajoutée de ce modèle. Par ailleurs, nous devons nuancer ce modèle à la vue des coefficients. Nous avons plutôt parlé de la proximité que la sémantique peut avoir avec la modélisation de graphe. En regardant nos coefficients, nous remarquons que plusieurs d'entre eux mesurent sémantiquement la même information que le graphe. Par exemple, on peut citer le coefficient X11 qui est le sujet 13, soit les éléments de causalités de la blessure qui est très significatif et important pour dissocier la classe de modularité 1 relative à la preuve de la classe de référence. Or, la causalité de la blessure est un élément central de la preuve, ce qui explique la force de ce coefficient. De même, la classe de modularité 5 est à un coefficient très fort pour la variable X9 en comparaison à la classe de référence. Étant donné que la variable X9 est le sujet 13 relatif à la rechute et que la modularité 5 est la communauté de jugements qui a trait à ce même thème, il est logique que les deux soient fortement reliés. Dans un autre ordre d'idées, nous soulignons que la classe de modularité 3, qui regroupe les jugements relatifs aux droits de la personne, a un coefficient de corrélation élevé avec la variable X4, qui est le sujet 3 relatif à la représentation en comparaison à notre catégorie de référence. Ici, le lien peut être déduit par le fait que la juste représentation est un élément central de la justice canadienne et québécoise. Par le fait, les enjeux de représentation par un syndicat, un avocat ou toute autre personne sont intimement liés au droit personnel des individus représentés. Si nous avons relevé quelques éléments de cette analyse, nous vous invitons à regarder les annexes 31 et 32 conjointement pour étayer l'analyse.

Dans ce chapitre, nous avons proposé une approche visant à mélanger la modélisation de sujet et la modélisation de graphe. Les différents modèles de régression nous ont prouvé qu'il existe un lien entre les deux dimensions juridique et sémantique. Cependant, en agrégeant les communautés du graphe, nous perdons aussi de l'information

qui peut s'avérer importante pour les professionnels. Nous notons aussi que le modèle s'il explique plus de la moitié de la variabilité de l'appartenance à une classe de modularité, d'autres variables pourrait rentrer en compte pour expliquer cette dernière. Malgré cela, la précision de 74% nous donne une bonne ligne de base. De plus, ce modèle nous apporte des informations quant aux enjeux que certaines pathologies apportent face au tribunal. Cette intelligence permet de mieux prévenir et préparer un argumentaire pour une partie. Finalement, il est essentiel de souligner que le mélange du LDA avec la régression multinomiale a pour but principal de montrer les liens entre les modèles de notre sujet. Or, la catégorisation proposée par le LDA est sensible aux données sur lesquelles est basé l'algorithme. Ainsi, une autre année de jugement entrainerait probablement une modification des sujets modélisés et possiblement aussi une différente modélisation de réseau de graphe. À ces fins, pour d'autres années de jugements, il faudrait reprendre tous les modèles et ne pas simplement reprendre les données de classification de ce travail. Néanmoins, ce cumul de méthode nous permet de mieux comprendre les structures sous-jacentes aux textes juridiques ainsi que d'envisager des applications de recherches de cas plus précises.

VI) Application concrète, travaux futurs et limites

Si notre approche a montré plusieurs points d'intérêts, nous proposons ce chapitre pour discuter des différents enjeux liés à notre travail et proposer des pistes d'améliorations que nous pourrions proposer dans le futur et nous finirons par évoquer les applications concrètes de notre travail dans la pratique juridique actuelle.

La première considération que nous devons aborder repose sur des considérations purement pratiques. Dans ce mémoire, nous avons fait des choix sur l'allocation de notre temps et de nos ressources. Ainsi, nous avons décidé de prioriser nos analyses à la préparation des données. De ce fait, notre méthode d'extraction des données basée sur des RegEx est certes suffisamment efficace, mais à demander plusieurs contre-vérifications et certains oublis ont de l'être corrigé au cours de nos processus d'analyse. Dans le futur, nous pourrions élargir le graphe avec d'autres années, domaines de droit, juridictions, etc. Or,

dans ce cadre, l'extraction des données de citations doit s'automatiser et être plus précise. Nous avons estimé utiliser un modèle de *deep learning* comme ChatGPT, mais ce dernier étant trop général n'arrivait pas à comprendre toute la subtilité des citations juridiques. Il existe des modèles supervisés réentraînés sur des corpus juridiques comme BERTlaw ou LegalBert, que nous n'avons pas explorés pour cette tâche, mais qui peuvent être des pistes de solution pour faciliter cette dernière. Outre cela, on peut considérer des modèles supervisés plus simples entrainés sur des données de citations québécoises et canadiennes pour effectuer cette tâche. L'enjeu avec cette dernière approche réside dans la complexité d'obtenir des données annotées de droit canadien relativement aux jurisprudences. Dans ce cas, la solution probable est d'annoter nous-mêmes un grand nombre de données pour composer nos échantillons d'entrainement, de test et de validation. Quoi qu'il en soit, cette tâche spécifique d'extraction des données de citations demeure cruciaux dans pour de nombreuses tâches juridiques qui dépassent le simple cas de ce mémoire.

Une autre considération qui est relative aux données concerne la modélisation du graphe. Afin de ne pas déborder en analyse, nous avons choisi l'année de jurisprudence 2017 pour des raisons explicitées plus haut. De plus, nous nous sommes basés sur les décisions issues du tribunal administratif du travail pour constituer notre corpus et nous avons ignoré les autres instances. Ce choix a influencé notre décision dans la représentation en graphe orienté entre les textes du corpus et les citations qu'il contient. Cependant, si nous souhaitons augmenter la taille de notre graphe dans le futur, nous devons en réalité introduire deux nouvelles dimensions dans ce dernier, une dimension temporelle et une dimension hiérarchique de la décision. En effet, si d'autres années de jugements rentrent dans le graphe, nous devons alors considérer une modélisation adaptée à cela. En effet, certaines décisions peuvent être infirmées par une loi ou une autre décision d'une année à une autre. Cela indique que tout un regroupement de jugement peut s'avérer caduc subitement, malgré leurs grands nombres. La nature longitudinale des données doit être intégrée alors dans le graphe en prenant en compte dans le poids des liens, par exemple la date du jugement. Par exemple, on pourrait considérer que plus la décision est éloignée

dans le temps, moins son poids est important, qui est généralement en accord avec les tribunaux qui préfèrent des décisions récentes pour élaborer leurs réflexions. Concernant l'aspect hiérarchique des citations, il est important de comprendre que la *common law* canadienne repose sur une pyramide juridique. Cela veut dire qu'un ordre d'importance des tribunaux existe, commençant en bas par les tribunaux administratifs et finissant en haut par la Cour suprême du Canada. En général, cette hiérarchie détermine la position du tribunal dans le fil d'appel d'un cas. De ce fait, on ne peut considérer sur le même plan une décision du tribunal administratif du travail et une décision de la cour d'appel. De la même manière que pour la temporalité, nous devons prendre ces éléments en compte dans la pondération des décisions pour de futurs travaux afin d'affiner notre cas d'utilisation. Pour ces cas précis, nous devrions solliciter l'expertise d'avocats et de juriste pour nous assurer que notre modélisation demeure cohérente avec l'état du droit actuel et futur.

Pour continuer dans les éléments relatifs à la modélisation de graphe, il serait intéressant de se pencher sur la doctrine. La doctrine est l'ensemble des références qui ne sont pas des lois ou des jurisprudences. Dans notre situation, nous sommes confrontés à un droit technique qui relève fréquemment de la médecine. À mesure que nous avons consulté notre corpus pour faire nos vérifications, nous avons constaté que des citations de revue médicales spécialisées ou de livres médicaux particuliers apparaissaient à fréquence régulière. Même si ces citations ne demeurent pas majoritaires, ces informations peuvent nous aider à détailler le réseau et à le diviser en communautés. Cela peut aider nos classifications et de potentielles prédictions en vue d'améliorer le réseau de graphes.

Concernant nos données, nous devons aussi souligner que la plateforme SOQUIJ regroupe beaucoup de données publiques de jugement. Cependant, des bases de données privées demeurent qui bloque l'accès à des décisions ou les rendent payantes. Ainsi, nous n'avons pas l'ensemble des données de jugement, mais celle qui nous est accessible. Le risque de cela réside dans le fait de potentiellement passer à côté de jugements majeurs qui ne nous sont pas accessible. En revanche, cette faiblesse peut être atténuée en ayant un grand échantillon de jugements, ce qui diminue la probabilité de rater une décision qui ne serait pas accessible par les bases de données publiques.

En ce qui a trait à notre modélisation de sujet, nous avons conservé le paramètre β, qui correspond à la distribution à priori des sujets dans les textes, par défaut. En vue d'améliorer les performances de ce dernier, nous pourrions considérer l'optimisation de ce paramètre pour analyser l'impact de celui-ci sur la performance générale du modèle.

Outre cela, nous pourrions essayer d'autres modèles de modélisation sémantique peut-être plus adaptés à de grandes quantités de données. En effet, des modèles denses comme BERTtopic pourraient s'appliquer même si ces derniers comportent des enjeux, comme le fait que ce dernier assume qu'un seul sujet est attribué par texte. Nous savons aussi qu'il existe des implémentations de *word embeddings* avec du LDA qui pourraient avoir des résultats intéressants à explorer. Ces approches nécessitent cependant de revoir notre processus de dimension nettoyage des données pour s'adapter aux enjeux que portent ces méthodes. En outre, nous avons choisi de nous concentrer sur un modèle de régression logistique multinomiale pour associer la sémantique au juridique. Notre objectif est de conserver le plus possible la nature explicative de nos modèles, or il serait tout de même intéressant d'expérimenter des modèles de *boosting* ou de réseau de neurones afin de voir les différences de performances et de cohérence de la classification des textes dans des classes de modularités.

De plus, nous adressons un enjeu soulevé par la littérature. En effet, cette dernière exprime le besoin d'un niveau de granularité des données plus précis pour les avocats dans les outils technologiques mis à leurs dispositions. On peut donc imaginer que ce travail n'est que la base d'un ensemble d'outils visant à amener cette précision. Deux fonctionnalités nous semblent intéressantes à explorer. La première serait relative au niveau de précision dans le graphe. Nous pourrions imaginer rechercher les paragraphes cités des jurisprudences pour indiquer directement à l'utilisateur le passage du texte pertinent à consulter ou analyser. Cela peut être réalisé en isolant les citations précises afin de retrouver leurs positions dans les textes d'origines en utilisant des vecteurs de mots afin de les appareiller dans les textes d'origine. La deuxième fonctionnalité repose sur la détection des cas caducs. Plus plutôt dans ce chapitre, nous avons évoqué la possibilité du potentiel changement de la norme de droit pouvant modifier l'interprétation de décision passée voir

les rendre invalides. Dans cette optique, il serait intéressant de pouvoir signaler ces décisions afin d'éviter une perte de temps en analyse de la part des utilisateurs. Nous pensons pouvoir réaliser cela en combinant l'analyse des communautés et la compétence de spécialiste dans le domaine. Ces deux éléments conjoints permettraient de compléter la solution proposée aux spécialistes pour s'adapter plus précisément aux besoins exigés par l'industrie.

Il convient de souligner que, depuis le début de la rédaction de ce mémoire, les plateformes de jurisprudence CANLII et SOQUIJ ont ajouté les relations entre les jugements. Ce facteur simplifie l'accès aux données de citations en plus d'en assurer une certaine qualité. Dans la future, il serait intéressant de se baser sur ces données de graphe afin d'affiner nos modèles. De plus, notre approche permet de contribuer à la parution des évolutions du droit. En effet, avec notre approche, nous proposons une vue d'ensemble sur une année de jugement qui permet aux experts en intelligence juridiques de se concentrer sur les points saillants de l'année. Nous pensons que tous ces éléments cumulés peuvent contribuer à une amélioration globale de l'accès à l'information juridique.

Finalement, si nous avons abordés des aspects théoriques, notre approche nous permet déjà d'envisager des améliorations réelles dans la pratique actuelle du droit du travail. Deux grands axes se dessinent alors. Le premier est relatif à la recherche de jurisprudence et de lois pour construire une procédure vis-à-vis du TAT. En effet, aujourd'hui les moteurs juridiques n'inclus pas une structure de graphe et d'analyse sémantique comme nous le proposons. De plus, les procureurs, juges et avocats sont souvent amené à consulter les textes papiers de lois pour définir un cadre propice à la demande des parties. Grâce à notre approche nous serions capables dans un cas de perte d'audition par exemple, de chercher les textes qui sont sémantiquement les plus proche de notre cas. Puis nous pourrions rechercher dans le graphe ces textes, nous permettant rapidement de voir la structure des décisions passées et des lois qui y sont rattachés, accélérant le processus d'identification d'un argumentaire pertinent. Cet élément, permettrait d'accélérer drastiquement le processus de recherche et d'aider le désengorgement des tribunaux qui aujourd'hui pose problème dans notre province. Le deuxième axe concret concerne les

entreprises, le législateur et leurs compréhensions des accidents de travails. En effet, dans notre mémoire nous avons pu révéler des enjeux comme la rechute des cas de surdité par exemple, qui est plus présente que pour d'autres pathologies. Cela nous indique un enjeu spécifique à cette pathologie qui demande une prévention accrue des employeurs voire une adaptation du cadre juridique. Ainsi, ces nouvelles connaissances permettent une meilleure prévention des accidents et des maladies professionnels entrainant une diminution des frais juridique des entreprises, un désengorgement des tribunaux et surtout une amélioration de la qualité de vie des travailleurs touchés par ces pathologies. Si à cela nous pouvons ajouter des données relatives à l'issue des procès dans nos modèles, nous pourrions obtenir une approche innovante unique au Québec pour détecter les angles morts de la loi actuelle sur ces sujets. En résumé, nous pensons que notre approche novatrice apporterait un gain majeur pour la santé et la sécurité au travail dans au Québec, au Canada et possiblement dans tous les états de *common law*.

Si, dans cette partie nous soulignons un ensemble de défis et de limites posés par notre recherche, nous avons aussi proposé des pistes de solutions pour les combler. Il est important de souligner que le domaine de l'intelligence juridique demeure tout de même jeune et que l'utilisation du NLP dans le processus l'est encore plus. De ce fait, de nombreuses possibilités restent à explorer qui peuvent inclure des modèles d'apprentissage profond. En revanche, l'utilisation de ces modèles reste à définir avec clarté, car il serait potentiellement dangereux de laisser la prise de décisions à ces derniers. Dans une optique d'extraction de données, on peut les considérer, mais nous devons essayer de privilégier des modèles explicatifs pour appuyer les experts dans leurs tâches quotidiennes. C'est dans ce sens que nous proposons d'explorer d'autres avenues tout en gardant cet objectif en tête afin de toujours privilégier la lisibilité des modèles sur leurs performances. Finalement, nous avons pu démontrer dans notre revue de littérature que les besoins applicatifs de la profession étaient spécifiques. Ainsi, il serait intéressant d'impliquer la communauté de pratique dans le développement des outils qui leur est adressée pour s'assurer que nous produisons l'intelligence utile à la construction et à la bonne conduite du système juridique québécois et canadien.

VII) Conclusion

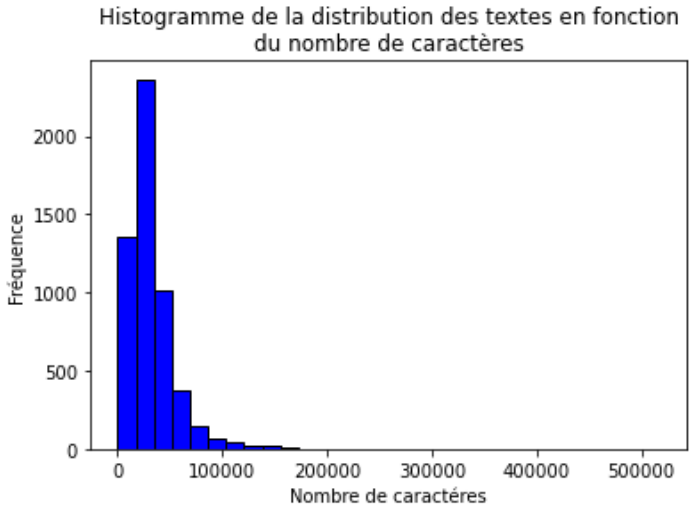
Au cours de ce mémoire, nous avons œuvré à apporter une approche computationnelle des enjeux de droit relatifs aux accidents et maladies du travail afin de générer de l'intelligence pour les entreprises. Dans notre revue de littérature, nous avons constaté que la majorité des travaux issue du NLP pour le droit ont une portée expérimentale et ne s'adresse que rarement aux besoins des avocats. En effet, la recherche se concentre sur des modèles de langues larges avec comme tâches principalement relevées, la prédiction de résultats de procès, la création d'assistants juridiques virtuels. Or, nous avons constaté que plusieurs chercheurs soulignent que ces applications et ces démarches, quoique potentiellement utiles pour les novices dans le domaine, laissent souvent les professionnels pantois quant à leurs utilisations pour la réalisation de leurs tâches quotidiennes. De plus, nous avons abordé l'enjeu éthique que ces modèles boites noires revêtent. Dans la mesure où il est impossible d'expliquer la prise de décision de ces modèles, nous pensons qu'il y a un risque à fonder des choix juridiques sur ces modèles, et ce, pour tous les participants au système juridique. Ainsi, dans l'optique de proposer une meilleure approche de la recherche de cas dans le cadre des maladies et accidents professionnels, nous recommandons d'utiliser un mélange d'approches supervisées et de réseau de graphes. Nous avons alors procédé à l'obtention des jugements du tribunal du travail concernant le sujet spécifique des maladies et accidents professionnels pour l'année 2017 en vue de vérifier le fonctionnement de ces modèles. Outre cela, la sélection de ces approches résulte de l'analyse de notre système juridique reposant sur le principe de la *common law*, qui considère les précédents des tribunaux avec autant d'importance que les textes de loi votés par les députés. Par le fait, nous avons décidé de modéliser une portion sémantique avec l'approche de l'allocation latente de Dirichlet qui permet, à l'aide de vecteurs de mots, de capturer la nature sémantique des jugements que nous analysons. La nature sémantique nous a montré que cette dernière capte dans ses sujets un mélange d'enjeux de procédures et des mots relatifs aux pathologies. Il est intéressant de remarquer qu'après de nombreuses étapes de nettoyages et de diminution de dimensions, nous avons trouvé une grande partie des maladies prescrites par les règlements relatifs aux maladies professionnelles, nous

indiquant la précision de notre division de sujets. Cependant, comme nous l'avions inféré, nous n'avons pas pu extraire la sensibilité juridique que comportent les jugements étudiés. C'est en constituant le réseau de graphe que nous avons pu isoler les communautés dans notre corpus, nous indiquant des regroupements autour de lois et de précédents. À la lecture des textes qui composait les 21 communautés, nous avons raisonnablement pu isoler et identifier chacune d'entre elles autour d'éléments de preuves, de partages de préjudices, etc. Nous avons accompagné cette étude d'une analyse des distances dans le graphe et des similarités de jacquards entre les objets de notre corpus. Cependant, nous avons remarqué que la dimensionnalité affectait fortement ces mesures et les rendait difficilement interprétables. En revanche, principalement avec l'utilisation de la similarité et la distance de Jaccard, nous pouvons déterminer quels textes du corpus sont les plus centraux. En outre, la similarité peut être un outil utile pour recommander des textes avec le plus de similarités dans leurs citations. Par ailleurs, nous n'avons pas pu montrer que la quantité de citations que comporte un texte était reliée à l'importance que celui-ci a dans le champ de la jurisprudence québécoise avec les données que nous avons à notre disposition. Finalement, nous avons regardé si nous pouvions faire converger la modélisation de sujet avec la modélisation des communautés de nœuds que le réseau de graphe nous a permis de dénicher. À ces fins, nous avons utilisé une régression logarithmique multinomiale utilisant les probabilités d'appartenance à chaque sujet comme variables explicatives pour prédire le numéro de communauté à laquelle appartient un texte. L'ensemble des manipulations que nous avons opérées pour réduire la dimensionnalité et augmenter la significativité de ce modèle ont porté leurs fruits, et, malgré un pseudo r carré de 52%, nous obtenons un taux de bonne classification de 74%. De surcroît, les coefficients de corrélations comportent de meilleurs scores de significativités, nous permettant de déduire que la modélisation et la modélisation de graphe sont corrélées. Enfin, cette approche nous permet de générer des éléments d'intelligence pertinents pour la compréhension des cas de maladie et d'accidents professionnels. Finalement, nous avons pu ouvrir nos horizons sur les travaux futurs pour approfondir ce mémoire. Nous soulignons spécifiquement deux axes. Le premier point consiste à étendre

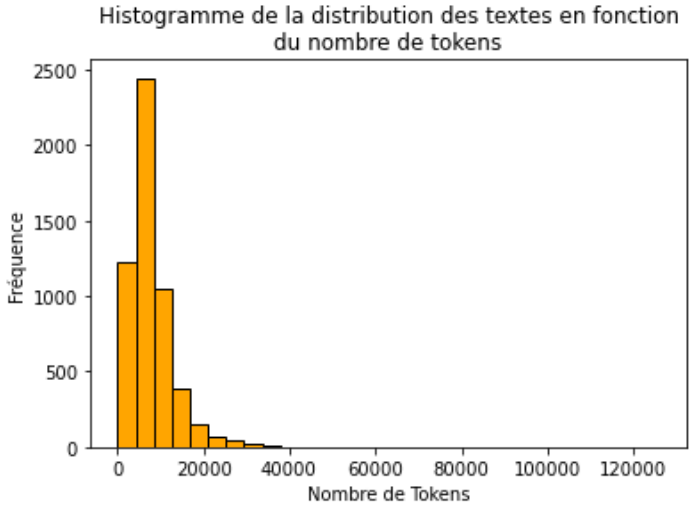
le réseau de graphes en intégrant d'autres années de décisions judiciaires. Cela nous permettrait de suivre l'évolution des normes juridiques au fil des ans, ce qui nous aiderait à affiner nos analyses. Le deuxième axe concerne la pondération des décisions de justice en fonction de l'instance de laquelle elles sont issues. En accord avec la théorie canadienne de la hiérarchie judiciaire, nous pensons que ce travail nous permet de donner plus d'impact à des décisions venues des tribunaux les plus importants nous permettant de nuancer la valeur apportée par une décision. Enfin, nous pensons qu'avec ces deux apports, nous serions capables de mieux détecter les jugements ou lois rendus caducs par le temps, le législateur et les juges. Ce facteur est d'une importance primordiale pour ne pas baser un argumentaire sur des faits de droits qui ne sont plus d'actualités, rendant nos analyses fausses.

En somme, ce mémoire a contribué à mettre en lumière une nouvelle perspective sur l'utilisation des méthodes statistiques dans le cadre de l'intelligence d'affaire légale. Sans nier la pertinence ni le rôle potentiel des modèles de langue larges pour les années futures, nous avons montré la faiblesse et la dangerosité de ces modèles pour le domaine juridique. Notre approche favorise la mise en avant la complexité des raisonnements que comporte le système légal. Si une grande partie des travaux s'attache à prédire l'issue des procès, nous avons démontré l'importance prédominante pour les parties prenantes de mieux visualiser les nuances que comporte la loi et son application. En outre, nous ouvrons la possibilité d'utilisé des modèles entrainés prédictifs en complément de notre approche pour améliorer l'intelligence juridique dans le secteur des accidents de travail. De plus, nous avons proposé des cas concrets qui à travers l'utilisation de nos modèles permettrait déjà un meilleur fonctionnement de nos institutions juridiques. Ainsi, c'est pour l'ensemble de ces raisons que nous pensons que notre approche à une place cruciale dans la fluidification de la justice québécoise et de la santé au travail. Finalement, nous pensons pouvoir élargir notre approche à d'autre secteur du droit dans le Canada et les pays adoptant la *Common law* , ce qui constitue une avancée majeure pour la modernisation des tribunaux dans les années à venir.

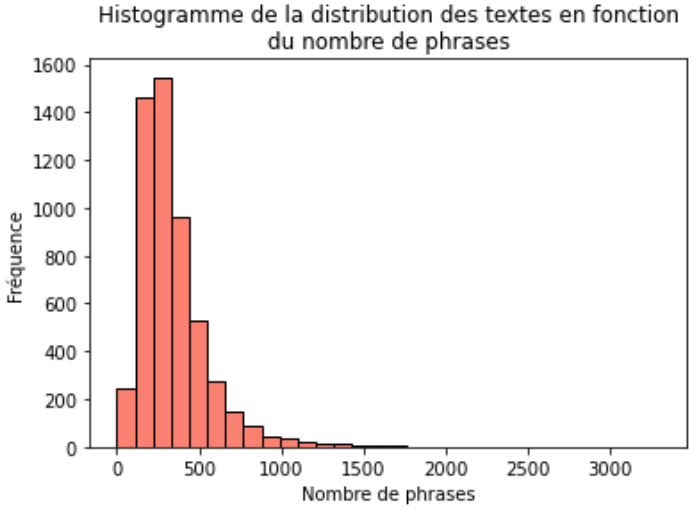
Annexes



Annexe 1 : Histogramme de la distribution des textes en fonction du nombre de caractères



Annexe 2 : Histogramme de la distribution des textes en fonction du nombre de mots

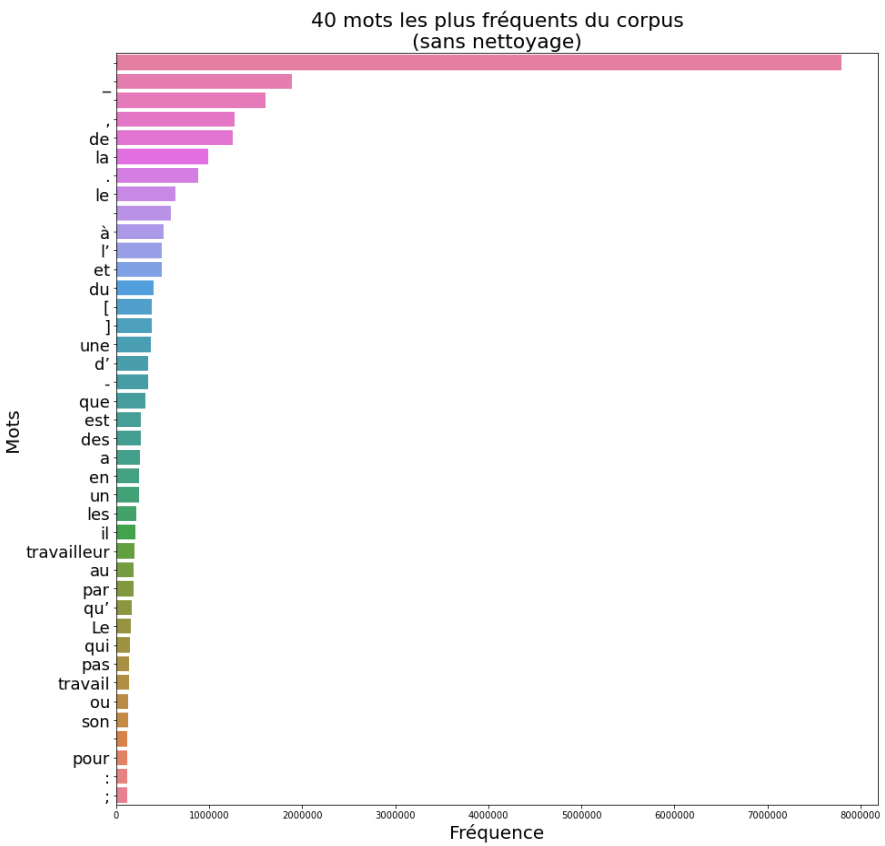


Annexe 3 : histogramme de la distribution des textes en fonction du nombre de phrases

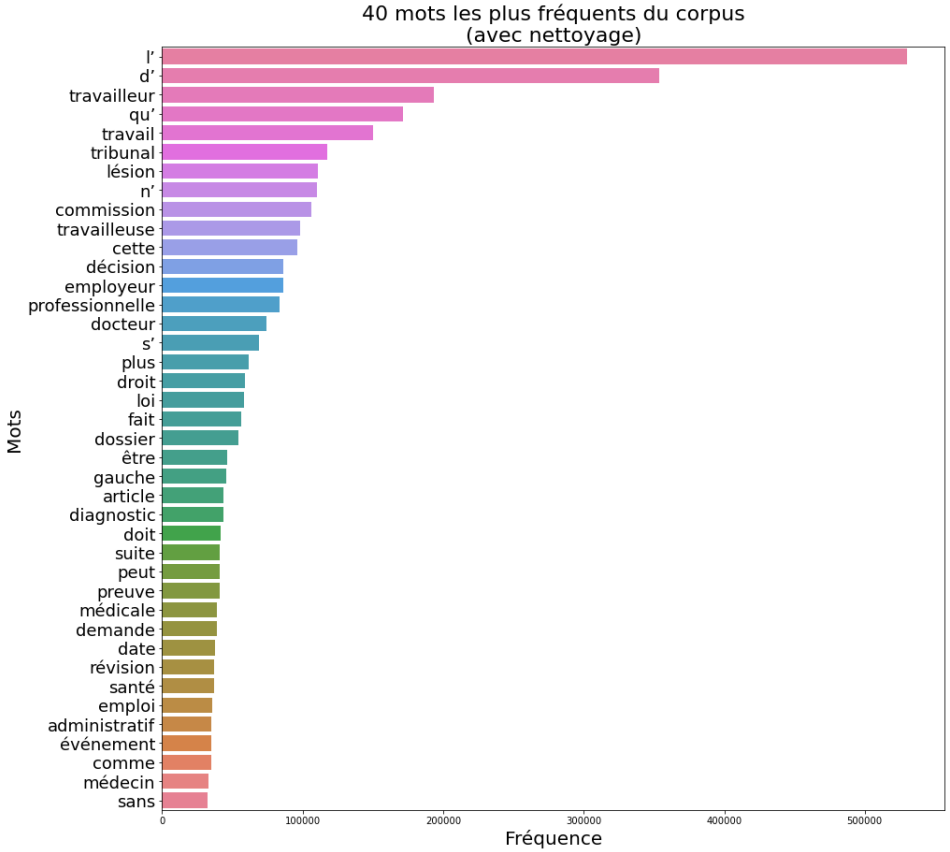
Liste des stop-words NLTK (french) :

'furent', 'à', 'qu', 'étées', 'soit', 'fussiez', 'notre', 'fussent', 'et', 'au', 'auraient', 'serions', 'par', 'eusse', 'étant', 'soyons', 'eut', 'tu', 'étaient', 'soient', 'auront', 'seriez', 'fut', 'sa', 'une', 'ce', 'son', 'nos', 'j', 'eussiez', 'qui', 'avez', 's', 'ayants', 'ces', 'ayante', 'avions', 'de', 'mais', 'aurais', 'avais', 'c', 'ayantes', 'eussent', 'en', 'aie', 'ayons', 'fûmes', 'vous', 'avons', 'pour', 'que', 'ai', 'elle', 'la', 'leur', 'aurons', 'serez', 'était', 'd', 'aurai', 'aient', 'eux', 'ne', 'les', 'ou', 'étants', 'seront', 'moi', 'auriez', 'seraient', 'lui', 'fûtes', 'même', 'ayant', 'ait', 'aurez', 'te', 'le', 'fus', 'été', 'eussions', 'serai', 'aura', 'eûmes', 'sera', 'tes', 'avait', 'l', 'serais', 'fusses', 'vos', 'il', 'm', 'auras', 'je', 'êtes', 'ayez', 'du', 'avec', 'mon', 'fussions', 't', 'ils', 'me', 'suis', 'avaient', 'eurent', 'des', 'votre', 'étés', 'n', 'aies', 'eus', 'ma', 'dans', 'aurions', 'serons', 'se', 'eue', 'étions', 'étantes', 'as', 'toi', 'étais', 'soyez', 'étée', 'fusse', 'ont', 'eusses', 'sommes', 'fût', 'nous', 'ses', 'on', 'un', 'étiez', 'sur', 'serait', 'eûtes', 'sont', 'sois', 'aurait', 'est', 'mes', 'y', 'eues', 'pas', 'es', 'ton', 'ta', 'eût', 'aviez', 'étante', 'seras', 'eu', 'aux'

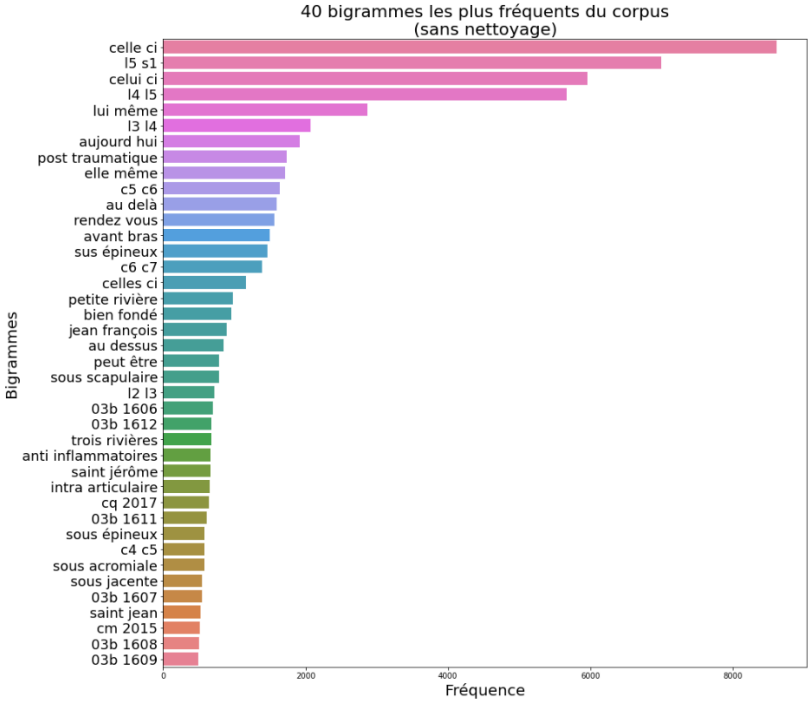
Annexe 4 : liste des stop-words français de NLTK



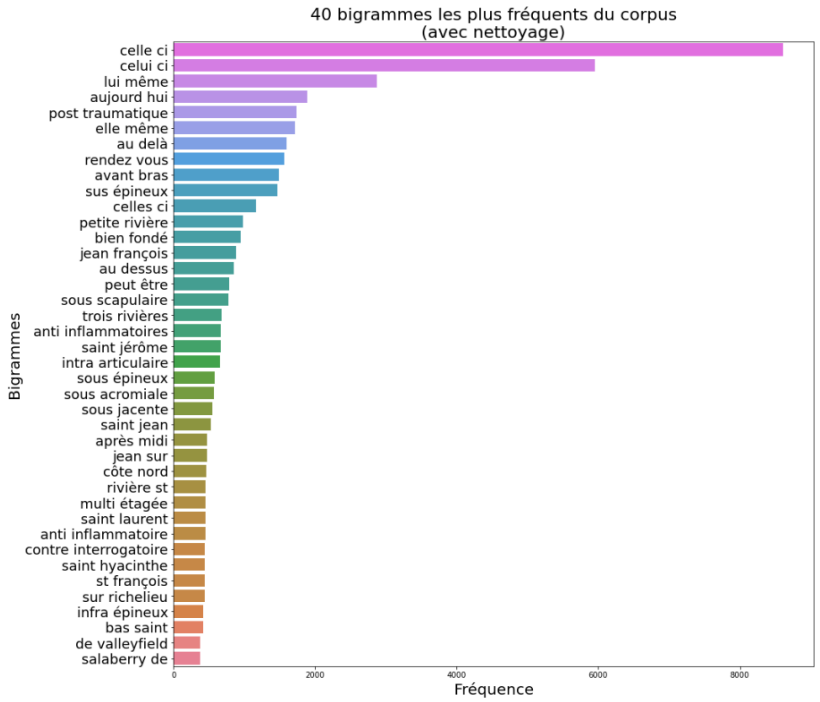
Annexe 5: Histogramme des 40 Tokens les plus fréquents non filtrés



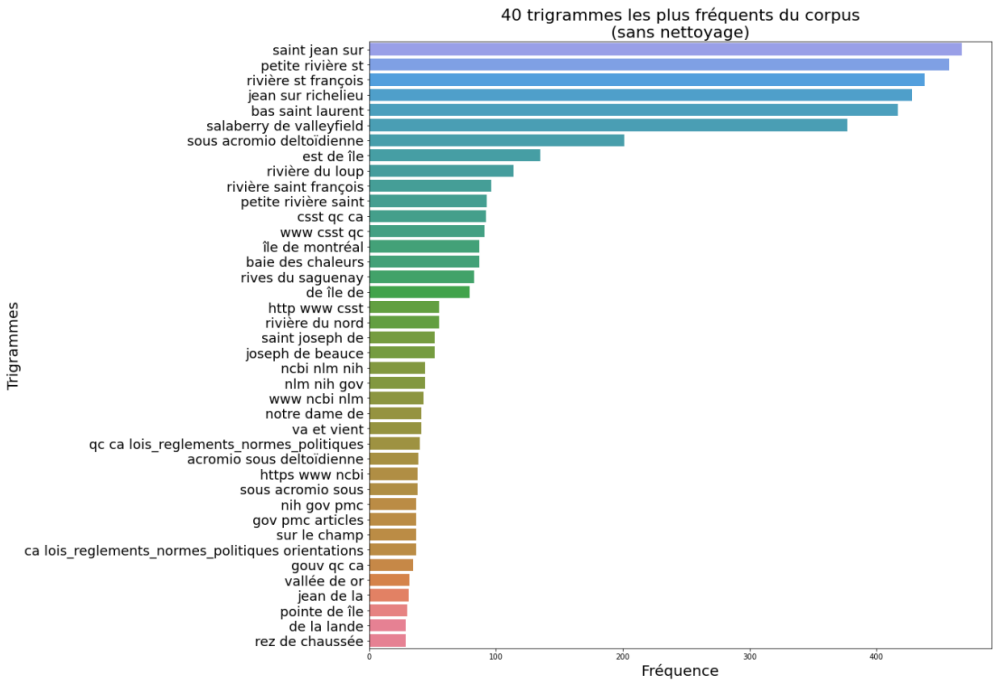
Annexe 6: Histogramme des 40 Tokens les plus fréquents filtrés



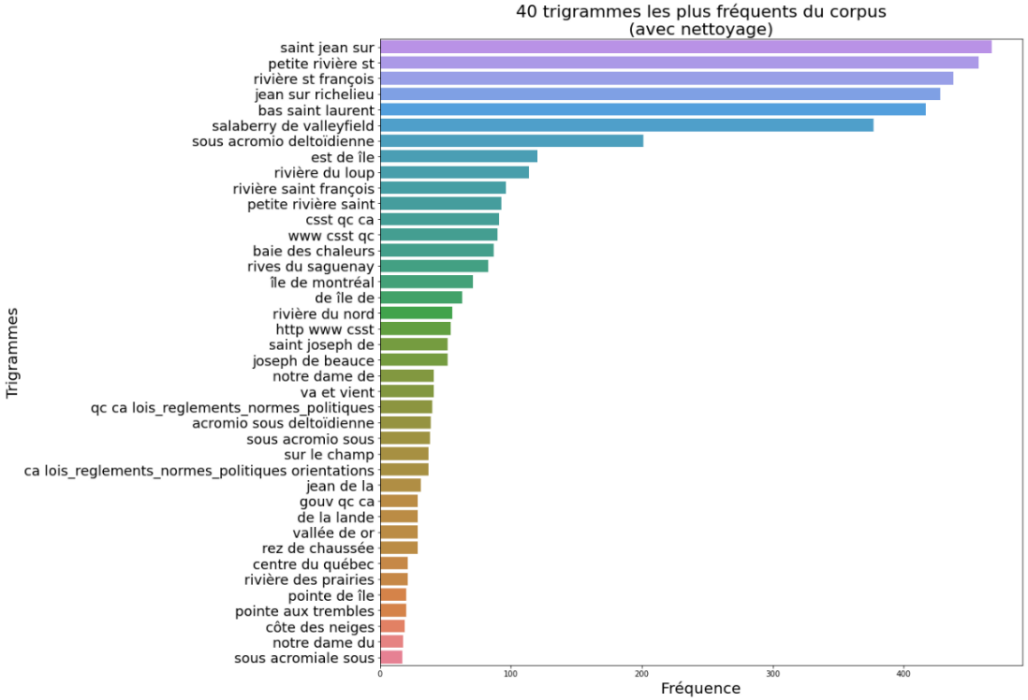
Annexe 7 : Histogramme des 40 bigrammes les plus fréquents non filtrés



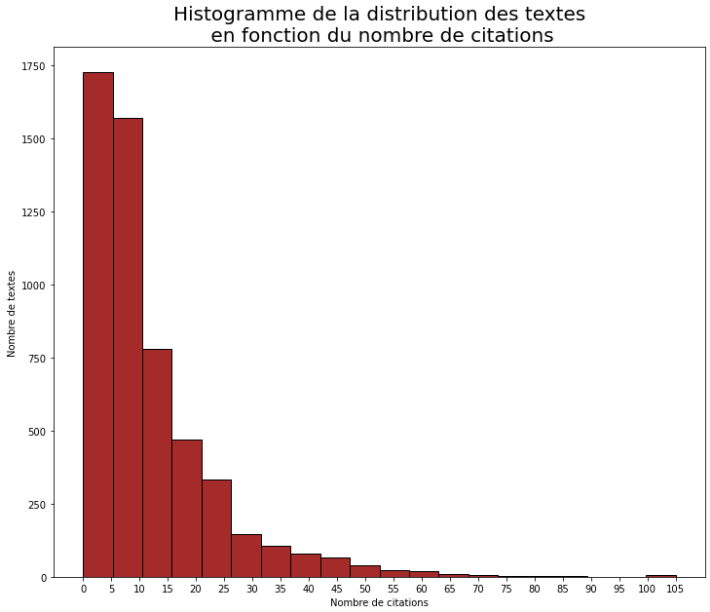
Annexe 8: Histogramme des 40 bigrammes les plus fréquents filtrés



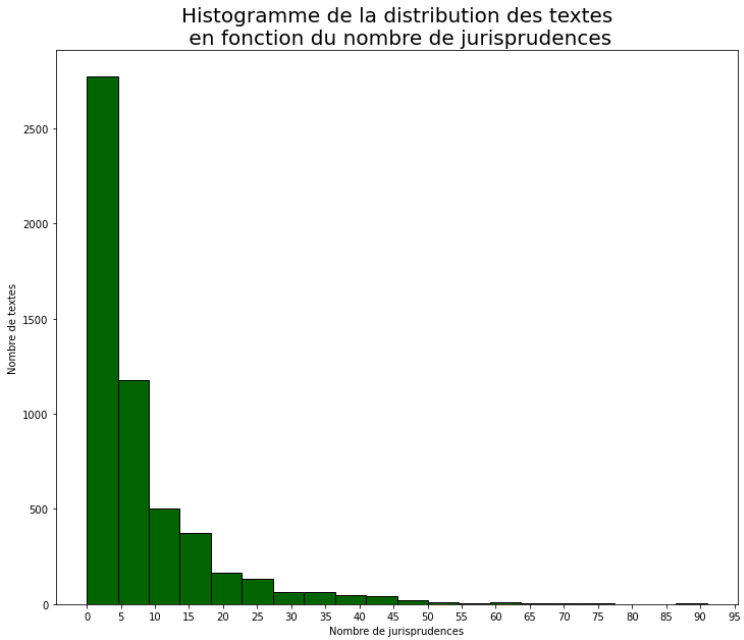
Annexe 9 : Histogramme des 40 trigrammes les plus fréquents non filtrés



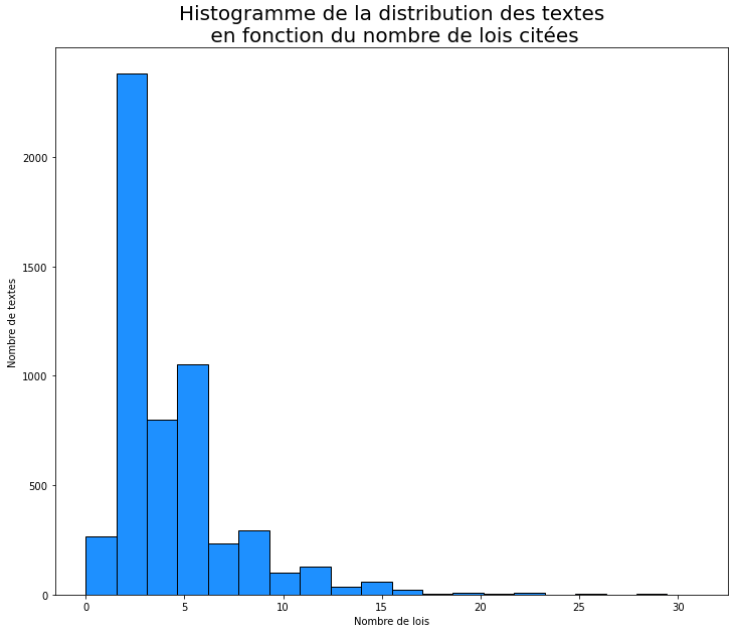
Annexe 10 : Histogramme des 40 trigrammes les plus fréquents filtrés



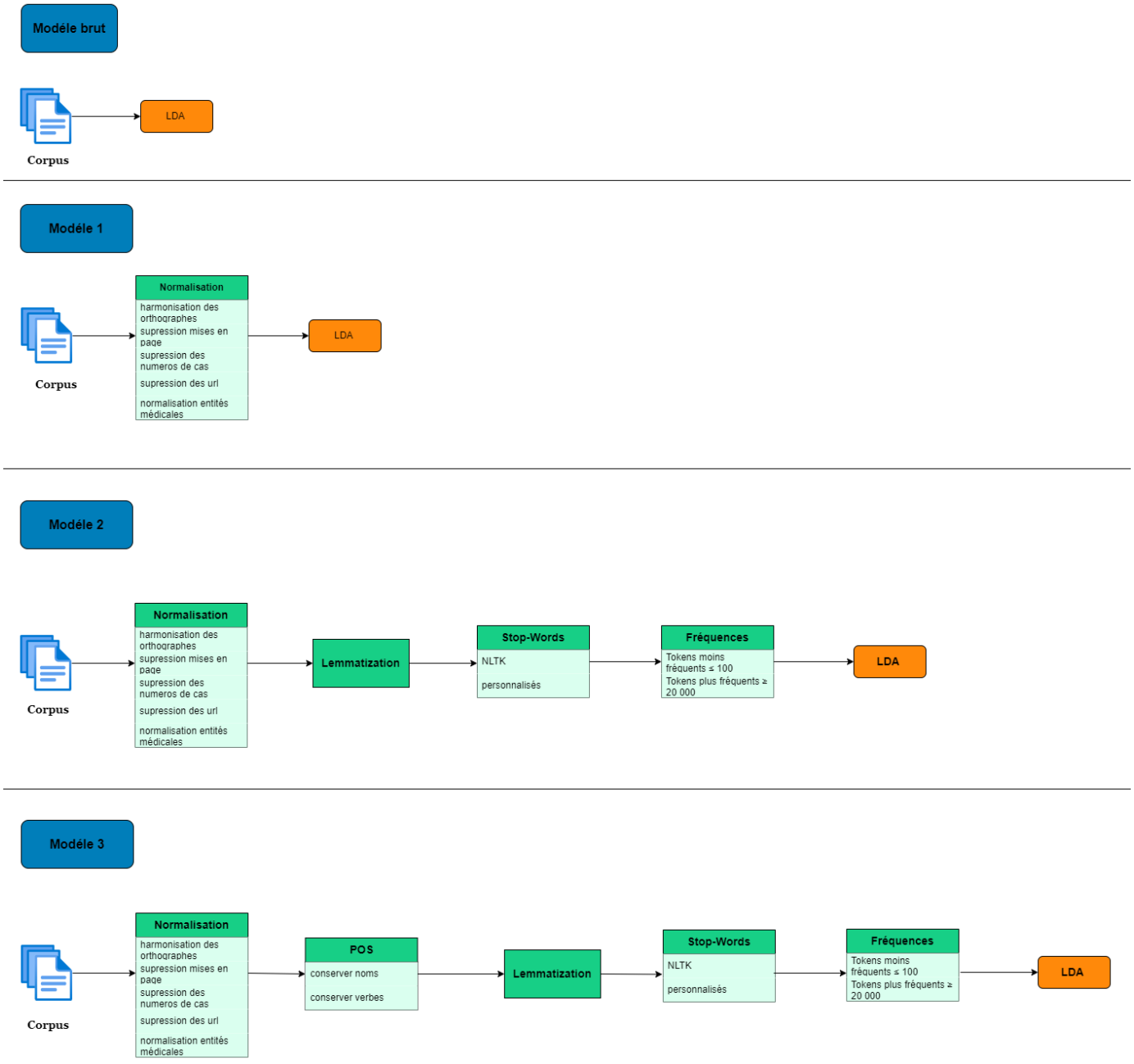
Annexe 11 : histogramme de la distribution des textes en fonction du nombre de citations



Annexe 12 : histogramme de la distribution des textes en fonction du nombre de jurisprudences citées



Annexe 13 : histogramme de la distribution des textes en fonction du nombre de lois citées



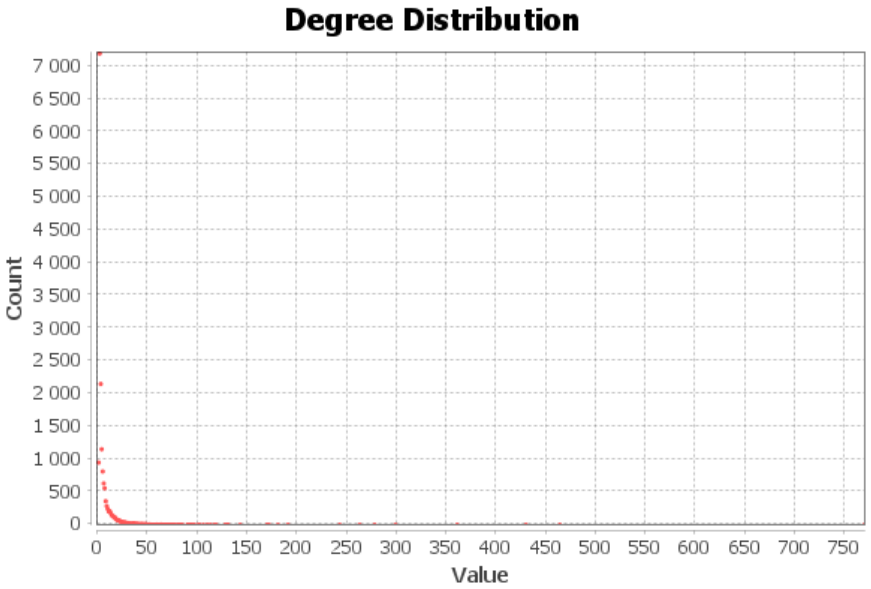
Annexe 14 : Diagramme des nettoyages des données pour les différents modèles de LDA

Modèle 1										
Numéro du sujet	**Mot 1**	**Mot 2**	**Mot 3**	**Mot 4**	**Mot 5**	**Mot 6**	**Mot 7**	**Mot8**	**Mot 9**	**Mot 10**
0	0.027*"date"	0.017*"docteur"	0.017*"travailleur"	0.014*"fonctionnelles"	0.014*"limitations"	0.012*"pour"	0.010*"avec"	0.010*"permanente"	0.009*"commission"	0.009*"rapport"
1	0.019*"pour"	.016*"dans"	0.014*"employeur"	0.013*"travail"	0.009*"elle"	0.008*"monsieur"	0.007*"date"	0.006*"avec"	0.006*"plus"	0.006*"fait"
2	0.027*"lombaire"	0.023*"date"	0.017*"discale"	0.016*"hernie"	0.014*"travailleur"	0.012*"vertebrev"	0.012*"entorse"	0.012*"sacrum"	0.011*"docteur"	0.010*"avec"
3	0.029*"pour	0.023*"travailleur"	0.014*"commission"	0.012*"date"	0.011*"aide"	0.011*"domicile"	0.011*"dans"	0.009*"réadaptation"	0.009*"remboursement"	0.009*"frais"
4	0.026*"date"	0.025*"décision"	0.018*"tribunal"	0.016*"commission"	0.016*"dans"	0.014*"pour"	0.012*"révision"	0.010*"elle"	0.008*"article"	0.008*"délai"
5	0.070*"travailleur"	0.030*"travail"	0.025*"date"	0.017*"pour"	0.015*"professionnelle"	0.015*"dans"	0.014*"commission"	0.013*"tribunal"	0.010*"maladie"	0.010*"employeur"
6	0.030*"date"	0.017*"docteur"	0.014*"gauche"	0.011*"droit"	0.011*"douleur"	0.011*"avec"	0.010*"droite	0.009*"diagnostic"	0.009*"examen"	0.009*"elle"
7	0.041*"date"	0.022*"travailleur"	0.021*"emploi"	0.016*"commission"	0.015*"droit"	0.014*"revenu"	0.014*"lésion"	0.013*"indemnité"	0.012*"dans"	0.012*"tribunal"
8	0.041*"date"	0.022*"travailleur"	0.020*"aggravation"	0.020*"lésion"	0.019*"récidive"	0.019*"rechute"	0.011*"professionnelle"	0.011*"travail"	0.010*"commission"	0.010*"avec"
9	0.038*"date"	0.022*"commission"	0.014*"examen"	0.014*"médical"	0.014*"assignation"	0.013*"médecin"	0.013*"employeur"	0.012*"temporaire"	0.011*"suspendre"	0.010*"suspension"
10	0.019*"date"	0.018*"droite"	0.018*"déchirure"	0.016*"coiffe"	0.013*"rotateurs"	0.012*"lésion"	0.011*"docteur"	0.011*"travailleur"	0.010*"dans"	0.010*"tendinite"
11	0.024*"lésion"	0.022*"employeur"	0.020*"date"	0.018*"professionnelle"	0.015*"travailleur"	0.014*"dans"	0.011*"déficience"	0.010*"tribunal"	0.010*"commission"	0.009*"cette"
12	0.040*"genou"	0.029*"date"	0.021*"droit"	0.016*"travailleur"	0.016*"interne"	0.014*"déchirure"	0.013*"gauche"	0.010*"pour"	0.010*"docteur"	0.009*"lésion"
13	0.016*"dans"	0.013*"maladie"	0.011*"travail"	0.009*"pour"	0.008*"sont"	0.007*"plus"	0.007*"cette"	0.006*"docteur"	0.006*"carpien"	0.006*"elle"
14	0.040*"elle"	0.039*"travailleuse"	0.024*"travail"	0.020*"date"	0.013*"dans"	0.011*"pour"	0.010*"lésion"	0.009*"tribunal"	0.008*"fait"	0.008*"madame"

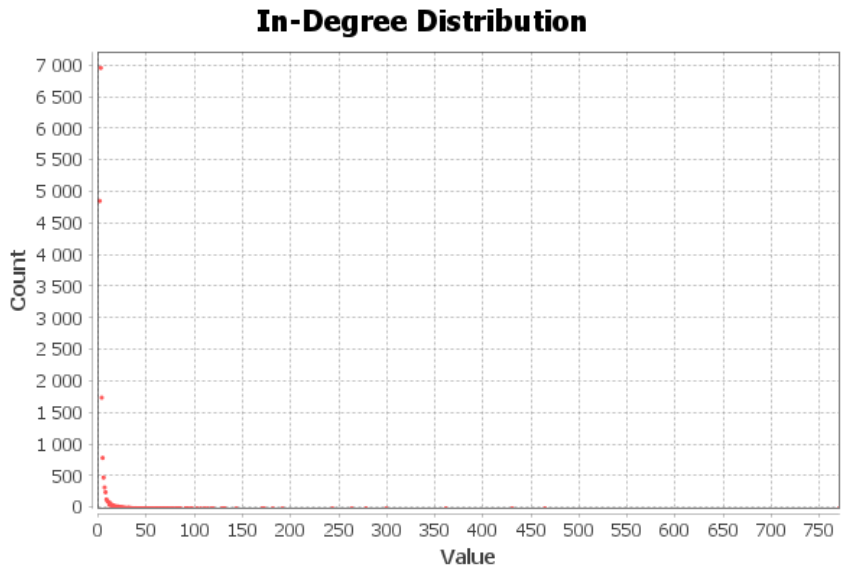
Annexe 15 : Tableau résumé des vecteurs de mots pour le modèle 1

Modèle 2										
Numéro du sujet	**Mot 1**	**Mot 2**	**Mot 3**	**Mot 4**	**Mot 5**	**Mot 6**	**Mot 7**	**Mot8**	**Mot 9**	**Mot 10**
0	0.031*"revenir"	0.028*"indemnité"	0.018*"exercer"	0.014*"remplacemer"	0.011*"capacité"	0.010*"capable"	0.009*"convenable"	0.009*"remplacement"	0.008*"prestation"	0.008*"recevoir"
1	0.009*"trouble"	0.007*"psychologique"	0.005*"rencontre"	0.005*"adaptation"	0.005*"dernier"	0.005*"situation"	0.005*"prendre"	0.004*"aussi"	0.004*"bien"	0.004*"mesure"
2	0.009*"survenir"	0.007*"travailler"	0.006*"indique"	0.005*"prendre"	0.005*"occasion"	0.005*"blessur"	0.005*"journée"	0.005*"vers"	0.005*"semaine"	0.005*"avant"
3	0.023*"déchirure"	0.019*"coiffe"	0.018*"rotateur"	0.016*"tendinite"	0.014*"tendon"	0.007*"biceps"	0.007*"tendinopathie"	0.007*"magnétique"	0.007*"supérieur"	0.007*"acromio"
4	0.022*"blessur"	0.020*"présomption"	0.010*"survenir"	0.009*"existence"	0.008*"soudain"	0.008*"imprévu"	0.008*"application"	0.008*"apparition"	0.007*"cause"	0.007*"blessure"
5	0.042*"coût"	0.025*"prestation"	0.022*"imputation"	0.017*"imputer"	0.017*"période"	0.015*"partag"	0.014*"handicap"	0.014*"conséquence"	0.010*"partage"	0.010*"unité"
6	0.036*"discal"	0.029*"hernie"	0.022*"vertebrev"	0.022*"sacrum"	0.016*"vertebreiv"	0.012*"entors"	0.010*"sténose"	0.009*"dégénératif"	0.009*"léger"	0.008*"discopathie"
7	0.015*"mouvement"	0.013*"tâche"	0.011*"effectuer"	0.010*"main"	0.007*"répétitif"	0.006*"temps"	0.006*"coud"	0.006*"façon"	0.006*"poste"	0.006*"poignet"
8	0.038*"cervical"	0.009*"dorsal"	0.008*"entors"	0.008*"symptôme"	0.007*"normal"	0.007*"traumatisme"	0.007*"supérieur"	0.007*"léger"	0.007*"entorse"	0.005*"région"
9	0.023*"délai"	0.015*"contestation"	0.014*"raisonnable"	0.010*"prévoir"	0.010*"contester"	0.009*"relever"	0.009*"défaut"	0.009*"demander"	0.008*"recevoir"	0.008*"déposer"
10	0.017*"poignet"	0.013*"fracture"	0.011*"main"	0.010*"chevill"	0.009*"pied"	0.008*"syndrome"	0.007*"séquelle"	0.007*"doigt"	0.006*"déficit"	0.006*"pourcentage"
11	0.029*"déficience"	0.016*"coût"	0.011*"conséquence"	0.011*"survenance"	0.010*"biomédical"	0.009*"prestation"	0.009*"handicap"	0.008*"déjà"	0.008*"handicapé"	0.007*"chez"
12	0.069*"genou"	0.025*"interne"	0.024*"déchirure"	0.013*"ménisqu"	0.012*"méniscal"	0.011*"ligament"	0.009*"externe"	0.008*"arthrose"	0.007*"antérieur"	0.007*"fémoro"
13	0.014*"activité"	0.013*"coût"	0.012*"tiers"	0.010*"prestation"	0.007*"véhicule"	0.007*"attribuable"	0.007*"circonstance"	0.007*"transfert"	0.006*"imputation"	0.006*"survenir"
14	0.011*"soin"	0.009*"entors"	0.008*"capacité"	0.007*"indemnité"	0.007*"physique"	0.006*"flexion"	0.006*"consolider"	0.006*"charge"	0.006*"amélioration"	0.006*"conclusion"
15	0.019*"aide"	0.014*"domicil"	0.013*"frais"	0.012*"assistance"	0.010*"besoin"	0.010*"soin"	0.010*"remboursement"	0.010*"personnel"	0.009*"réadaptation"	0.009*"service"
16	0.043*"aggravation"	0.042*"rechute"	0.038*"récidiv"	0.019*"initial"	0.008*"modification"	0.007*"relation"	0.007*"alléguer"	0.007*"nouveau"	0.006*"présence"	0.006*"reconnaître"
17	0.008*"erreur"	0.007*"cour"	0.007*"nouveau"	0.007*"question"	0.007*"party"	0.006*"syndicat"	0.006*"fond"	0.005*"vice"	0.005*"procédure"	0.005*"québec"
18	0.022*"bruit"	0.014*"surdité"	0.013*"auditif"	0.012*"chez"	0.008*"exercé"	0.007*"pulmonaire"	0.007*"atteindre"	0.006*"maladie"	0.006*"causer"	0.006*"amiant"

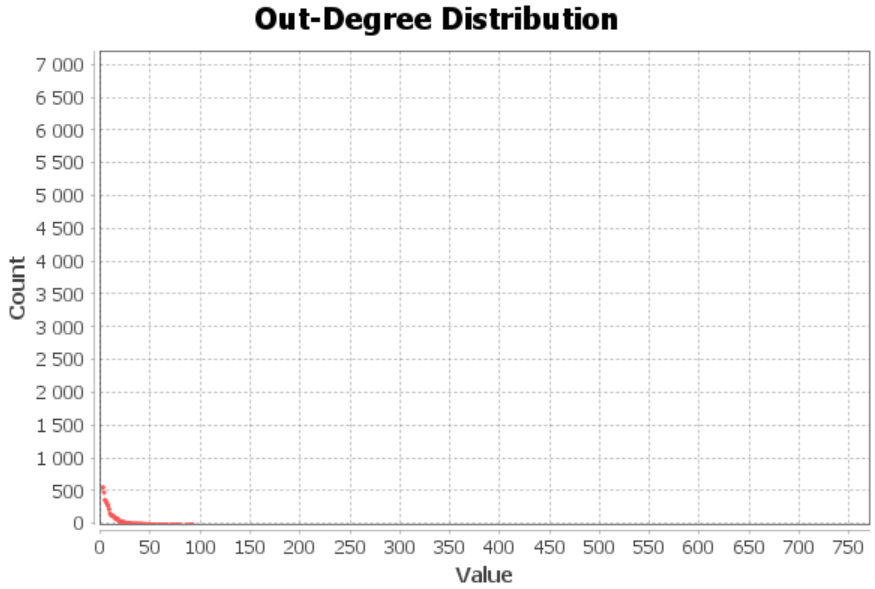
Annexe 16 : Tableau résumé des vecteurs de mots pour le modèle 2



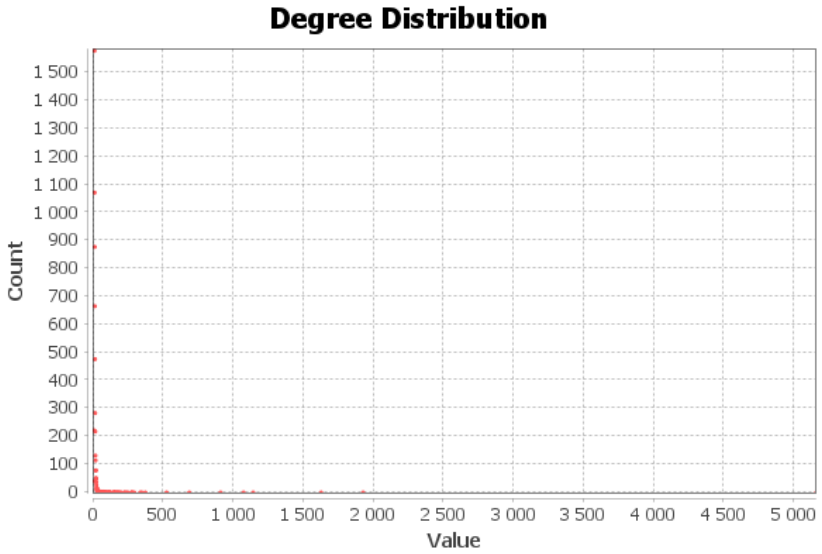
Annexe 17 : distribution du degré pour le graphe de jurisprudence



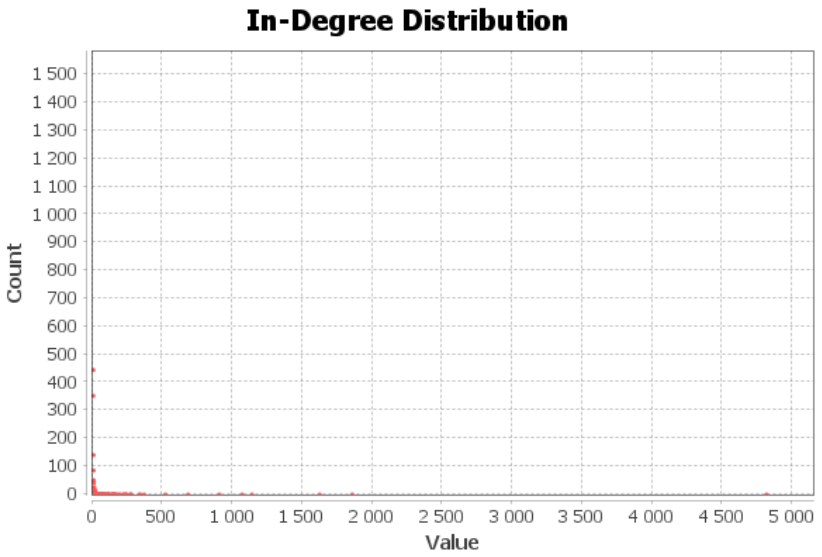
Annexe 18 : distribution du degré entrant pour le graph de jurisprudence



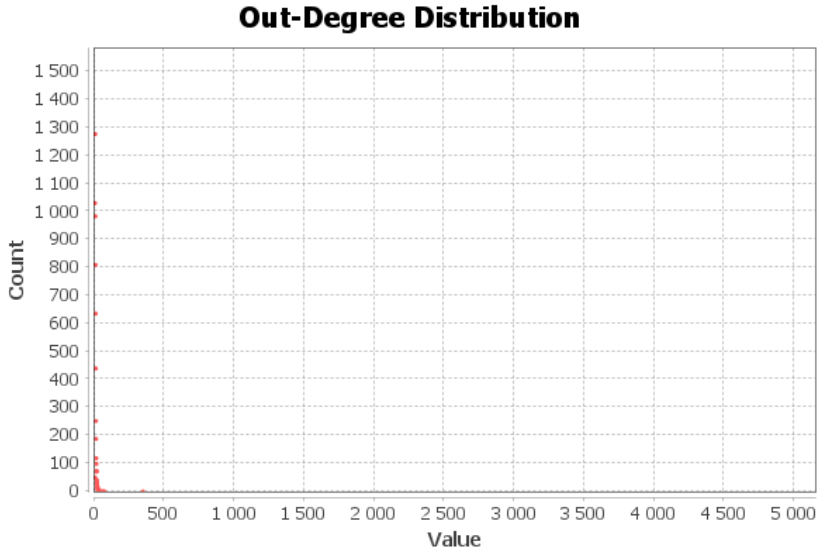
Annexe 19 : distribution du degré sortant pour le graph de jurisprudence



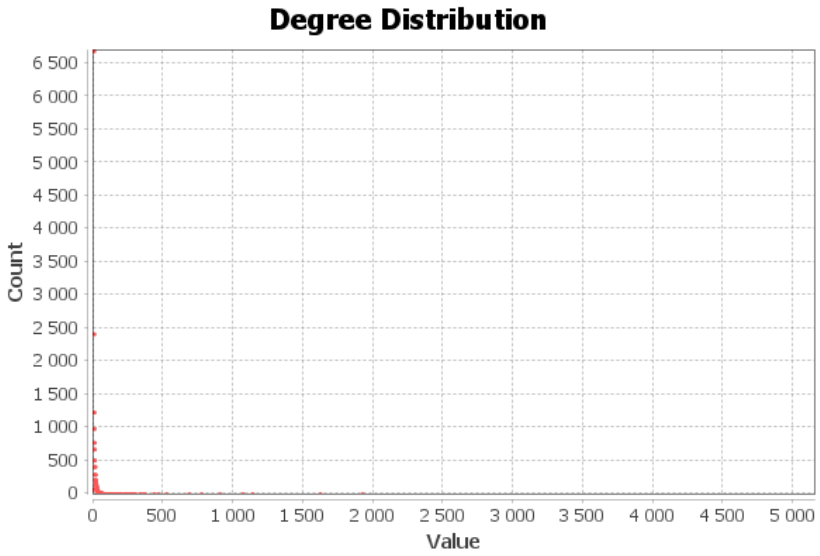
Annexe 20 : distribution du degré pour le graphe des lois



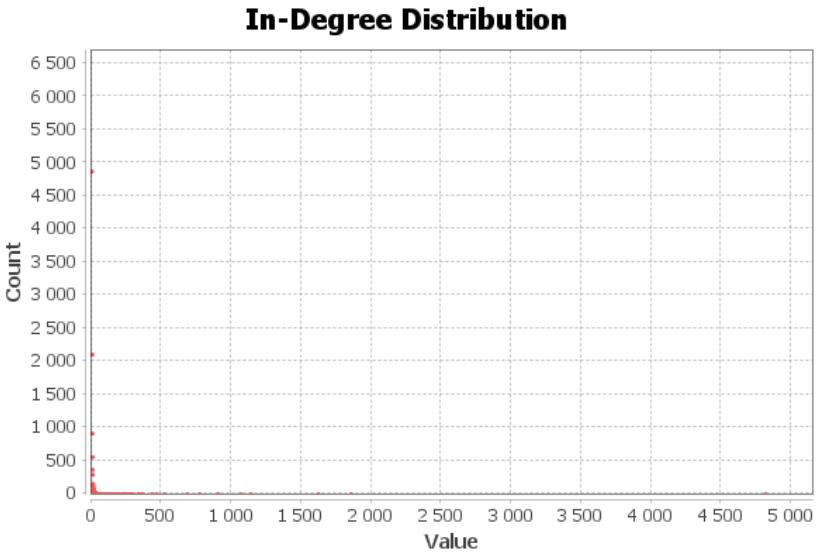
Annexe 21: distribution du degré entrant pour le graph des lois



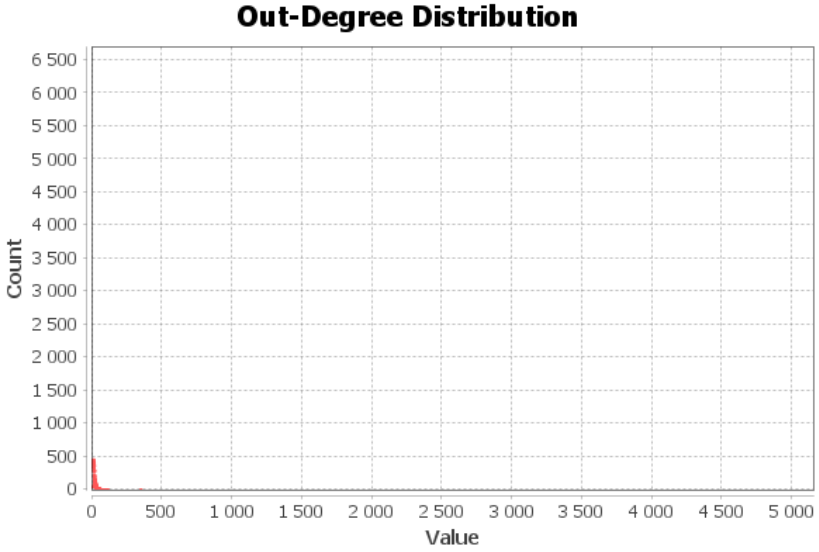
Annexe 22 : distribution du degré sortant pour le graph des lois



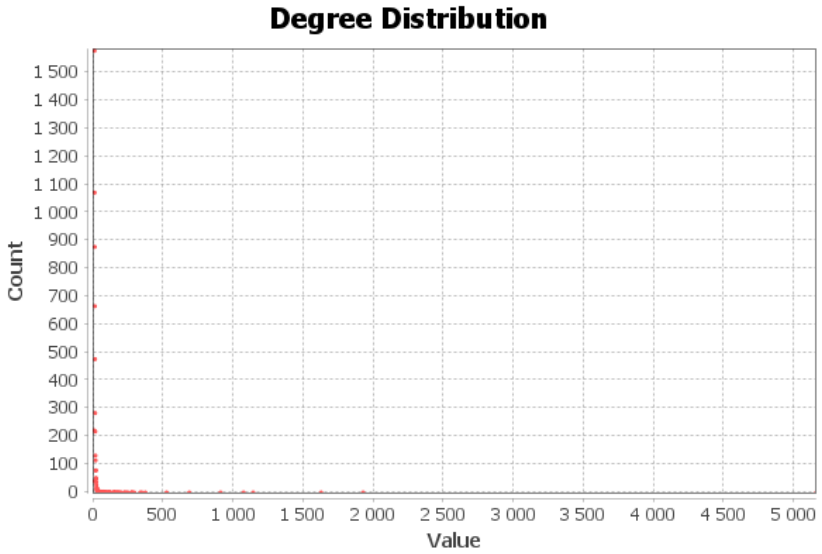
Annexe 23 : distribution du degré pour le graphe complet



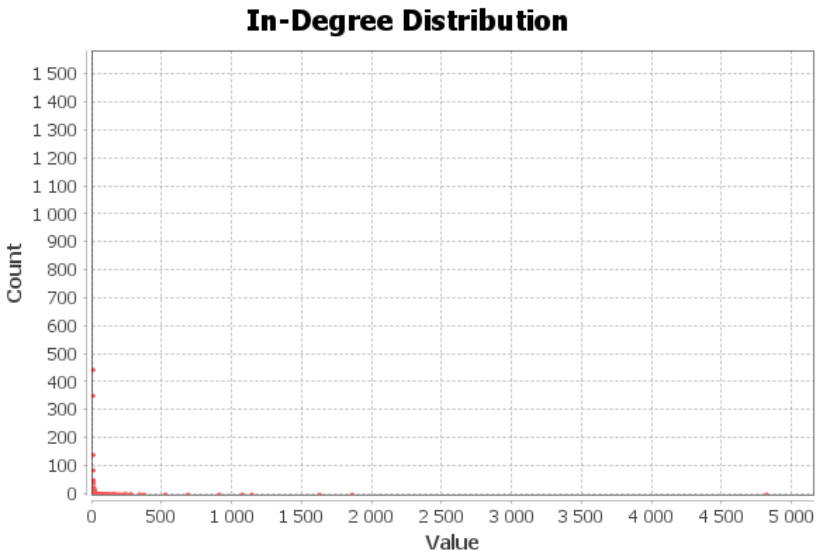
Annexe 24 : distribution du degré entrant pour le graphe complet



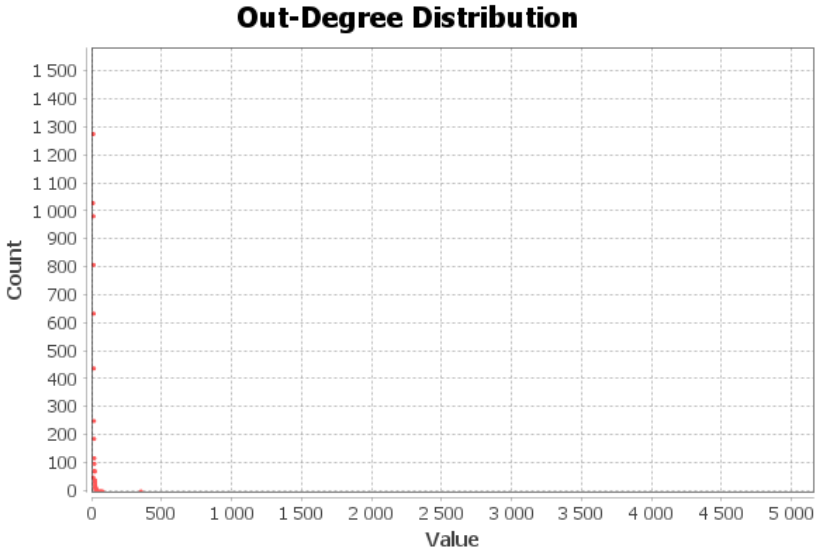
Annexe 25 : distribution du degré sortant pour le graphe complet



Annexe 26 : distribution du degré pour le graphe des lois



Annexe 27: distribution du degré entrant pour le graph des lois



Annexe 28 : distribution du degré sortant pour le graph des lois

MNLogit Regression Results

Dep. Variable:	modularity_class	No. Observations:	5403
Model:	MNLogit	Df Residuals:	5243
Method:	MLE	Df Model:	150
Date:	Mon, 31 Mar 2025	Pseudo R-squ.:	0.5355
Time:	16:37:27	Log-Likelihood:	-5289.1
converged:	True	LL-Null:	-11387.
Covariance Type:	nonrobust	LLR p-value:	0.000

modularity_class=2	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5870	0.303	-1.939	0.053	-1.180	0.006
x1	1.5637	0.548	2.852	0.004	0.489	2.638
x2	-5.5227	0.924	-5.977	0.000	-7.334	-3.712
x3	-8.4991	2.203	-3.858	0.000	-12.817	-4.182
x4	1.0539	0.731	1.442	0.149	-0.379	2.487
x5	-0.0359	0.580	-0.062	0.951	-1.173	1.102
x6	-3.6758	0.787	-4.674	0.000	-5.217	-2.134
x7	-1.6438	0.528	-3.113	0.002	-2.679	-0.609
x8	-6.5039	0.964	-6.745	0.000	-8.394	-4.614
x9	-0.4576	0.879	-0.521	0.603	-2.180	1.265
x10	0.5657	0.510	1.110	0.267	-0.433	1.565
x11	5.5430	0.472	11.748	0.000	4.618	6.468
x12	-1.8257	0.653	-2.797	0.005	-3.105	-0.546
x13	1.4942	0.705	2.120	0.034	0.113	2.875
x14	0.6437	0.483	1.331	0.183	-0.304	1.591
x15	2.1568	0.642	3.359	0.001	0.898	3.415

modularity_class=4	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4109	0.431	-3.271	0.001	-2.256	-0.566
x1	2.0290	0.608	3.335	0.001	0.837	3.221
x2	6.4341	0.562	11.456	0.000	5.333	7.535
x3	-8.5158	4.600	-1.851	0.064	-17.531	0.500
x4	-5.4431	3.343	-1.628	0.103	-11.995	1.109
x5	1.6349	0.679	2.406	0.016	0.303	2.966
x6	-5.1817	1.761	-2.943	0.003	-8.633	-1.731
x7	-5.9673	2.067	-2.887	0.004	-10.018	-1.916
x8	0.0732	0.653	0.112	0.911	-1.207	1.354
x9	-6.6460	2.176	-3.055	0.002	-10.910	-2.382
x10	-0.7114	0.801	-0.888	0.374	-2.281	0.858
x11	-1.7322	1.162	-1.490	0.136	-4.010	0.546
x12	-3.9351	1.550	-2.539	0.011	-6.973	-0.897
x13	2.7451	0.680	4.036	0.000	1.412	4.078
x14	1.1051	0.531	2.080	0.037	0.064	2.146
x15	-4.4955	1.789	-2.513	0.012	-8.001	-0.990

modularity_class=6	coef	std err	z	P>|z|	[0.025	0.975]
const	-11.9342	7.823	-1.526	0.127	-27.266	3.398
x1	-2.6212	22.339	-0.117	0.907	-46.405	41.162
x2	-7.1038	54.016	-0.132	0.895	-112.973	98.765
x3	9.0798	8.038	1.130	0.259	-6.675	24.835
x4	-12.4097	31.909	-0.389	0.697	-74.950	50.131
x5	-10.9790	35.837	-0.306	0.759	-81.219	59.261
x6	9.6110	8.050	1.194	0.233	-6.167	25.389
x7	6.6147	8.557	0.773	0.440	-10.157	23.387
x8	8.4677	8.185	1.035	0.301	-7.574	24.510
x9	14.5211	9.420	1.541	0.123	-3.943	32.985
x10	10.0989	8.459	1.194	0.233	-6.480	26.678
x11	8.4294	9.151	0.921	0.357	-9.506	26.365
x12	-1.9810	14.925	-0.133	0.894	-31.234	27.272
x13	-6.3747	40.635	-0.157	0.875	-86.017	73.268
x14	-15.4774	26.493	-0.584	0.559	-67.402	36.447
x15	-10.9908	35.846	-0.307	0.759	-81.248	59.266

modularity_class=7	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6810	0.320	-2.129	0.033	-1.308	-0.054
x1	1.1065	0.601	1.840	0.066	-0.072	2.285
x2	-7.5992	1.476	-5.148	0.000	-10.492	-4.706
x3	1.0592	0.442	2.395	0.017	0.192	1.926
x4	1.9142	0.659	2.905	0.004	0.623	3.206
x5	1.0297	0.518	1.989	0.047	0.015	2.044
x6	3.2056	0.401	7.997	0.000	2.420	3.991
x7	-3.9954	0.627	-6.374	0.000	-5.224	-2.767
x8	-2.3889	0.610	-3.915	0.000	-3.585	-1.193
x9	-2.0142	0.994	-2.027	0.043	-3.962	-0.066
x10	2.9020	0.470	6.174	0.000	1.981	3.823
x11	-2.1558	0.759	-2.840	0.005	-3.644	-0.668
x12	-0.0888	0.515	-0.172	0.863	-1.099	0.921
x13	-0.3840	0.966	-0.398	0.691	-2.277	1.509
x14	-0.9446	0.620	-1.523	0.128	-2.160	0.271
x15	-2.5957	1.073	-2.419	0.016	-4.699	-0.492

modularity_class=8	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8780	0.435	-4.314	0.000	-2.731	-1.025
x1	0.5180	0.880	0.589	0.556	-1.206	2.242
x2	-3.7650	1.277	-2.949	0.003	-6.267	-1.263
x3	-2.8630	1.066	-2.686	0.007	-4.952	-0.774
x4	1.8405	0.781	2.357	0.018	0.310	3.371
x5	1.2793	0.670	1.911	0.056	-0.033	2.592
x6	-2.8691	0.782	-3.667	0.000	-4.403	-1.336
x7	-0.9426	0.661	-1.427	0.154	-2.238	0.352
x8	-3.2096	0.775	-4.144	0.000	-4.728	-1.692
x9	2.4928	0.949	2.626	0.009	0.632	4.353
x10	-0.2641	0.762	-0.347	0.729	-1.757	1.229
x11	1.2138	0.717	1.693	0.090	-0.192	2.619
x12	6.9989	0.529	13.223	0.000	5.961	8.036
x13	-0.2477	1.248	-0.199	0.843	-2.693	2.198
x14	0.0544	0.739	0.074	0.941	-1.394	1.502
x15	2.2358	0.746	2.997	0.003	0.774	3.698

modularity_class=9	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1976	0.337	-3.559	0.000	-1.857	-0.538
x1	1.5047	0.587	2.562	0.010	0.353	2.656
x2	-1.0364	0.624	-1.661	0.097	-2.259	0.186
x3	-9.6356	2.018	-4.774	0.000	-13.591	-5.680
x4	6.2116	0.587	10.574	0.000	5.060	7.363
x5	1.5385	0.519	2.966	0.003	0.522	2.555
x6	-3.0438	0.668	-4.557	0.000	-4.353	-1.735
x7	-1.0438	0.522	-2.001	0.045	-2.066	-0.022
x8	-0.9698	0.505	-1.921	0.055	-1.959	0.019
x9	2.0104	0.815	2.466	0.014	0.413	3.608
x10	1.5036	0.527	2.851	0.004	0.470	2.537
x11	2.3801	0.527	4.519	0.000	1.348	3.413
x12	3.0379	0.463	6.561	0.000	2.130	3.945
x13	0.7256	0.794	0.914	0.361	-0.831	2.282
x14	0.0662	0.543	0.122	0.903	-0.997	1.130
x15	-0.2310	0.752	-0.307	0.759	-1.705	1.243

modularity_class=14	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5644	0.355	1.589	0.112	-0.132	1.260
x1	-3.7279	1.853	-2.012	0.044	-7.359	-0.097
x2	-8.5534	2.413	-3.545	0.000	-13.283	-3.824
x3	-3.3796	0.879	-3.845	0.000	-5.102	-1.657
x4	-2.8012	1.798	-1.558	0.119	-6.324	0.722
x5	-1.9571	1.034	-1.892	0.058	-3.985	0.070
x6	-18.1930	5.713	-3.184	0.001	-29.391	-6.995
x7	-6.7473	1.676	-4.025	0.000	-10.032	-3.462
x8	-11.5712	1.522	-7.603	0.000	-14.554	-8.588
x9	-6.0157	2.077	-2.896	0.004	-10.087	-1.945
x10	-6.2216	1.640	-3.794	0.000	-9.435	-3.008
x11	-1.3638	0.772	-1.767	0.077	-2.877	0.149
x12	-7.4590	1.555	-4.796	0.000	-10.507	-4.411
x13	-3.4903	1.528	-2.284	0.022	-6.485	-0.496
x14	-3.0151	1.164	-2.591	0.010	-5.296	-0.734
x15	5.0272	0.658	7.644	0.000	3.738	6.316

modularity_class=16	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.8645	1.773	-2.743	0.006	-8.340	-1.389
x1	-10.4921	20.981	-0.500	0.617	-51.613	30.629
x2	1.7512	3.491	0.502	0.616	-5.092	8.594
x3	-13.2258	12.969	-1.020	0.308	-38.644	12.193
x4	4.5571	2.264	2.013	0.044	0.119	8.995
x5	0.0901	3.348	0.027	0.979	-6.472	6.652
x6	-23.8178	15.294	-1.557	0.119	-53.793	6.157
x7	-2.0672	5.113	-0.404	0.686	-12.088	7.954
x8	6.4421	1.845	3.492	0.000	2.827	10.057
x9	-6.7990	19.426	-0.350	0.726	-44.873	31.275
x10	2.8903	3.360	0.860	0.390	-3.696	9.477
x11	1.4075	2.609	0.540	0.589	-3.705	6.520
x12	-7.7238	5.426	-1.423	0.155	-18.359	2.911
x13	-2.4654	7.634	-0.323	0.747	-17.427	12.496
x14	-16.5993	24.340	-0.682	0.495	-64.305	31.107
x15	-8.0503	4.245	-1.897	0.058	-16.370	0.269

modularity_class=18	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6340	0.367	-1.725	0.084	-1.354	0.086
x1	0.6995	0.664	1.053	0.292	-0.603	2.002
x2	-17.7381	5.979	-2.967	0.003	-29.457	-6.019
x3	-3.1428	1.035	-3.038	0.002	-5.171	-1.115
x4	-0.3620	1.327	-0.273	0.785	-2.962	2.238
x5	-0.3844	0.647	-0.594	0.552	-1.652	0.883
x6	-9.8974	1.832	-5.403	0.000	-13.488	-6.307
x7	-1.8675	0.617	-3.026	0.002	-3.077	-0.658
x8	-7.7868	2.403	-3.240	0.001	-12.497	-3.077
x9	9.3285	0.751	12.429	0.000	7.857	10.799
x10	-0.3501	0.647	-0.541	0.588	-1.618	0.918
x11	-0.9629	0.788	-1.222	0.222	-2.507	0.581
x12	-5.1759	1.224	-4.228	0.000	-7.575	-2.776
x13	-0.1205	0.968	-0.124	0.901	-2.018	1.777
x14	-0.1170	0.587	-0.199	0.842	-1.268	1.034
x15	-2.0630	1.270	-1.624	0.104	-4.553	0.427

modularity_class=21	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2315	0.543	-2.267	0.023	-2.296	-0.167
x1	-0.1207	1.226	-0.099	0.922	-2.523	2.281
x2	-23.8772	13.668	-1.747	0.081	-50.665	2.911
x3	-3.3977	1.592	-2.135	0.033	-6.517	-0.278
x4	-10.2240	6.242	-1.638	0.101	-22.458	2.010
x5	-0.9619	1.161	-0.829	0.407	-3.237	1.314
x6	-1.4057	0.952	-1.477	0.140	-3.271	0.460
x7	-0.9030	0.780	-1.157	0.247	-2.432	0.626
x8	-12.3551	6.584	-1.876	0.061	-25.260	0.550
x9	-4.2585	2.009	-2.119	0.034	-8.197	-0.320
x10	1.3547	0.779	1.738	0.082	-0.173	2.882
x11	-10.1081	3.553	-2.845	0.004	-17.071	-3.145
x12	-1.1629	1.155	-1.006	0.314	-3.427	1.102
x13	1.8575	1.176	1.580	0.114	-0.447	4.162
x14	1.0460	0.804	1.301	0.193	-0.530	2.622
x15	-1.0974	1.899	-0.578	0.563	-4.820	2.625

Annexe 29 : tableau récapitulatif des coefficients de corrélation de la régression multinomiale – modèle brut

MNLogit Regression Results

Dep. Variable:	modularity_class	No. Observations:	5403
Model:	MNLogit	Df Residuals:	5283
Method:	MLE	Df Model:	114
Date:	Thu, 03 Apr 2025	Pseudo R-squ.:	0.5249
Time:	13:51:57	Log-Likelihood:	-4005.7
converged:	True	LL-Null:	-8430.6
Covariance Type:	nonrobust	LLR p-value:	0.000

modularity_class=1	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5999	10.225	-0.156	0.876	-21.640	18.440
Topic_0	1.3942	10.288	0.136	0.892	-18.769	21.558
Topic_1	-7.8534	10.320	-0.761	0.447	-28.080	12.374
Topic_2	-8.6512	10.427	-0.830	0.407	-29.088	11.786
Topic_3	1.5958	10.292	0.155	0.877	-18.577	21.768
Topic_4	-0.1543	10.317	-0.015	0.988	-20.375	20.066
Topic_5	-0.4601	10.302	-0.045	0.964	-20.653	19.732
Topic_6	-5.3110	10.288	-0.516	0.606	-25.476	14.854
Topic_7	-0.1751	10.295	-0.017	0.986	-20.353	20.003
Topic_8	1.5531	10.277	0.151	0.880	-18.589	21.695
Topic_9	1.3233	10.291	0.129	0.898	-18.846	21.493
Topic_10	1.1710	10.312	0.114	0.910	-19.040	21.382
Topic_11	-0.1344	10.328	-0.013	0.990	-20.376	20.108
Topic_12	-2.7542	10.343	-0.266	0.790	-23.026	17.517
Topic_13	6.6273	10.286	0.644	0.519	-13.533	26.788
Topic_14	-1.4376	10.292	-0.140	0.889	-21.609	18.734
Topic_15	0.9558	10.283	0.093	0.926	-19.198	21.109
Topic_16	1.2639	10.293	0.123	0.902	-18.911	21.438
Topic_17	1.6605	10.302	0.161	0.872	-18.531	21.852
Topic_18	1.3948	10.284	0.136	0.892	-18.760	21.550

modularity_class=2	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6650	12.555	-0.212	0.832	-27.273	21.943
Topic_0	0.1109	12.639	0.009	0.993	-24.661	24.883
Topic_1	-6.2375	12.677	-0.492	0.623	-31.084	18.609
Topic_2	-2.7940	12.628	-0.221	0.825	-27.545	21.957
Topic_3	1.4152	12.619	0.112	0.911	-23.319	26.149
Topic_4	1.1421	12.661	0.090	0.928	-23.673	25.958
Topic_5	-0.0058	12.672	-0.000	1.000	-24.843	24.831
Topic_6	-4.1466	12.629	-0.328	0.743	-28.899	20.606
Topic_7	0.3024	12.647	0.024	0.981	-24.486	25.091
Topic_8	-1.9390	12.608	-0.154	0.878	-26.650	22.772
Topic_9	3.9625	12.628	0.314	0.754	-20.789	28.714
Topic_10	0.3725	12.666	0.029	0.977	-24.452	25.197
Topic_11	-1.1123	12.675	-0.088	0.930	-25.955	23.730
Topic_12	0.8868	12.650	0.070	0.944	-23.907	25.680
Topic_13	2.0355	12.627	0.161	0.872	-22.712	26.783
Topic_14	7.5232	12.627	0.596	0.551	-17.226	32.272
Topic_15	-0.4085	12.669	-0.032	0.974	-25.239	24.422
Topic_16	0.4932	12.646	0.039	0.969	-24.292	25.278
Topic_17	-0.0451	12.690	-0.004	0.997	-24.917	24.827
Topic_18	2.7110	12.626	0.215	0.830	-22.035	27.457

modularity_class=3	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1593	9.864	-0.219	0.827	-21.492	17.173
Topic_0	1.3464	9.930	0.136	0.892	-18.115	20.808
Topic_1	-3.5906	9.925	-0.362	0.718	-23.043	15.862
Topic_2	-9.5013	10.048	-0.946	0.344	-29.196	10.193
Topic_3	5.9260	9.922	0.597	0.550	-13.520	25.372
Topic_4	1.7342	9.947	0.174	0.862	-17.761	21.229
Topic_5	0.2566	9.945	0.026	0.979	-19.235	19.748
Topic_6	-4.0596	9.918	-0.409	0.682	-23.498	15.379
Topic_7	0.3286	9.934	0.033	0.974	-19.142	19.799
Topic_8	-0.1300	9.912	-0.013	0.990	-19.557	19.297
Topic_9	3.5774	9.929	0.360	0.719	-15.883	23.038
Topic_10	0.8948	9.949	0.090	0.928	-18.604	20.394
Topic_11	0.6402	9.955	0.064	0.949	-18.871	20.151
Topic_12	-0.5499	9.950	-0.055	0.956	-20.052	18.952
Topic_13	2.5634	9.924	0.258	0.796	-16.887	22.014
Topic_14	3.8141	9.920	0.384	0.701	-15.628	23.256
Topic_15	0.1468	9.931	0.015	0.988	-19.318	19.612
Topic_16	0.8531	9.938	0.086	0.932	-18.625	20.331
Topic_17	1.8864	9.943	0.190	0.850	-17.602	21.375
Topic_18	0.7299	9.934	0.073	0.941	-18.739	20.199

modularity_class=4	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.8112	19.504	-0.195	0.845	-42.039	34.416
Topic_0	-0.0970	19.671	-0.005	0.996	-38.652	38.457
Topic_1	-7.0040	19.704	-0.355	0.722	-45.624	31.616
Topic_2	0.4211	19.589	0.021	0.983	-37.973	38.816
Topic_3	0.2807	19.758	0.014	0.989	-38.444	39.006
Topic_4	1.4701	19.676	0.075	0.940	-37.094	40.034
Topic_5	1.6240	19.652	0.083	0.934	-36.893	40.141
Topic_6	-14.8849	20.121	-0.740	0.459	-54.320	24.551
Topic_7	-1.8276	19.718	-0.093	0.926	-40.473	36.818
Topic_8	-5.9692	19.598	-0.305	0.761	-44.381	32.442
Topic_9	-0.9447	19.537	-0.048	0.961	-39.236	37.347
Topic_10	2.0670	19.700	0.105	0.916	-36.545	40.679
Topic_11	-2.7843	19.805	-0.141	0.888	-41.602	36.033
Topic_12	5.2337	19.661	0.266	0.790	-33.301	43.769
Topic_13	3.1388	19.604	0.160	0.873	-35.284	41.561
Topic_14	-3.2838	19.635	-0.167	0.867	-41.768	35.201
Topic_15	-0.5980	19.679	-0.030	0.976	-39.168	37.972
Topic_16	0.8201	19.675	0.042	0.967	-37.742	39.382
Topic_17	4.8126	19.625	0.245	0.806	-33.651	43.276
Topic_18	9.0801	19.615	0.463	0.643	-29.365	47.526

modularity_class=5	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7347	13.803	-0.198	0.843	-29.788	24.318
Topic_0	1.6754	13.885	0.121	0.904	-25.538	28.889
Topic_1	-18.5673	15.170	-1.224	0.221	-48.300	11.165
Topic_2	-1.7310	13.899	-0.125	0.901	-28.973	25.511
Topic_3	0.5151	13.955	0.037	0.971	-26.837	27.867
Topic_4	0.7996	13.928	0.057	0.954	-26.500	28.099
Topic_5	1.3315	13.899	0.096	0.924	-25.911	28.574
Topic_6	-9.5785	13.913	-0.688	0.491	-36.847	17.690
Topic_7	0.5990	13.895	0.043	0.966	-26.634	27.832
Topic_8	-4.6455	13.965	-0.333	0.739	-32.017	22.726
Topic_9	12.0350	13.887	0.867	0.386	-15.183	39.253
Topic_10	1.0015	13.910	0.072	0.943	-26.261	28.264
Topic_11	-0.0388	13.939	-0.003	0.998	-27.359	27.281
Topic_12	1.5782	13.906	0.113	0.910	-25.677	28.833
Topic_13	1.1356	13.885	0.082	0.935	-26.079	28.350
Topic_14	-3.3158	13.928	-0.238	0.812	-30.615	23.983
Topic_15	0.5039	13.881	0.036	0.971	-26.703	27.710
Topic_16	1.6687	13.891	0.120	0.904	-25.557	28.895
Topic_17	3.0139	13.908	0.217	0.828	-24.246	30.274
Topic_18	-0.4721	13.905	-0.034	0.973	-27.726	26.782

modularity_class=6	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.9910	20.777	-0.240	0.810	-45.713	35.731
Topic_0	2.4893	20.913	0.119	0.905	-38.498	43.477
Topic_1	-18.7225	22.928	-0.817	0.414	-63.661	26.216
Topic_2	-0.2980	20.910	-0.014	0.989	-41.280	40.684
Topic_3	-6.8328	21.766	-0.314	0.754	-49.494	35.829
Topic_4	1.8955	20.983	0.090	0.928	-39.230	43.021
Topic_5	3.5878	20.896	0.172	0.864	-37.368	44.543
Topic_6	0.0579	20.893	0.003	0.998	-40.892	41.007
Topic_7	3.3340	20.908	0.159	0.873	-37.644	44.312
Topic_8	-7.6767	21.745	-0.353	0.724	-50.295	34.942
Topic_9	0.2303	20.985	0.011	0.991	-40.899	41.359
Topic_10	3.2456	20.959	0.155	0.877	-37.833	44.324
Topic_11	3.2273	20.947	0.154	0.878	-37.829	44.284
Topic_12	1.4967	21.025	0.071	0.943	-39.711	42.704
Topic_13	-6.2331	21.263	-0.293	0.769	-47.908	35.442
Topic_14	2.4459	20.912	0.117	0.907	-38.541	43.433
Topic_15	4.1845	20.878	0.200	0.841	-36.735	45.104
Topic_16	4.5117	20.881	0.216	0.829	-36.415	45.438
Topic_17	4.3097	20.894	0.206	0.837	-36.643	45.262
Topic_18	2.1923	21.000	0.104	0.917	-38.967	43.352

Annexe 30 : tableau récapitulatif des coefficients de corrélation de la régression multinomiale – modèle avec agrégation de classe de modularité

MNLogit Regression Results

Dep. Variable:	modularity_class	No. Observations:	5403
Model:	MNLogit	Df Residuals:	5307
Method:	MLE	Df Model:	90
Date:	Thu, 03 Apr 2025	Pseudo R-squ.:	0.5215
Time:	15:26:58	Log-Likelihood:	-4034.2
converged:	True	LL-Null:	-8430.6
Covariance Type:	nonrobust	LLR p-value:	0.000

modularity_class=1	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9851	0.276	-3.564	0.000	-1.527	-0.443
x1	0.7345	0.468	1.569	0.117	-0.183	1.652
x2	-8.5883	0.862	-9.958	0.000	-10.279	-6.898
x3	-9.6100	2.205	-4.358	0.000	-13.932	-5.288
x4	0.9572	0.554	1.729	0.084	-0.128	2.042
x5	-0.8336	0.538	-1.549	0.121	-1.889	0.221
x6	-5.9029	0.774	-7.629	0.000	-7.419	-4.386
x7	-0.7819	0.499	-1.567	0.117	-1.760	0.196
x8	0.8593	0.353	2.438	0.015	0.168	1.550
x9	0.7194	0.832	0.865	0.387	-0.910	2.349
x10	-0.6817	0.444	-1.534	0.125	-1.553	0.189
x11	6.0512	0.436	13.887	0.000	5.197	6.905
x12	-2.1622	0.633	-3.417	0.001	-3.403	-0.922
x13	0.6507	0.613	1.062	0.288	-0.550	1.852
x14	0.6312	0.448	1.408	0.159	-0.247	1.510
x15	0.9887	0.563	1.755	0.079	-0.116	2.093

modularity_class=2	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3026	0.424	-5.431	0.000	-3.134	-1.472
x1	-0.2460	0.833	-0.295	0.768	-1.879	1.387
x2	-6.5691	1.211	-5.424	0.000	-8.943	-4.195
x3	-3.1685	1.047	-3.025	0.002	-5.221	-1.116
x4	1.0857	0.693	1.567	0.117	-0.273	2.444
x5	0.7799	0.630	1.239	0.215	-0.454	2.014
x6	-4.5015	0.750	-6.004	0.000	-5.971	-3.032
x7	-0.0623	0.643	-0.097	0.923	-1.323	1.198
x8	-2.2437	0.743	-3.019	0.003	-3.700	-0.787
x9	3.5661	0.913	3.905	0.000	1.776	5.356
x10	-1.5417	0.722	-2.137	0.033	-2.956	-0.128
x11	1.6515	0.706	2.341	0.019	0.269	3.034
x12	7.1463	0.500	14.281	0.000	6.166	8.127
x13	-0.9945	1.184	-0.840	0.401	-3.316	1.326
x14	0.1442	0.714	0.202	0.840	-1.255	1.543
x15	2.3103	0.709	3.259	0.001	0.921	3.700

modularity_class=3	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3936	0.313	-4.459	0.000	-2.006	-0.781
x1	0.5589	0.509	1.099	0.272	-0.438	1.556
x2	-4.3929	0.560	-7.841	0.000	-5.491	-3.295
x3	-10.2183	2.046	-4.994	0.000	-14.229	-6.208
x4	5.1568	0.459	11.240	0.000	4.258	6.056
x5	0.9440	0.455	2.076	0.038	0.053	1.835
x6	-4.8397	0.625	-7.741	0.000	-6.065	-3.614
x7	-0.4336	0.500	-0.867	0.386	-1.414	0.546
x8	-0.9469	0.482	-1.964	0.050	-1.892	-0.002
x9	2.8476	0.766	3.716	0.000	1.346	4.349
x10	-0.0651	0.465	-0.140	0.889	-0.977	0.847
x11	1.8414	0.526	3.500	0.000	0.810	2.873
x12	3.0202	0.422	7.160	0.000	2.193	3.847
x13	-0.3374	0.699	-0.483	0.629	-1.707	1.032
x14	0.0812	0.498	0.163	0.870	-0.895	1.057
x15	0.0905	0.710	0.127	0.899	-1.301	1.482

modularity_class=4	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1694	0.337	0.503	0.615	-0.491	0.830
x1	-4.3957	1.767	-2.488	0.013	-7.859	-0.933
x2	-11.5911	2.407	-4.815	0.000	-16.309	-6.873
x3	-3.6672	0.834	-4.399	0.000	-5.301	-2.033
x4	-3.3239	1.658	-2.005	0.045	-6.574	-0.074
x5	-2.6382	1.021	-2.583	0.010	-4.640	-0.637
x6	-19.3678	5.655	-3.425	0.001	-30.452	-8.284
x7	-5.9496	1.645	-3.616	0.000	-9.174	-2.725
x8	-9.2264	1.418	-6.506	0.000	-12.006	-6.447
x9	-4.7349	2.018	-2.347	0.019	-8.690	-0.780
x10	-7.3507	1.592	-4.618	0.000	-10.471	-4.231
x11	-1.0908	0.763	-1.430	0.153	-2.586	0.404
x12	-6.9017	1.496	-4.612	0.000	-9.834	-3.969
x13	-4.3200	1.492	-2.896	0.004	-7.244	-1.396
x14	-3.0113	1.145	-2.630	0.009	-5.256	-0.767
x15	4.7472	0.592	8.013	0.000	3.586	5.908

modularity_class=5	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0051	0.350	-2.875	0.004	-1.690	-0.320
x1	-0.0812	0.602	-0.135	0.893	-1.261	1.098
x2	-20.6762	6.055	-3.415	0.001	-32.543	-8.809
x3	-3.5489	1.017	-3.491	0.000	-5.541	-1.556
x4	-1.2357	1.280	-0.966	0.334	-3.744	1.272
x5	-0.9294	0.606	-1.534	0.125	-2.117	0.258
x6	-11.3204	1.811	-6.250	0.000	-14.870	-7.771
x7	-1.1044	0.598	-1.847	0.065	-2.277	0.068
x8	-6.3550	2.214	-2.870	0.004	-10.695	-2.015
x9	10.2995	0.708	14.544	0.000	8.912	11.687
x10	-1.7094	0.601	-2.845	0.004	-2.887	-0.532
x11	-0.5539	0.773	-0.716	0.474	-2.070	0.962
x12	-5.2021	1.228	-4.237	0.000	-7.608	-2.796
x13	-0.8329	0.906	-0.919	0.358	-2.609	0.943
x14	-0.0741	0.559	-0.132	0.895	-1.170	1.022
x15	-2.1643	1.231	-1.758	0.079	-4.577	0.249

modularity_class=6	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5522	0.522	-2.975	0.003	-2.575	-0.529
x1	-0.9354	1.172	-0.798	0.425	-3.233	1.362
x2	-20.7012	7.767	-2.665	0.008	-35.924	-5.478
x3	-3.8505	1.565	-2.460	0.014	-6.919	-0.782
x4	-9.3410	5.640	-1.656	0.098	-20.396	1.714
x5	-1.5877	1.126	-1.410	0.158	-3.794	0.619
x6	-3.4240	0.905	-3.782	0.000	-5.199	-1.650
x7	-0.1436	0.754	-0.190	0.849	-1.621	1.333
x8	-10.4149	5.952	-1.750	0.080	-22.080	1.251
x9	-3.1831	1.975	-1.612	0.107	-7.053	0.687
x10	-0.1721	0.725	-0.237	0.812	-1.593	1.248
x11	-10.4560	3.751	-2.788	0.005	-17.807	-3.105
x12	-1.0782	1.130	-0.954	0.340	-3.294	1.137
x13	0.9800	1.111	0.882	0.378	-1.197	3.157
x14	1.0593	0.769	1.378	0.168	-0.447	2.566
x15	-1.2607	1.873	-0.673	0.501	-4.931	2.410

Annexe 31 : tableau récapitulatif des coefficients de corrélation de la régression multinomiale – modèle avec agrégation de classe de modularité et sélection de variables

Sujets aprés selection de variables		
Nom de variable	**Numéro du sujet d'origine**	**Nom du Sujet**
X1	0	**genou**
X2	1	**prestation de handicape**
X3	2	**Frais de soins**
X4	3	**Représentation**
X5	4	**Condition Psychologique**
X6	6	**Indemnité de remplacement de revenu**
X7	7	**Capacité de retour**
X8	8	**Partage de prestation et responsabilité**
X9	9	**Rechute**
X10	11	**Entorse cervicale**
X11	13	**Causalité de la blessure**
X12	14	**Délai de contestation**
X13	15	**Déchirure de la coiffe des rotateurs**
X14	16	**Blessure au dos**
X15	18	**Maladie de l'audition**

Annexe 32 : Tableau récapitulatif des variables du modèle de régression avec agrégation des classes de modularité et sélection de variables

Bibliographie

Akbik, Alan, Tanja Bergmann, Roland Vollgraf. (2019). «*Pooled Contextualized Embeddings for Named Entity Recognition*. » Dans Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 724–728. Association for Computational Linguistics. Récupéré de https://aclanthology.org/N19-1078.

Bauer Emmanuel., Stammbach Dominik, Gu Nianlong, Ash Eliott. (2023). «*Legal Extractive Summarization of U.S. Court Opinions*». Récupéré de https://arxiv.org/abs/2305.08428

Basu Poulami, Dinabandhu Bhandari, Koustav Sengupta (2024). « Using Machine Learning Algorithms With TF-IDF to Generate Legal Petitions », International Journal of Engineering Research and Technology (IJERT), février.

Benveniste Émile (1964). « *Ferdinand de Saussure à l'École des Hautes Études* ». Annuaires de l'École pratique des hautes études, 72(1964), 20–34.

Bermylle John U. Razon, Geoffrey A. Solano, Lorenz Timothy B. Ranera (2022). « *Topic Modelling Supreme Court Case Decisions using Latent Dirichlet Allocation* », *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 284–289. Récupéré de https://doi.org/10.1109/ICTC55196.2022.9952945.

Blei David, Ng Andrew, Jordan Michael (2003). «*Latent Dirichlet Allocation*». Journal of Machine Learning Research, 3(4-5), 993–1022. Récupéré de https://doi.org/10.1162/jmlr.2003.3.4-5.993.

Blondel D. Vincent, Guillaume, Jean-Loup., Lambiotte, Renaud, Lefebvre, Etienne (2008). «*Fast unfolding of communities in large networks*». Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. Récupéré de https://doi.org/10.1088/1742-5468/2008/10/p10008

Breton-Demeule Charles. (2021). « *La jurisprudence* ». Chaire de rédaction juridique Louis-Philippe-Pigeon, Université Laval.

Budahazy Ryan, Lu Cheng, Yihuan Huang, Andrew Johnson, Pengyu Li, Joshua Vendrow, Zhoutong Wu, Denali Molitor, Elizaveta Rebrova, Deanna Needell (2021). « Analysis of Legal Documents via Non-negative Matrix Factorization Methods », arXiv:2104.14028 [cs.LG]. Récupéré de https://doi.org/10.48550/arXiv.2104.14028.

Celard Pedro, Lorenzo Iglesias Eva, Sorribes-Fdez, Jose Manuel, Romero, Ruben, Seara Vieira Adrian, Borrajo, Lourdes (2024). «*Modified LDA vector and feedback analysis for short query Information Retrieval systems*». Logic Journal of the IGPL, 32(1), jzae044. Récupéré de https://doi.org/10.1093/jigpal/jzae044.

Chalkidis Ilias, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, Nikolaos Aletras (2022). « *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English* », dans Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, p. 4310–4330, Dublin, Ireland. Récupéré de https://doi.org/10.18653/v1/2022.acl-long.297.

Chalkidis Ilias., Fergadiotis Manos, Malakasiotis Prodromos, Aletras Nikolaos, Androutsopoulos, Ion (2020). « *LEGAL-BERT: The Muppets straight out of Law School* ». Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2898–2904). Association for Computational Linguistics. Récupéré de https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chen Wenchao, Chen Bo, Liu Yingpi, Cao Xuefei, Zhao Qianru, Zhang Hao, Tian Long (2022). «M*ax-Margin Deep Diverse Latent Dirichlet Allocation With Continual Learning*». IEEE Transactions on Cybernetics, 52(7), 5639–5653. Récupéré de https://doi.org/10.1109/TCYB.2020.3044915

Chen, Yubo, Liheng Xu, Kang Liu, Daojian Zeng, Jun Zhao (2015). « *Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks* », dans Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, p. 167–176, Beijing, China. Récupéré de https://doi.org/10.3115/v1/P15-1017.

Christopoulou Fenia, Makoto Miwa, Sophia Ananiadou (2018). « *A Walk-Based Model on Entity Graphs for Relation Extraction* », dans Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, p. 81–88, Melbourne, Australia. Récupéré de https://doi.org/10.18653/v1/P18-2014.

Clark John, Holton Derrek Allan (1991). «*A First Look at Graph Theory*». World Scientific Publishing Co. Pte. Ltd.

CNESST. (2023). « *Orientations en imputation* ». ISBN 978-2-550-94667-0. Récupéré de https://www.cnesst.gouv.qc.ca, consulté le 20 février 2025.

CNESST. (2024), « *indemnité de remplacement de revenu* ». Démarches et formulaires, travailleuses et travailleurs, indemnité et remboursement, indemnité. Récupéré de https://www.cnesst.gouv.qc.ca/, consulté le 15 décembre 2024.

Deroy Aniket, Kripabandhu Ghosh, Saptarshi Ghosh (2024). « Applicability of Large Language Models and Generative Models for Legal Case Judgement Summarization », Artificial Intelligence and Law, Springer. Récupéré de https://doi.org/10.48550/arXiv.2407.12848 .

Devaraj Pranav Nataraj, Rakesh Teja P V, Manoj Kumar R (2023). « *Development of a Legal Document* AIChatbot », Vellore Institute of Technology, Chennai, India.

Didwania Krish, Durga Toshniwal, Amit Agarwal (2024). « *Unveiling Themes in Judicial Proceedings: A Cross-Country Study Using Topic Modeling on Legal Documents from India and the UK* », arXiv:2406.00040 [cs.CL]. Récupéré de https://doi.org/10.48550/arXiv.2406.00040.

Eto Masaki (2019). «*Extended co-citation search: Graph-based document retrieval on a co-citation network containing citation context information*». Information Processing & Management, 56(6), 102046. Récupéré de https://doi.org/10.1016/j.ipm.2019.05.007.

Farea Amer, Tripathi Shailesh, Glazko Galina, Emmert-Streib Frank (2024). «*Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research*». Engineering Applications of Artificial Intelligence, 136(Part A), 108877. Récupéré de https://doi.org/10.1016/j.engappai.2024.108877.

Gardner Anne Von Der Lieth (1984). An artificial intelligence approach to legal reasoning, Stanford University, Stanford, CA, USA.

Gerasimenko Nikolai, Chernyavskiy, Alexander, Nikiforova Maria, Ianina Anastasia, Vorontsov Konstantin (2023). «*Incremental Topic Modeling for Scientific Trend Topics Extraction*». Récupéré de https://doi.org/10.28995/2075-7182-2023-22-88-103

Gouvernement du Canada. (2022). « *Rapport annuel 2022 – Accidents du travail dans les secteurs relevant de la compétence fédérale* ». Récupéré de https://www.canada.ca/fr/emploi-developpement-social/services/sante-securite/rapports/2022-accidents.html.

Greco Candida M., Aandrea Tagarelli (2024). « *Evaluating GPT-3.5's Awareness and Summarization Abilities for European Constitutional Texts with Shared Topics* », arXiv:2401.14524 [cs.CL]. Récupéré de https://doi.org/10.48550/arXiv.2401.14524

Guha, Neel, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, Zehua Li (2023). « *LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models* », dans Proceedings of the 37th International Conference on Neural Information Processing Systems, Curran Associates Inc., p. 44123–44279, Red Hook, NY, USA. arXiv preprint arXiv:2308.11462.

Hamiter Mickie, (2024). « *Perte auditive* ». Manuel Merck, Version pour les patients. Récupéré de https://www.merckmanuals.com

Huang, Zihan, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E. Ho, Mark S. Krass, Matthias Grabmair. (2021). «*Context-Aware Legal Citation Recommendation Using Deep Learning*», Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL 2021), 10 pages. Récupéré de https://doi.org/10.48550/arXiv.2106.10776.

Khazaeli, Soha, Janardhana Punuru, Chad Morris, Sanjay Sharma, Bert Staub, Michael Cole, Sunny Chiu-Webster,Dhruv Sakalley. (2021). «A Free Format Legal Question Answering System», Dans Proceedings of the Natural Legal Language Processing Workshop 2021, 107–113. Association for Computational Linguistics. Récupéré de https://aclanthology.org/2021.nllp-1.11.

Kort Fred (1957). « Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the 'Right to Counsel' Cases », American Political Science Review, vol. 51, no 1, p. 1–12. Récupéré de https://doi.org/10.2307/1951767

Kuru Onur, Can Ozan Arkan, Yuret Deniz (2016). «CharNER: Character-Level Named Entity Recognition». In Y. Matsumoto & R. Prasad (Eds.), Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 911–921). Osaka, Japan: The COLING 2016 Organizing Committee. Récupéré de https://aclanthology.org/C16-1087/

Lee C. Y. (1961). «*An Algorithm for Path Connections and Its Applications. IRE Transactions on Electronic Computers*», EC-10(3), 346–365. Récupéré de https://doi.org/10.1109/TEC.1961.5219222.

Lluelles Didier, Ringuette Josée (2014). « *Guide des références pour la rédaction juridique (8e éd.)*». ISBN: 978-2-89400-344-2.

Liu Hanwen, Kou Huazizhen, Yan Chao, Qi Lianyong (2019). « *Link prediction in paper citation network to construct paper correlation graph* ». EURASIP Journal on Wireless Communications and Networking, 2019. Récupéré de https://doi.org/10.1186/s13638-019-1561-7

Mahari Robert Zev (2021). «*AutoLAW: Augmented Legal Reasoning through Legal Precedent Prediction*». Récupéré de https://doi.org/10.48550/arXiv.2106.16034.

Mahari Robert, Dominik Stammbach, Elliott Ash, Alex Pentland (2023). « The Law and NLP: Bridging Disciplinary Disconnects », dans Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, p. 3445–3454, Singapore.

Mimno David, Wallach Hanna, Talley Edmund, Leenders Miriam, McCallum Andrew (2011). «*Optimizing Semantic Coherence in Topic Models*». Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 262–272). Association for Computational Linguistics. Récupéré de https://aclanthology.org/D11-1024/.

Moreno-Sánchez Isabel, Font-Clos Francesc, Corral Álvaro (2016). « *Large-Scale Analysis of Zipf's Law in English Texts*». PLOS ONE, 11(1), e0147073. Récupéré de https://doi.org/10.1371/journal.pone.0147073.

Nadkarni Prakash M., Lucila Ohno-Machado, Wendy W. Chapman (2011). « *Natural language processing: an introduction* », Journal of the American Medical Informatics Association, vol. 18, no 5, p. 544–551. Récupéré de https://doi.org/10.1136/amiajnl-2011-000464

Nagel Stuart (1965). « *Predicting Court Cases Quantitatively* », Michigan Law Review, vol. 63, no 8, p. 1411. Récupéré de https://repository.law.umich.edu/mlr/vol63/iss8/6

Nguyen Thien Huu, Ralph Grishman (2018). « *Graph Convolutional Networks with Argument-Aware Pooling for Event Detection* », Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI Press, New Orleans, Louisiana, USA, Article 724. ISBN 978-1-57735-800-8.

Lockard Katherine, Robert Slater, Brandon Sucrese (2023). Using NLP to Model U.S. «*Supreme Court Cases*», SMU Data Science Review, vol. 7, no 1, article 4. Récupéré de https://scholar.smu.edu/datasciencereview/vol7/iss1/4.

Pand Banoj Kumar, Madhulita Mohapatra (2019). « *Creating Legal Petitions Using TF-IDF and Machine Learning Algorithms* », Dogo Rangsang Research Journal, vol. 9, no 1, janvier-avril. ISSN 2347-7180.

Shounak Paul, Pawan Goyal, Saptarshi Gosh (2021). «*LeSICiN: A Heterogeneous Graph-based Approach for Automatic Legal Statute Identification from Indian Legal Documents*». arXiv preprint arXiv: 2112.14731. Récupéré de https://arxiv.org/abs/2112.14731

Rice Dan, Rhodes Jesse H., Nteta Tatishe (2019). «*Racial bias in legal language, Research & Politics*», vol. 6, no 2. Récupéré de https://doi.org/10.1177/2053168019848930.

Röder Michael, Both Aandreas, Hinneburg Alexander (2015). «Exploring the Space of Topic Coherence Measures». Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (pp. 399–408). Association for Computing Machinery. Récupéré de https://doi.org/10.1145/2684822.2685324.

Saura Jose Ramon (2019). «*A Three-Stage Methodological Process of Data Text Mining: A UGC Business Intelligence Analysis*». Récupéré de https://doi.org/10.13140/RG.2.2.11093.06880

Segal Jeffrey A. (1984). « *Predicting Supreme Court Cases Probabilistically: The Search and Seizure Cases, 1962-1981* », American Political Science Review, vol. 78, no 4, p. 891–900. Récupéré de https://doi.org/10.2307/1955796

Syed, Shaheen, Spruit Marco (2017). «*Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation*». In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 165–174). Récupéré de https://doi.org/10.1109/DSAA.2017.61.

Tang Li, Simon Clematide. (2021). «*Searching for Legal Documents at Paragraph Level: Automating Label Generation and Use of an Extended Attention Mask for Boosting Neural Models of Semantic Similarity*», Proceedings of the Natural Legal Language Processing Workshop (NLLP 2021), 114–122. Récupéré de https://doi.org/10.18653/v1/2021.nllp-1.12.

Tang Yanran, Qiu Ruihong, Liu Yilun, Li Xue, Huang Zi (2023). «*CaseGNN: Graph Neural Networks for Legal Case Retrieval with Text-Attributed Graphs*». arXiv preprint arXiv:2312.11229. Récupéré de https://arxiv.org/abs/2312.11229

Thomson Reuters (2022). *2022 legal Departement Operations (LDO) Index, 7th Annual Edition*. Version mise à jour le 13 octobre 2022, Toronto, Récupéré de https://www.thomsonreuters.com/en/press-releases.

S Sidney Ulmer. (1963).« *Quantitative analysis of judicial processes: Some practical and theoretical applications*». Law and Contemporary Problems, 28, 164.

Uray Martina, Skočaj Danijel, Roth Peter M., Bischof Horst, Leonardis Ales (2007). «*Incremental LDA Learning by Combining Reconstructive and Discriminative Approaches*». Proceedings of the British Machine Vision Conference. Récupéré de https://doi.org/10.5244/C.21.44.

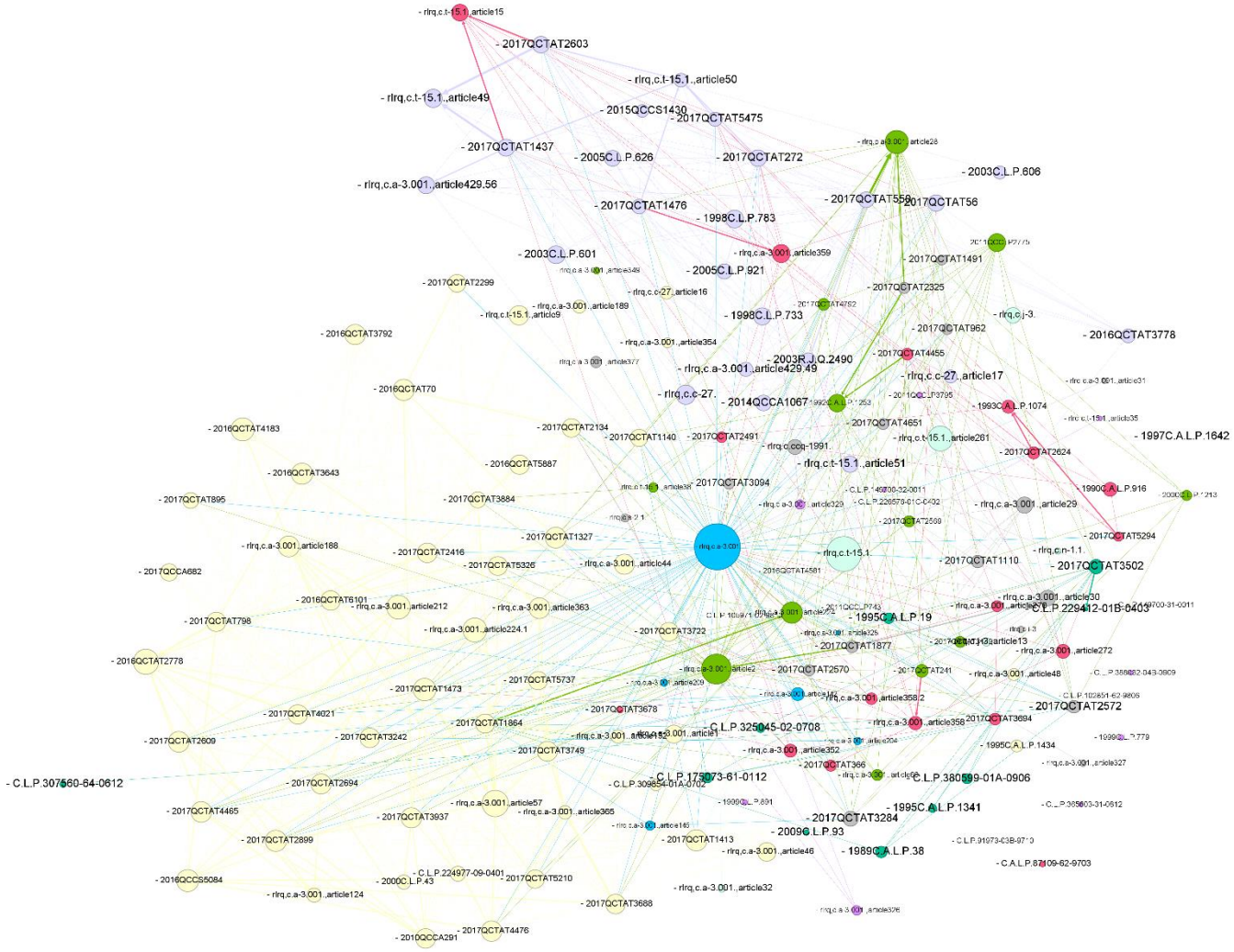
Wang Qiqi, Zhao Kaiqi, Amor Robert, Liu Benjamin, Wang Ruofan, Carpuat, M., de Marneffe, M.-C., & Meza Ruiz, I. V. (2022). « *D2GCLF: Document-to-Graph Classifier for Legal Document Classification*». In Findings of the Association for Computational Linguistics: NAACL 1 2022 (pp. 2208–2221). Association for Computational Linguistics. Récupéré de *https://doi.org/10.18653/v1/2022.findings-naacl.170*

Wehnert, Sabine, Viju Sudhi, Shipra Dureja, Libin Kutty, Saijal Shahania,t Ernesto W. De Luca. (2021). «*Legal Norm Retrieval with Variations of the BERT Model Combined with TF-IDF Vectorization*», Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 285–294. São Paulo, Brazil: Association for Computing Machinery. Récupéré de https://doi.org/10.1145/3462757.3466104.

Yang Zichao, Yang Diyi, Dyer Chris, He Xiadong, Smola Alex, Hovy Eduard (2016). «*Hierarchical Attention Networks for Document Classification*». In Proceedings of the 2016 Conference of the North American 1 Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480–1489). Association for Computational Linguistics. Récupéré de https://doi.org/10.18653/v1/N16-1174

Zeng, Daojian, Kang Liu, Yubo Chen, Jun Zhao (2015). « *Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks* », dans Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, p. 1753–1762, Lisbon, Portugal. Récupéré de https://doi.org/10.18653/v1/D15-1203.

Zhong, Haoxi, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, Maosong Sun (2020). « *How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence* », dans Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, p. 5218–5230. Récupéré dans https://aclanthology.org/2020.acl-main.466/


- rlrq,c.t-15.1.,article15
- 2017QCTAT2603
- rlrq,c.t-15.1.,article50
- rlrq,c.t-15.1.,article49
- 2015QCCS1430
- 2017QCTAT5475
- 2017QCTAT1437
- 2005C.L.P.626
- 2017QCTAT272
- rlrq,c.a-3.001.,article28
- 2003C.L.P.606
- rlrq,c.a-3.001.,article429.56
- 2017QCTAT1476
- 2017QCTAT558
- 2017QCTAT56
- 1998C.L.P.783
- 2003C.L.P.601
- rlrq,c.a-3.001.,article359
- 2017QCTAT1491
- 2005C.L.P.921
- 2017QCTAT2299
- rlrq,c.c-27.,article16
- 2017QCTAT2325
- rlrq,c.a-3.001.,article189
- rlrq,c.t-15.1.,article9
- 1998C.L.P.733
- rlrq,c.j-3.
- 2016QCTAT3792
- 2017QCTAT962
- 2016QCTAT3778
- rlrq,c.a-3.001.,article354
- 2003R.J.Q.2490
- 2017QCTAT4455
- rlrq,c.a-3.001.,article429.49
- rlrq,c.c-27.,article17
- 2016QCTAT70
- rlrq,c.c-27.
- 2014QCCA1067
- 1993C.A.L.P.1074
- 2016QCTAT4183
- 2017QCTAT2134
- 2017QCTAT4651
- 1997C.A.L.P.1642
- 2017QCTAT1140
- 2017QCTAT2491
- rlrq,c.ccq-1991.
- rlrq,c.t-15.1.,article261
- 2017QCTAT2624
- 2016QCTAT5887
- rlrq,c.t-15.1.,article51
- 2016QCTAT3643
- 2017QCTAT3094
- 1990C.A.L.P.916
- 2017QCTAT895
- 2017QCTAT3884
- rlrq,c.a-3.001.,article29
- 2017QCTAT1327
- 2017QCTAT5294
- rlrq,c.a-3.001.,article188
- 2017QCTAT2416
- rlrq,c.t-15.1.
- rlrq,c.n-1.1.
- 2017QCTAT5326
- rlrq,c.a-3.001.,article44
- 2017QCTAT1110
- 2017QCTAT3502
- 2017QCCA682
- rlrq,c.a-3.001.,article30
- 2016QCTAT6101
- rlrq,c.a-3.001.,article212
- rlrq,c.a-3.001.,article363
- 1995C.A.L.P.19
- C.L.P.229412-01B-0403
- 2017QCTAT798
- rlrq,c.a-3.001.,article224.1
- 2017QCTAT3722
- 2017QCTAT1877
- rlrq,c.a-3.001.,article272
- 2016QCTAT2778
- rlrq,c.a-3.001.,article2
- 2017QCTAT2570
- 2017QCTAT5737
- rlrq,c.a-3.001.,article48
- 2017QCTAT1473
- rlrq,c.a-3.001.,article358.2
- 2017QCTAT2572
- 2017QCTAT4021
- 2017QCTAT3678
- 2017QCTAT1864
- C.L.P.325045-02-0708
- rlrq,c.a-3.001.,article358
- 2017QCTAT3694
- 2017QCTAT2609
- 2017QCTAT3242
- 1995C.A.L.P.1434
- 2017QCTAT3749
- rlrq,c.a-3.001.,article352
- 2017QCTAT366
- C.L.P.307560-64-0612
- 2017QCTAT2694
- C.L.P.175073-61-0112
- C.L.P.380599-01A-0906
- 2017QCTAT4465
- rlrq,c.a-3.001.,article57
- 1995C.A.L.P.1341
- 2017QCTAT3937
- rlrq,c.a-3.001.,article365
- 2017QCTAT3284
- 2017QCTAT2899
- 2009C.L.P.93
- 2017QCTAT1413
- 1989C.A.L.P.38
- rlrq,c.a-3.001.,article46
- 2016QCCS5084
- C.L.P.224977-09-0401
- 2000C.L.P.43
- 2017QCTAT5210
- rlrq,c.a-3.001.,article124
- rlrq,c.a-3.001.,article32
- 2017QCTAT3688
- 2017QCTAT4476
- 2016QCCA291