HEC MONTRÉAL

Analyse de données textuelles et extraction de thèmes dans les retours d'employés d'une entreprise

par
Patrick Mesana

Analytique d'affaires

Mémoire présenté en vue de l’obtention du grade de maîtrise ès sciences (M.Sc.)

Résumé

La gestion des ressources humaines (GRH) a beaucoup évolué ces dernières années. Le gestionnaire a aujourd'hui à sa disposition, une grande quantité d'information et des outils pour tenter de comprendre ses employés afin de leurs offrir le meilleur environnement de travail possible. Cependant, lorsqu'il s'agit de données textuelles, les méthodes quantitatives sont encore à la frontière de la recherche et de l'industrie. Extraire les thèmes d'importance d'un grand nombre de retours d'employés est une tâche de traitement et de compréhension du langage.

Dans notre étude, nous avons utilisé les méthodes classiques de vectorisation de documents connues de la littérature, notamment TF-IDF. Nous avons aussi développé un nouveau modèle de poids se basant sur l'information mutuelle et utilisant les étiquettes fournis dans notre jeu de donnée. Ces étiquettes représentent les grands axes de la GRH. Nous avons ensuite développé un outil de visualisation, permettant d'explorer ces représentations vectorielles réduites à 2D par l'algorithme t-SNE. L'interactivité de notre outil nous a permis d'avoir une idée subjective mais utile, d'un thème dans nos données. Finalement, nous avons comparé une méthode de regroupement de points, DBSCAN, avec la méthode LDA qui est l'état de l'art dans la modélisation de thèmes.

Mots Clés : Traitement du langage, Extraction de thèmes, Retours d'employés, Analyse statistique, DBSCAN, LDA, Information mutuelle

Liste des tableaux

Liste des figures

Remerciements

Je tiens à remercier ma femme, Marie, de sa patience et de ses encouragements quotidiens. J'ai aussi une pensée pour ma petite fille, Anna, qui a commencé sa vie en même temps que son papa retournait aux études. Ce fut une aventure que nous avons partagée tous les trois et sans elles, elle n'aurait pas beaucoup de sens.

Je remercie aussi ma famille, en particulier mes parents, qui sont toujours et encore des repères dans ma vie.

Finalement, je remercie mon directeur, Gilles Caporossi, de sa bienveillance, des bonnes idées et conseils qu'il m'a apportés. J'espère sincèrement qu'on continuera à travailler ensemble.

Table des matières

1 Introduction

Les entreprises de nos jours, quelle que soit leur taille, peuvent accumuler un grand nombre de données. Nous nous sommes intéressés à un problème bien spécifique de l'une d'entre elles. La société en question, est spécialisée dans la collecte de retours d'employés sur leur environnement de travail, sur leur bien-être en entreprise et en dehors. Elle collecte ces retours dans différentes entreprises avec différents domaines d'expertise. Beaucoup de ces retours sont textuels et c'est ce sur quoi nous allons nous concentrer dans ce mémoire. L'objectif de ce projet de recherche consiste à trouver les thèmes ou tendances dans les retours textuels d'employés d'une entreprise.

Nous avons utilisé la science des données afin d'extraire de l'information nouvelle, en tentant d'apporter un éclaircissement ("insights" en anglais) aux entreprises concernées. Nous avons employé des méthodes classiques d'exploration, ainsi que des méthodes d'apprentissage machine. Nous regarderons aussi comment le contexte de la gestion des ressources humaines peut aider notre analyse. Ceci nous amènera à répondre à la question plus générale que pose ce mémoire :

Comment analyser quantitativement les milliers de retours d'employés d'une entreprise et en extraire une information utile pour ses gestionnaires?

2 Revue de littérature

Dans cette revue, nous voulons élaborer l'intérêt des entreprises à l'analyse de données et en particulier les données textuelles. Nous parlerons des méthodes couramment implémentées en entreprises et celles qui sont toujours à la frontière de la recherche scientifique et de l'industrie. Ceci devrait nous permettre d'avoir des attentes raisonnables, tout en laissant la porte ouverte à des approches innovantes.

2.1 Utilité des données textuelles pour une entreprise

Pendant longtemps, la science des données s'est concentrée sur les données dites structurées, c'est a dire des données organisées de façon à ce que l'on puisse facilement les retrouver. Par exemple, quelqu'un est facilement identifiable par son nom, son prénom et sa date de naissance. Récemment, un intérêt grandit pour les données non structurées. Par exemple, des images, des documents, etc. Nous commencerons par situer cet intérêt dans le cadre de la gestion des ressources humaines.

2.1.1 L'informatisation de la gestion des ressources humaines

La gestion des ressources humaines (GRH) a beaucoup évolué ces dernières années, les pratiques ont grandement changé, notamment grâce à l'utilisation des technologies du numérique. Rapidement, les entreprises se sont rendu compte de l'avantage compétitif de moderniser les pratiques de GRH, même si la valeur ajoutée peut varier d'une entreprise à une autre (Powell and Dent-Micallef 1997). La communication entre les employées s'est également modernisée, l'utilisation de courriel dans un premier temps, puis plus récemment la messagerie instantanée.

Une entreprise continuellement connectée permet à ses gestionnaires de sonder les employés et d'améliorer la gestion de l'entreprise. En effet, le gestionnaire peut évaluer la satisfaction de son équipe beaucoup plus fréquemment qu'auparavant, par exemple à travers l'élaboration de métriques de satisfaction ou d'engagement. Un des enjeux majeurs d'une entreprise est de retenir ses employés (Ramlall 2004). Des méthodes d'analyse similaires à celles employées pour mesurer la satisfaction des consommateurs (Rust, Stewart et al. 1996) sont

couramment utilisées dans la stratégie interne de nombreuses entreprises. Elles deviennent un atout compétitif.

La collecte de données doit être faite méthodiquement, il faut construire des questionnaires pertinents, poser les bonnes questions. L'industrie a investi massivement dans ce domaine et une expertise s'est formée (Sanchez 2007). Elle vise à aider les entreprises à rendre ces questionnaires efficaces. Parfois, les questions à choix multiples ne suffisent pas, d'où l'intérêt d'avoir des questions ouvertes où la personne peut répondre sans contrainte.
Néanmoins, le plus gros défi des entreprises n'est pas d'accumuler les données, mais bien de les exploiter. Les données non structurées telles les données textuelles présentent des défis supplémentaires.

2.1.2 Comment exploiter des données textuelles?

La démocratisation de l'internet a fait considérablement augmenter la quantité de documents textuels stockés numériquement. Il est très laborieux de lire avec attention tous les textes que possède une entreprise afin d'en extraire les sujets ou tendances d'importance (Grimmer and Stewart 2013). Pourtant, comme discuté, être capable de bien comprendre ses employés ou ses clients, est un enjeu majeur dans la société actuelle.

Avec les moyens technologiques d'aujourd'hui, la recherche informatique par mots clés est la méthode la plus simple à implémenter et elle est très efficace pour de nombreux problèmes. Par exemple, il existe un moteur de recherche ouvert et gratuit appelé Apache Lucene que les entreprises peuvent intégrer à leurs systèmes informatiques. La fonctionnalité de recherche de Wikipedia en est une application (McCandless, Hatcher et al. 2010). Cependant, la condition obligatoire de la recherche par mot clé est que l'utilisateur doit avoir une idée, plus ou moins précise, des sujets d'importances, c'est-à-dire qu'il doit être capable d'associer des mots à des thèmes. Si certains mots ne font pas partie de son vocabulaire, il passera vraisemblablement à côté de documents qui pourraient l'intéresser. Par conséquent, les analyses et interprétations qui découlent de ces recherches contiendront un biais.

Pour cette raison, des outils de visualisation ont été inventés afin d'apporter plus d'objectivité. Par exemple, le logiciel Wordstat permet de voir un corpus de documents sous

différents angles. Ces outils permettent notamment de lier des mots, des documents, et des informations structurées que l'entreprise possède. L'utilisation de ces outils implique tout de même un degré élevé d'expertise du domaine. Des formations techniques de haut niveau sont également nécessaires.

2.2 La science des données textuelles

L'apparition d'outils de "text mining" n'est en fait que la partie émergente de l'iceberg. Cela fait bien longtemps que des chercheurs se demandent comment exploiter une grande quantité de données textuelles. Durant la Deuxième Guerre mondiale, l'armée américaine a capturé un bon nombre de papiers scientifiques allemands et ils se sont demandé quel était le meilleur moyen de les indexer. De là est né le domaine de la récupération d'information (IR en anglais), mais c'est l'arrivée du "World Wide Web" et les premiers moteurs de recherche qui ont donné une toute autre dimension à ce domaine (Baeza-Yates and Ribeiro-Neto 1999).

2.2.1 Traitement du langage

La première question que l'on se pose lorsqu'on tente d'analyser des documents est; comment quantifier le texte?

On définit d'abord l'entité textuelle, qu'on nomme plus communément un document, et l'ensemble complet des documents forme un corpus. Ensuite, la technique la plus utilisée est de considérer un document comme un sac de mots (BOW en anglais). On perd instantanément la structure des phrases, mais on peut de cette façon, compter les mots des documents et former un vocabulaire.

Un vocabulaire plus riche sémantiquement ne signifie pas forcément un vocabulaire avec plus de mots. On peut être amené à vouloir comptabiliser seulement certaines catégories de mots. Par exemple, nous pourrions vouloir conserver uniquement les noms et les adjectifs. Pour y parvenir, nous avons besoin de l'information grammaticale de chaque mot d'une phrase, qu'a priori nous n'avons pas. Heureusement, il existe des algorithmes d'étiquetage grammatical ("Part of speech tagging" ou POS en anglais) non supervisé. C'est un problème complexe, car certains mots pris séparément peuvent avoir plusieurs sens, l'étiquette dépend alors du contexte de la

phrase, voire du corpus. Toutefois, le problème peut être isolé et traité séparément avec un certain succès, il existe aujourd'hui des solutions de plus en plus performantes par apprentissage machine (Andor, Alberti et al. 2016).

L'étiquetage grammatical peut aussi servir à une autre tâche très populaire dans le traitement du langage, la lemmatisation. Son objectif est de réduire les mots dérivés d'une même racine en leur forme commune. Par exemple, les mots au pluriel seront ramenés au singulier. C'est une tâche cousine de la désuffixation (stemming en anglais), qui est un processus transformant un mot en sa racine. La lemmatisation est considérée comme moins brutale, car la transformation se fait par une analyse morphologique des mots et un vocabulaire extérieur au corpus. Si on couple un algorithme de POS avec un processus de lemmatisation se basant sur une base de données lexicale telle Wordnet (Miller 1995), on peut réduire drastiquement la complexité du langage du corpus. En conséquence, des mots qui apportent une information sémantique similaire auront un compte commun, et on peut espérer que toutes tâches subséquentes y trouveront un gain.

Il existe bien d'autres techniques pour nettoyer ou simplifier les documents: nous ne les listerons pas. Nous porterons notre attention sur une, particulièrement intéressante pour l'analyse de documents de petite taille, la construction de phrases de noms (noun phrase en anglais). Une phrase de noms est une expression constituée d'au moins un nom et dont les mots ont une "collocation" (cooccurrence successive) significative (Manning and Schütze 1999). Pour cette raison, ils sont souvent une combinaison de deux ou trois mots. Déterminer quels sont les mots qui constituent de bonnes phrases de noms dans un contexte précis est une question ouverte.

2.2.2 Vectorisation des documents

Il est populaire de modéliser les sacs de mots comme des vecteurs de nombres réels, car ceci nous permet d'utiliser l'outillage mathématique. La taille du vecteur correspond à la taille du vocabulaire du corpus. Chaque dimension du vecteur représente le poids d'un mot du vocabulaire dans le document. L'occurrence d'un mot dans un document est très souvent utilisée comme la valeur de ce poids. Il est aussi possible de mesurer une distance entre deux vecteurs. Plusieurs mesures de similarité existent; minimiser la distance euclidienne est un choix possible,

l'angle entre les vecteurs en est un autre. Ce sont le type de données et l'objectif de la tâche qui déterminent quelle est la meilleure option. La distance cosinus est très utilisée, car elle est moins affectée par l'absence de mots dans un document.

Il est intéressant de noter que l'occurrence des mots dans un corpus suit remarquablement la loi Zipf (Zipf 1935). Ainsi, les mots de la catégorie des déterminants sont très fréquents alors qu'ils apportent peu à la sémantique d'une phrase. D'où la nécessité de calculer autrement le poids d'un mot, la plus connue est TF-IDF (Salton and Buckley 1988). Le calcul consiste grossièrement à multiplier la fréquence d'un mot dans un document par un ratio prenant en compte l'occurrence du mot dans l'ensemble du corpus (Figure 2-1). Ce nombre peut être également perçu comme la quantité d'information que le mot apporte au document, cette quantité est dépendante de la rareté du mot dans le corpus.

$$W_{i,j} = TF_{i,j} \times log\left(\frac{N}{df_i}\right)$$

Figure 2-1 - Formule de TF-IDF
$TF_{i,j}$ est l'occurrence du mot i dans le document j
df_i est le nombre de documents contenant le mot i
N le nombre total de document

Une vectorisation à l'aide de TF-IDF devrait en théorie permettre des regroupements de documents sémantiquement plus justes. Toutefois, les vecteurs de mots n'offrent pas une description suffisamment réduite d'un document. Ils ne nous permettent pas de déduire une structure statistique inter-documents ou intra-documents. De plus, la dépendance avec un vocabulaire non exhaustif limite la capacité expressive d'une idée et son importance dans un ensemble de documents.

2.3 Extraction de thèmes dans un corpus

La science des données textuelles ne sert pas uniquement à la récupération de documents. Tout ce que nous avons présenté a de multiples applications. Bill MacCartney (2014) propose un diagramme situant les tâches connues du traitement du langage. Il identifie un sous-ensemble de tâches plus complexes de compréhension du langage naturel (NLU en

anglais). L'extraction de thèmes fait partie de cette catégorie, c'est aussi un domaine de recherche très actif.

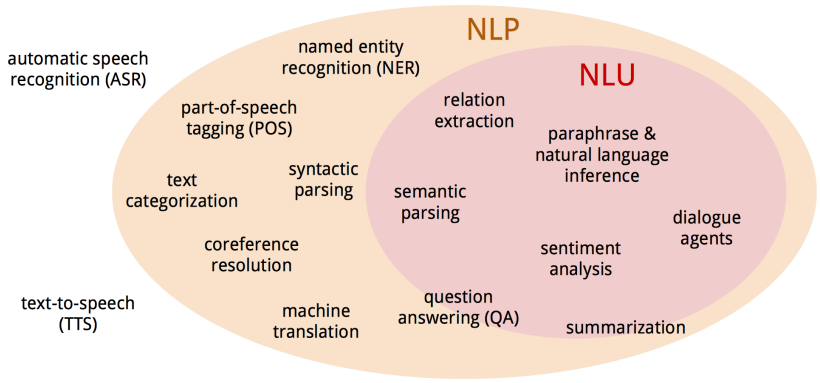


Figure 2-2 - Les tâches de traitement du language, bien connues de la communauté scientifique

2.3.1 Qu'est-ce qu'un thème ou un sujet?

C'est une question qui, de prime abord, peut paraitre triviale, mais dont la réponse peut varier suivant le contexte. Si on s'en tient à la définition formelle du dictionnaire Larousse, un thème ou sujet, est "ce qui fournit matière à quelque chose" ou "ce qui fait la matière d'une discussion, d'un écrit, d'une œuvre ; ce dont il s'agit". Donc, chercher un sujet dans un document revient à poser la question suivante : de quoi ce document s'agit-il?

Les mots qui constituent la réponse à cette question sont les représentants d'un ou plusieurs thèmes de ce document. C'est donc une idée abstraite, un concept, qui se matérialise par des mots. Ces mots sont remplis de sémantique et peuvent parfois suffire à identifier un thème. C'est le cas de la catégorie des noms, qui ont souvent un poids important dans la sémantique d'une phrase. Un document peut contenir un nombre arbitraire de sujets, ils seront assurément moins nombreux que le nombre de mots qu'il contient.

2.3.2 La modélisation mathématique de thème

Latent Semantic Indexing ou LSI (Deerwester, Dumais et al. 1990), fut une des premières méthodes notables dont l'objectif est de faire de la réduction de dimension sémantique. Elle se base sur la représentation vectorielle des documents, souvent produite par TF-IDF, et utilise l'algèbre linéaire pour décomposer ("Singular Value Decomposition" ou SVD) la matrice

correspondante au corpus. C'est-à-dire une matrice où une ligne est un document et où une colonne est un mot du vocabulaire.

$$\begin{bmatrix} T_{11} & \cdots & T_{1k} \\ \vdots & \ddots & \vdots \\ T_{m1} & \cdots & T_{mk} \end{bmatrix} \begin{bmatrix} W_{11} & \cdots & W_{1v} \\ \vdots & \ddots & \vdots \\ W_{k1} & \cdots & W_{kv} \end{bmatrix} \approx \begin{bmatrix} W_{11} & \cdots & W_{1v} \\ \vdots & \ddots & \vdots \\ W_{m1} & \cdots & W_{mv} \end{bmatrix}$$

Figure 2-3- Décomposition matricielle de la matrice correspondante au corpus
k est le nombre de sujets
m est le nombre de documents
v est la taille du vocabulaire

Cette décomposition sert à identifier un sous-espace vectoriel conservant un maximum de variance. LSI est très similaire à la méthode "Principal Components Analysis" (PCA) et peut être vue comme une méthode de réduction de dimension ou de compression de données textuelles. De plus, Deerwester et al. ont montré que les dimensions ou "features" de ce nouvel espace, pouvaient capturer des notions basiques linguistiques telles la synonymie ou la polysémie. Si on cherche une interprétation plus intuitive de LSI, on peut dire que des documents qui ont une forte cooccurrence de mots, en particulier des mots forts sémantiquement, vont globalement être similaires et pointeront dans la même direction.

2.3.3 La cooccurrence des mots

La cooccurrence de mots est une caractéristique importante de la modélisation mathématique de thèmes. De nombreux chercheurs ont tenté d'apporter une perspective statistique et de concevoir des modèles probabilistes pouvant expliquer la relation entre cooccurrence de mots et des variables explicatives. Latent Dirichlet Allocation ou LDA (Blei, Ng et al. 2003) est un modèle de référence. Contrairement à ses prédécesseurs (e.g : LSI) LDA permet de modéliser à la fois la relation entre les mots et les sujets, mais aussi la relation entre les documents et les sujets.

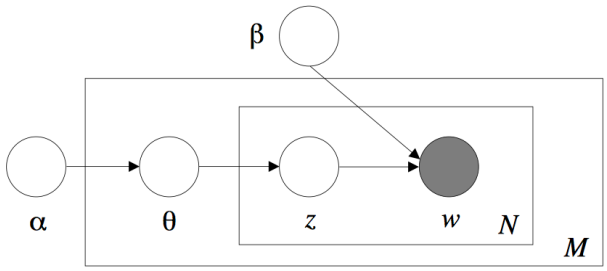


Figure 2-4- Modèle graphique de LDA
N est le nombre de mots dans un document
M est le nombre de documents
w correspond à un mot
z correspond à un jeton de thème
θ correspond à la proportion d'un sujet dans un document
α et β sont les paramètres fixes du modèle

LDA est un modèle d'adhésion mixte (Mixed Membership Model en anglais), c'est-à-dire qu'un document peut être membre de plusieurs sujets à la fois. L'équation bayésienne qui caractérise ce modèle permet d'inférer les paramètres d'une distribution de sujets pour chaque document.

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)}$$

Figure 2-5- Équation bayésienne de LDA
À gauche, la distribution a posteriori que l'on veut maximiser
α et β sont les paramètres fixes de la méthode
θ est le vecteur paramètre de la distribution des thèmes qui suit la loi de Dirichlet
w est un vecteur de mots (document)
z est un vecteur de jetons de thèmes

Seulement, l'inférence de ce type de modèle est un problème calculatoire d'une grande complexité. Il faudra faire une simulation Monte Carlo, ou bien faire appel à des méthodes approximatives avancées (Inference variationnelle) qui ajoutent leur lot d'incertitude.

En définitive, la direction prise par les modèles émergeants suit la définition d'un sujet donné précédemment. Un thème est défini par une distribution de mots qu'on retrouve dans les documents. Par contre, si on tente d'inférer les thèmes du corpus à partir du contenu des documents, et qu'on considère la cooccurrence comme la métrique à suivre, alors une autre question se pose : quel est l'impact de la taille des documents ?

2.3.4 Le cas particulier des textes courts

Depuis l'invention de LDA, de nombreuses variantes ont vu le jour, ajustant le modèle à des problèmes de plus en plus spécifiques, et un problème récurrent est celui des textes courts. Nous entendons par "texte court", un document ne contenant que quelques mots. Lorsqu'un document est suffisamment long, la fréquence d'observer deux mots respectivement corrélés à un même sujet peut être suffisamment élevée pour justifier l'existence de ce sujet dans un corpus. Dans un texte court, il est bien plus rare d'avoir une représentation distribuée d'un sujet complexe. Effectivement, la matrice représentant l'occurrence des mots dans les documents sera particulièrement creuse. Pour remédier à ce problème, les auteurs de Biterm (Yan, Guo et al. 2013) modélisent explicitement la cooccurrence des mots et sortent du paradigme par document et le remplacent par un découpage du corpus en contextes de courte taille. Dans le cas où les documents (ou pseudo-documents) sont tous échangeables suivant le théorème de Finetti (1990), ce modèle offre de bonnes performances sur des jeux de données de grande taille (e.g : twitter).

Dans la situation où l'ordre des documents importe, il est difficile de justifier ces modèles probabilistes. De plus, inférer les paramètres de ces distributions, suppose que les données n'ont pas d'autres variables explicatives qui biaiseraient nos résultats. Malgré la sophistication de ces modèles, leur applicabilité n'est donc pas si évidente.

De nos jours, il existe d'autres approches pour traiter les documents de petite taille. Les plus convaincantes tentent de profiter de grande quantité de données (e.g: Wikipedia) et utilisent des algorithmes d'apprentissage profond pour construire des espaces sémantiques de mots ou de documents (Le and Mikolov 2014).

Plutôt que de comparer des vecteurs d'occurrence de mots provenant du corpus, il est possible d’utiliser ces espaces construits préalablement (embeddings en anglais) pour calculer des distances entre les mots des documents, et induire la distance entre deux documents en utilisant des méthodes plus classiques (Kusner, Sun et al. 2015). Cette voie est prometteuse, mais il est encore difficile d'évaluer ces méthodes sur de nombreux problèmes.

2.4 Exploration et automatisation

À moins d'avoir des données étiquetées, il est difficile de valider les résultats d'une tâche d'extraction de sujets. Dans la plupart des situations, cette tâche est non supervisée. On peut être très ambitieux, c'est-à-dire réaliser une extraction sans avoir recours à une phase de sélection manuelle, donc totalement automatisée. On peut aussi utiliser la science des données pour simplement rendre la tâche manuelle beaucoup moins laborieuse. Dans tous les cas, le choix varie en fonction de l'application finale.

2.4.1 Visualisation du corpus

Une visualisation très populaire lorsqu'on veut voir rapidement quels mots apparaissent le plus dans un corpus est le nuage de mot (wordcloud en anglais). Il s'agit de placer des mots dans un espace fermé et la taille d'un mot dépend de l'importance qu'on veut bien allouer au mot. La figure 2-6 est une llustration d'un nuage de mots.

Figure 2-6- Exemple de nuage de mots

Les poids peuvent suivre une des métriques présentées, par exemple les poids TF-IDF. C'est une représentation artistique avec une utilité limitée.

Lorsque nous possédons des documents qui ont une évolution dans le temps, il est possible de regarder l'évolution des mots les plus significatifs. Si les documents sont annotés ou que nous procédons à des regroupements de documents par sujets, il est possible de détecter certaines tendances dans le corpus. La figure 2-7 est un diagramme tiré l'article "TextFlow" (Cui, Liu et al.

2011). Il montre lévolution de thèmes dans le temps, la taille du flux correspond à l'importance du thème dans le corpus.

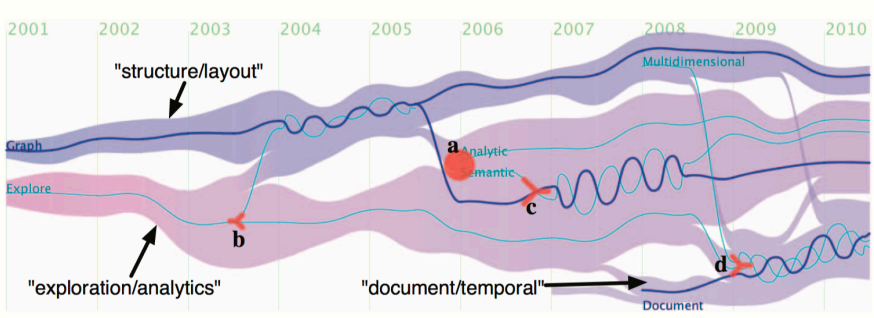


Figure 2-7- "TextFlow" et l'évolution de thèmes dans le temps

Finalement, on peut s'intéresser à voir les documents dans un espace vectoriel qu'on serait capable d'explorer et de comprendre visuellement, soit en 2D ou en 3D.

L'objectif est d'obtenir une représentation dans un espace réduit qui respecte le plus l'espace d'origine. Pour y parvenir, on peut tout simplement prendre nos documents en sac de mots sous forme vectorielle et appliquer les algorithmes de réduction de dimensions classiques. PCA (Principal Component Analysis) est une option possible, MDS (Multidimentional Scaling) peut être plus approprié dans notre cas. PCA utilise la distance euclidienne tandis que MDS peut se baser sur d'autres mesures de distance. Ces deux méthodes ont une caractéristique commune, elle tente de conserver la variance globale ou la dissimilarité des documents. Deux documents qui n'ont aucun mot en commun par exemple, ne devraient pas être proche en 2D. L'inconvénient, c'est que PCA et MDS capturent les variations globales aux dépens de celles locales, donc il est possible que des documents qui ont une certaine proximité se trouvent plutôt éloignés dans un sous espace. Une alternative, qui a gagné en popularité ces dernières années pour visualiser des espaces à beaucoup de dimension, est t-SNE (Maaten and Hinton 2008).

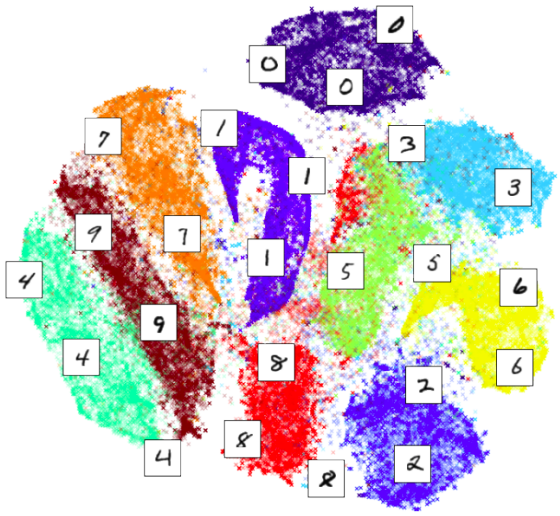


Figure 2-8- t-SNE du jeu de données MNIST
Il s'agit d'images en noir et blanc de nombres de 0 à 9.
On observe que les mêmes nombres se retrouvent dans l'espace

"t-Distributed Stochastic Neighbor Embedding" est une technique qui essaie de placer des points dans un sous-espace cible de façon aléatoire tout en conservant le voisinage des points dans l'espace de départ. L'astuce de SNE n'est pas de comparer directement les distances locales, mais de placer une gaussienne autour de chaque point et de transformer chaque distance (e.g : la distance cosinus) d'un point à un autre en une probabilité de voisinage. Si deux points sont proches, ils auront une forte probabilité d'être voisins. En appliquant le même principe dans l'espace cible, si les points sont parfaitement placés alors ils devraient avoir la même distribution de voisins que dans l'espace de départ. Pour se guider, l'algorithme utilise la descente du gradient sur la divergence Kullback-Leibler entre les deux distributions, celle dans l'espace de départ et celle de l'espace cible, car cette dernière change au fur et à mesure qu'on essaie d'améliorer la solution.

$$KL(P \parallel Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

Figure 2-9- Divergence KL qu'on tente de minimiser dans t-SNE
p_{ij} est la probabilité que le point i et j sont voisins dans l'espace de départ, de même pour q_{ij} dans l'espace cible

2.4.2 Les méthodes de regroupements

L'approche automatisée consiste à utiliser des méthodes de regroupement (clustering en anglais). Maintenant que nous savons comment transformer nos documents en points dans un espace vectoriel, nous pouvons potentiellement utiliser tous les algorithmes non supervisés connus de la littérature. Si on connait plus ou moins le nombre de groupes présents dans un corpus, alors l'algorithme k-means (MacQueen 1967) est un choix approprié. C'est un algorithme populaire dans la science des données qui permet d'obtenir un partitionnement de l'espace de document. Formellement, l'objectif est le suivant:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

Figure 2-10- L'objectif de K-means
$S = \{S_1, S_2 ..., S_k\}$ est une partition et S_i un groupe
μ_i est la moyenne des points dans S_i

Chaque portion de la partition trouvée peut être considérée comme le représentant d'un groupe sémantique. L'implémentation se résume en la répétition de deux étapes jusqu'à l'optimum. La première étape consiste à définir les centroïdes de chaque classe, la deuxième à affecter chaque document à la bonne classe. Cette mécanique est régulièrement appelée EM, pour Expectation Maximization en anglais. En pratique, k-means est fortement influencé par l'initialisation des centroïdes. Comme tout algorithme de partitionnement rigide, il est limité dans sa capacité d'exprimer des sujets hybrides (Steinbach, Karypis et al. 2000). Plutôt que de construire un partitionnement rigide, on peut aussi assigner un document à un sujet avec une probabilité.
C'est le cas des mixtures de gaussiennes (Bishop 2006) appliquées au regroupement de documents, dont l'implémentation est également basée sur EM.

Nous finirons par présenter une alternative aux méthodes de partitionnement (souple ou rigide). DBSCAN , "Density-based spatial clustering of applications with noise" (Birant and Kut 2007), se base sur la densité des points pour trouver des groupes de points intéressants. Il ne

prend pas d'hypothèse sur l'aspect global du corpus et se concentre sur la proximité des points, "nearby neighbours" en anglais. Son objectif est de trouver des points noyaux ("core points" en anglais) qui définissent un groupe. Pour se faire, l'algorithme a besoin de seulement deux paramètres, un nombre $minPts$ de points minimums à proximité d'un point noyau, et le rayon ε définissant le voisinage d'un point noyau.

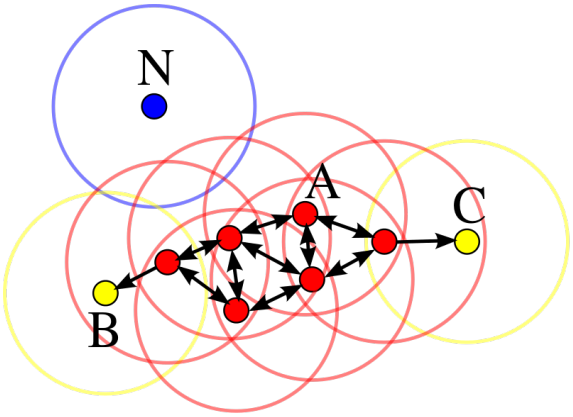


Figure 2-11- Ilustration de DBSCAN (Photo prise de Wikipédia)
minPts = 4

L'illustration ci-dessus montre les points noyaux en rouge, chaque point a au moins 4 points dans un rayon ε. Les points B et C en jaune font partie du même groupe sans être des points noyaux, et le point bleu est un point aberrant vis-à-vis de ce groupe.

Un avantage de cet algorithme est qu'il est déterministe. En effet, l'ordre de traitement des points n'importe pas, on arrivera toujours au même résultat. Contrairement à K-means, il se peut que plusieurs points soient considérés comme valeurs aberrantes à la fin. Ces points ne seront pas assignés à un groupe.

3 Méthodologie

Nous commencerons par décrire les données de l'étude, quels techniques nous avons utilisés pour les préparer. Ensuite, nous spécifierons la définition d'un thème dans notre contexte afin de trouver un terrain d'entente pour notre analyse. Nous avons utilisé des modèles de la littérature, mais nous proposerons aussi un modèle et une méthodologie adaptée à notre problème. L'extraction de thèmes se fera d'abord par exploration, à l'aide d'un outil d'analyse que nous avons développé, puis par une méthode de regroupement, plus objective et automatisé.

3.1 Description des données

Les données ont été fournies par une entreprise qui souhaite garder l'anonymat. Elles contiennent les enregistrements d'un outil de collecte de retours d'employés sur leur environnement de travail. Nous appellerons ces enregistrements, des "feedbacks" (de l'anglais), terme couramment utilisé dans une entreprise nord-américaine. Les données que nous possédons sont aussi dans la langue anglaise. Un feedback, est une réponse textuelle à une question, elle-même est le suivi d'une réponse à une première question à choix multiples. Ci-dessous, un schéma représentant la séquence de l'expérience utilisateur.

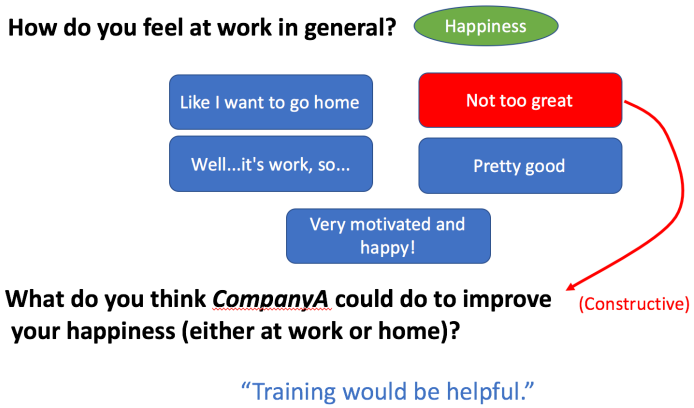


Figure 3-1- Scénario d'utilisation de l'outil de collecte.

La vignette verte correspond à une métrique associée à la question. La réponse est constructive (plutôt négative), alors une deuxième question est posée et la réponse de l'employé doit être textuelle. C'est de là que proviennent les données textuelles que nous allons analyser.
Suivant l'outil de collecte, les métriques (e.g: “Happiness”) correspondent aux grands axes de gestion des ressources humaines (RH). Chaque métrique a une paire de questions associées, une question constructive et une question positive. En effet, lorsqu'un employé répond à une question à choix multiplse, si sa réponse est très positive, l'outil le questionnera à nouveau pour qu'il donne une réponse textuelle dans le même esprit, de même si elle est négative. Sa réponse sera donc automatiquement associée à une métrique RH et un sentiment positif ou négatif.
Voici la description de quelques-unes de ces métriques et le type de questions de suivi :

- **Ambassadorship** : mesure comment l'employé fait rayonner l'entreprise à l'extérieur
    - Positif : Que recommandez-vous de notre entreprise?
    - Constructif : Qu'est-ce qui fait que vous ne recommanderez pas notre entreprise?
- **Happiness** : capture l'état de bien-être de l'employé
    - Positif : Qu'est-ce qui vous rend heureux ?
    - Constructif : Qu'est-ce qui nuis à votre bonheur que nous pourrions améliorer?
- **Relationship with managers** : mesure le ressenti de l'employé sur sa relation avec son gestionnaire.
    - Positif : Pourquoi vous pensez que vous avez une bonne relation avec votre gestionnaire ?
    - Constructif : Si vous pouviez changer une chose de l'interaction entre vous et votre gestionnaire, que'est-ce que ça serait ?

Nous insistons sur le fait que ces métriques RH sont définies par les concepteurs de l'outil de collecte. Le nombre de feedbacks textuels de l'entreprise sélectionnée après nettoyage est 13112, chaque feedback est associé à une métrique. Les feedbacks sont répartis sur une période de 16 mois.

3.2 Préparation du Corpus

Avant de commencer notre analyse, il faut préparer le corpus de documents afin qu'il soit exploitable. Nous avons tout d'abord procédé à un nettoyage des caractères spéciaux (e.g : emoticons) par un algorithme effectuant un simple filtrage de ces caractères.. Nous avons ensuite utilisé un algorithme d'étiquetage grammatical pour ne garder que les mots qui nous intéressaient. Plus précisément, nous avons utilisé la librairie Python Spacy qui implémente une version propriétaire, mais ouverte d'un algorithme supervisé, un perceptron, pour prédire correctement les entités grammaticales de chaque mot grâce à un corpus (en anglais) externe déjà annoté. Nous avons sélectionné deux stratégies de filtrage. Une première qui conserve uniquement les noms et les adjectifs, car ces catégories sont souvent celles qui ont le plus de sens dans une phrase. Une deuxième stratégie consiste à conserver toutes les catégories de mots excepté la ponctuation, les nombres ou autres symboles ayant passé les mailles du premier nettoyage (e.g : $, % ...).

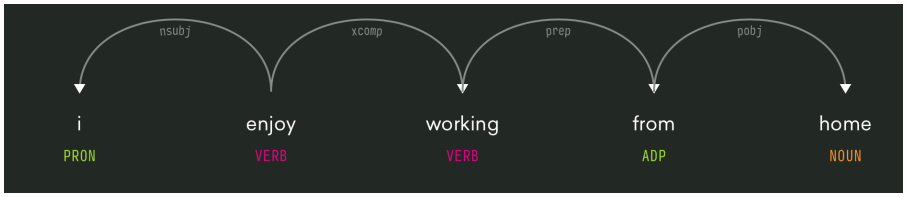


Figure 3-2- Exemple d'étiquetage grammatical réalisé par Spacy

Finalement, nous avons lemmatisé certains mots, cette fois-ci grâce à la libraire NLTK et à la base de données lexicale WordNet. En effet, cette librairie permet de transformer certains mots en leur lemme, par exemple la transformation du pluriel au singulier. Cette dernière étape réduit significativement la taille du vocabulaire sans grand impact sémantique, surtout si nous considérons un document comme un "sac de mots" comme nous l'avons fait.

working	work
isn't	be not
managers	manager
passed	pass

Figure 3-3- Exemple de mots à gauche et à droite leusr lemmes

3.3 Définition d'un thème dans notre contexte

Un thème est un concept exprimé par des mots qu'on retrouve dans des documents, dans notre cas un document est un retour d'employé. C'est aussi une réponse à une question de suivi. Les documents que nous avons à notre disposition sont souvent très courts. Ci-dessous, vous trouverez la distribution de la taille en nombre de mots des feedbacks du corpus avant prétraitement.

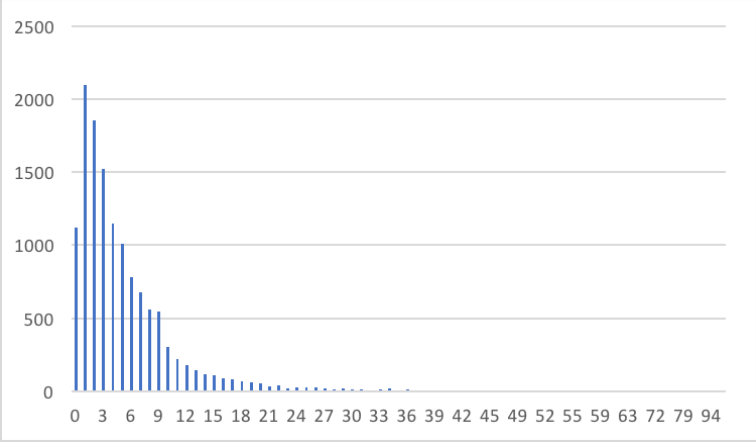


Figure 3-4- Distribution de la taille des feedbacks (en nombre de mots)

La petite taille des documents est une caractéristique spécifique à notre corpus, elle crée un lien très étroit entre les mots et les thèmes qu'on peut extraire. Les mots utilisés dans ces feedbacks sont l'évidence d'une intention d'un employé et cette intention est souvent unique par feedback. À cause du manque de cooccurrences de mots (les mêmes mots présents dans plusieurs documents), nous aurons peu d'évidence que certains mots sont liés entre eux et

forment un thème. Nous aurons également peu ou pas d'évidence que deux thèmes distincts par les mots, sont en fait qu'un seul thème.

Deuxièmement, nous sommes dépendants du vocabulaire employé dans l'entreprise. Nous voulons que notre analyse soit reproductible (sans beaucoup d'effort) à d'autres entreprises qui collecteraient des feedbacks, mais dont le domaine metier serait différent.

Finalement, nous pouvons espérer faire émerger uniquement les sujets qui sont exprimés dans le corpus par des réponses d'employés revenant assez fréquemment. Il sera difficile d'évaluer l'importance d'un thème dans l'entreprise, car nous nous concentrons sur une période dans le temps (16 mois). Une certaine volatilité des sujets principaux est à prévoir.

Ceci nous amène à poser trois hypothèses qui délimiteront notre recherche de thèmes :

- Nous cherchons en moyenne un thème par document.
- Nous cherchons uniquement des thèmes identifiables par le vocabulaire des employés.
- Nous cherchons des tendances.

Nous avons également besoin d'un formalisme précis pour comparer nos résultats. Afin de distinguer la tâche de regroupement de documents et la tâche d'extraction de thèmes, nous ne pouvons pas nous suffire d'un groupe de documents comme simple représentant d'un thème. Le formalisme que nous avons choisi, est de regarder à la fois les documents et les mots comme évidences de la présence d'un thème.

3.4 Lien entre métriques RH et thèmes du corpus

Jusqu'à maintenant, nous avons parlé de métriques RH indépendamment des thèmes. Une métrique RH n'est pas incompatible avec la définition d'un thème que nous venons de donner. Une métrique regroupe plusieurs documents, et on peut certainement regarder l'occurrence des mots de chaque métrique ainsi que leurs cooccurrences.

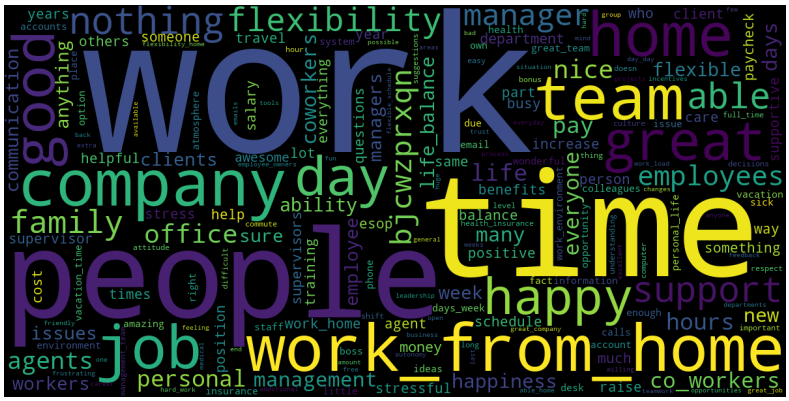


Figure 3-5- Wordcloud des feedbacks associés à la métrique RH "Happiness"

Bien évidemment, le but de cette recherche est de trouver des thèmes avec une granularité plus fine. On peut alors se poser la question; où se trouve un thème précis dans les métriques RH? Est-ce qu'un thème est exclusif à une métrique, ou bien est-il réparti sur l'ensemble des métriques RH?

Dans les limites de la définition donnée précédemment, un thème doit être représenté par une intersection de mots entre au moins deux documents. Supposons qu'on choisisse un mot comme le représentant d'un thème (e.g: "health"), on peut alors regarder la distribution des métriques pour ce mot. Soit,

$$P(M \mid T = "health")$$

où M est la variable aléatoire correspondant à une métrique et T la variable aléatoire correspondant à un terme (un mot). L'estimation de cette probabilité conditionnelle peut être affectée par la distribution des métriques. Nous avons donc vérifié que l'échantillon suit une loi uniforme.

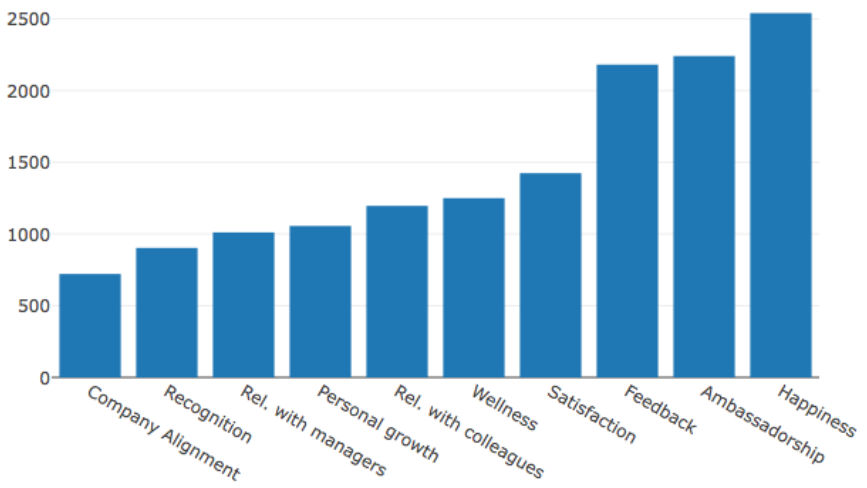


Figure 3-6- Nombre de documents pour chaque métrique

Nous pouvons observer une variance assez significative. Toutefois, les concepteurs de l'outil nous ont indiqué que l'algorithme générateur de questions a été conçu pour que la distribution des métriques tende vers une loi uniforme.

Nous avons ensuite observé que certains mots sont plutôt concentrés sur une ou deux métriques, tandis que d'autres sont présents dans toutes les métriques.

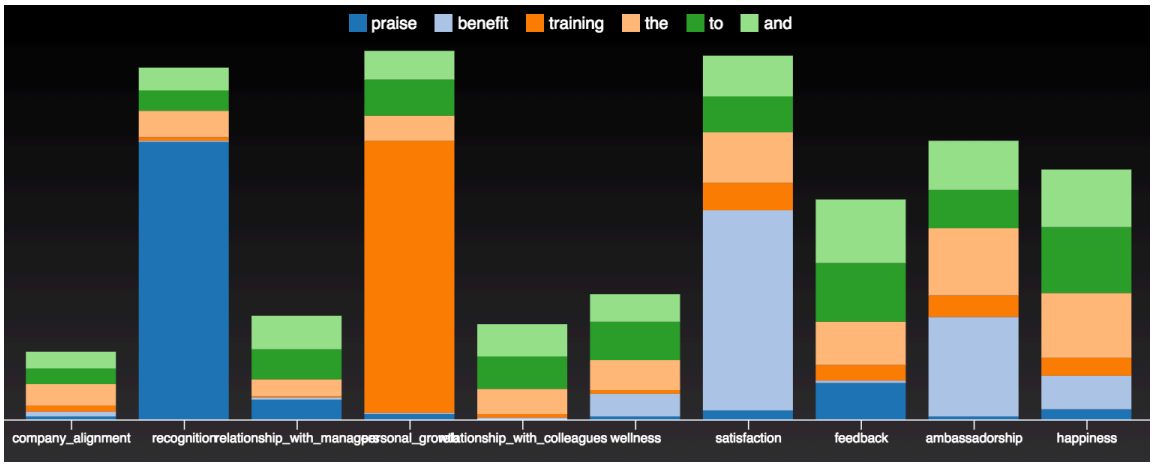


Figure 17 - Fréquence des métriques pour des mots ambigus et non ambigus

Nous faisons l'hypothèse, que certains mots utilisés par un employé sont plus ou moins ambigus vis-à-vis des questions qui leur sont posées. Les mots tels "the", "to", "and", couramment utilisé dans le langage, devraient se trouver dans toutes les métriques avec des proportions similaires. Il serait difficile de prédire la question en fonction de ces mots, c'est ce que nous voulons dire en les désignant comme ambigus. Tandis que des mots utilisés par les employés pour répondre précisément aux questions qui leur sont posées, tels "praise", "benefit", "training" , devraient se trouver majoritairement dans une ou deux métriques. Nous pensons que cette hypothèse est compatible avec le fait que les concepteurs de l'outil de collecte mettent au point des questions pour chercher de l'information spécifique. Une réponse valide à toutes les questions du système est peu probable et inconsistante avec la vision du produit.

Pour renforcer cette hypothèse, nous avons réalisé des tests d'indépendance sur 20 mots du corpus. Nous voulions vérifier que si nous prenons un mot qu'on considère ambigu et qu'on regarde les deux métriques où il apparait le plus, son niveau d'occurrence cumulé sera approximativement 20% (2 sur 10 métriques). Cela se traduit par la table d'incidence suivante.

MOT	STATISTIQUE KHI2	P-VALUE
THE	0.150829993	0.697743446
WORK	0.391498393	0.531512833
COMPANY	0.502871903	0.478240944
MANAGER	0.359048249	0.549035289
NOTHING	0.839311108	0.359593869
THING	0.074380165	0.785062869
PRAISE	2.461652794	0.116655974
HEALTH	1.949336577	0.162658359
CHAT	2.65569273	0.103179798
HOME	1.143785135	0.284853928
MISSION	3.096246302	0.078473041
CULTURE	2.144132653	0.143115899
THANK	2.94041935	0.086388302
FEEDBACK	1.771776844	0.183162333
BENEFIT	2.35254101	0.125078978
CHALLENGE	1.90217783	0.167834753
TRAINING	2.240374343	0.134448427
MENTORSHIP	4	**0.045500264**
SHE	3.658923358	0.05576966
HER	3.546178903	0.059682741

Table 3-1- Tests Khi Deux d'une sélection de mots

On peut observer que les mots tels "the" ou "work" sont utilisés dans toutes les métriques RH et ont des p-valeurs plutôt élevées tandis que des mots tels "praise" ou "mentorship" ont des p-valeurs très basses. Nous pensons donc que chercher des mots exclusifs aux métriques est une bonne méthode pour trouver des thèmes significatifs dans les feedbacks.

3.5 Modélisation

Mathématiquement, il est courant de voir une intersection de mots entre deux documents se traduire comme une distance entre les documents sous leur forme vectorielle. La première étape de la modélisation est donc de produire ces vecteurs inscrits dans l'espace des mots du vocabulaire du corpus, puis de définir une mesure de distance.

3.5.1 Modèles classiques de poids de mots

Le modèle de poids de mots le plus courant est l'occurrence des mots dans un document, en anglais il est nommé "Term Frequency" (TF). Ce modèle nécessite seulement de constituer le vocabulaire du corpus et de compter les mots qui apparaissent dans un document. Le modèle TF ne suppose pas que certains mots apportent plus ou moins de sémantique. Ceci ne reflète pas vraiment la réalité. Premièrement, certains mots du langage tels les articles sont utilisés dans tout contexte et apparaissent donc beaucoup plus souvent que le reste des mots du langage, sans apporter plus de sémantique. Deuxièmement, dans le contexte d'un texte court, il est rare de voir des mots apparaitrent plusieurs fois, excepté pour les mots fréquemment utilisés dans le langage. Donner plus de poids à un terme tel "the" dans un texte court ne nous permettra pas de comparer correctement deux documents.
Un autre modèle très populaire qui tente de remédier à ce problème est TF-IDF.

$$TFIDF_{i,j} = TF_{i,j} \times log\left(\frac{N}{df_i}\right)$$

Où $TFIDF_{i,j}$ est le poids TF-IDF pour le mot i du document j.

$$df_i = \left| \frac{N}{\{d \in D : t \in d\}} \right|$$

C'est-à-dire, le nombre N de documents d du corpus D sur le nombre de documents où le terme t apparait.

Dans notre contexte, $TF_{i,j}$ a une faible variance à cause de la petite taille des textes. Nous décidons alors de poser $TF_{i,j} = 1$ lorsque le mot est présent, même s'il est présent plusieurs fois dans le document. Nous avons alors,

$$TFIDF_t = log\left(\frac{N}{df_t}\right) = -log\left(\frac{df_t}{\mathrm{N}}\right)$$

On note que le poids du mot ne dépend plus du document et $\frac{df_t}{\mathrm{N}}$ est simplement la fréquence du mot dans le corpus. On peut dire que cette quantité est un estimateur de la probabilité $P(T)$ d'observer le terme t et $-log\left(\frac{df_t}{\mathrm{N}}\right)$ est un estimateur de la quantité d'information $I(T)$.

3.5.2 Modèle de poids adapté utilisant l'information mutuelle

Les modèles de poids que nous venons de présenter sont fréquemment utilisés dans la littérature, car ils sont agnostiques du type de corpus que l'on souhaite analyser. Cependant, certains poids TF-IDF peuvent paraitre arbitraires si on regarde l'ordre des mots qui en découle.

MOT	TF-IDF (BINAIRE)
BLUE	7.871845
EMERGENCY	7.689524
MOTIVATION	7.535373
EQUIPMENT	6.996376
GOAL	4.613749
OPPORTUNITY	3.790924
THE	1.097393

Table 3-2- Poids TF-IDF d'une sélection de mots

Nous avons conçu un modèle de poids qui prend en compte la dépendance d'un mot avec une métrique RH. Nous espérons que ce modèle sera plus adapté aux spécificités de notre corpus.
Plutôt que d'utiliser les scores khi deux, qui nécessitent une configuration particulière pour modéliser notre hypothèse, nous avons choisi une mesure dont la signification est plus intuitive, il s'agit de l'information mutuelle.
Avant de regarder la formulation exacte, nous voulons justifier ce choix. Plus tôt, nous avons utilisé $TFIDF(T)$ et choisis la formulation qui le rend équivalent à la quantité d'information $I(T) = -log\ P(T)$. Ceci nous permet de voir ce poids comme une statistique, nous avons voulu conserver cette perspective.
Tout d'abord, pour comprendre comment intégrer la notion de contexte, on a regardé $P(T)$ comme la probabilité marginale de $P(T, C)$ où C est un contexte d'utilisation d'un terme T.
Cela suppose qu'on a une liste exhaustive des contextes où se trouve un terme. On a alors,

$$P(T) = \sum_{i}^{n} P(T, C = c_i)$$

où c_i est le contexte i.
Nous avons supposé que les métriques RH sont de tels contextes, car on s'intéresse à l'utilisation d'un terme dans le corpus, uniquement sous le spectre des métriques RH définies. On a donc,

$$P(T = t) = P(T = t, M = m_1) + P(T = t, M = m_2) + \ldots + P(T = t, M = m_{10})$$
$$= (P(M = m_1 \mid T = t) + P(M = m_2 \mid T = t) + \ldots + P(M = m_{10} \mid T = t)) \times P(T = t)$$

Où m_i est la métrique RH i.

Cette décomposition en probabilités conditionnelles rejoint l'analyse que nous avons faite lorsque nous avons établi un lien entre un thème et une métrique RH. Si un mot est complètement exclusif à la première métrique, on aura:

$$P(T = t) = P(M = m_1, T = t) = P(M = m_1 \mid T = t) \times P(T = t)$$

Par cet calcul, nous voulons montrer que la forme de la distribution de $M|T$ et l'ambiguïté d'un terme T sont liés.

Si la distribution des métriques pour un terme est uniforme (on suppose aussi que M est uniforme), alors le terme est complètement ambigu. À l'opposé, un terme n'est pas ambigu s'il est exclusif à un contexte comme l'exemple ci-dessus. Puisque nous voulons un concept qui capture cette idée, l'entropie conditionnelle $H(M|T)$ est parfaitement adapté. Elle représente l'espérance de la quantité d'information conditionnelle de M sachant T.

$$H(M|T) = -\sum_{M} P(M \mid T) \log P(M \mid T)$$

$H(M|T)$ est élevé lorsque $P(M \mid T)$ est proche d'une loi uniforme (terme très ambigue). $H(M|T)$ est faible lorsque $P(M \mid T)$ sera concentré sur quelques valeurs de M (terme peu ambigu).

La corrélation entre variables aléatoires vue sous le prisme de la théorie de l'information est bien connue. Elle est plus généralement définie comme l'information mutuelle entre deux variables.

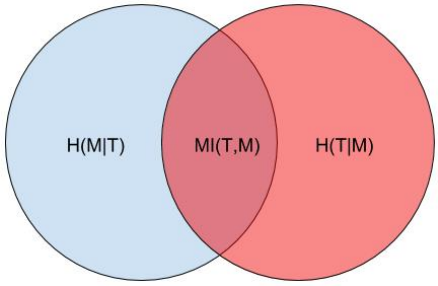


Figure 3-7- Diagramme de Venn de l'information mutuelle MI(T,M)

L'information mutuelle IM (ou MI pour "Mutual Information" en anglais) s'intéresse au ratio suivant:

$$r = \frac{P(T, M)}{P(T)\,P(M)}$$

Si les variables aléatoires T et M sont indépendantes, $r = 1$ et $-log\, r = 0$.
Nous allons maintenant développer la formule de MI pour montrer qu'elle est directement liée avec l'entropie conditionnelle $H(M|T)$.

$$
\begin{aligned}
IM(T,M) &= \sum_M \sum_T P(T,M)\ \log\left(\frac{P(T,M)}{P(T)\,P(M)}\right) \\
&= \sum_M \sum_T P(T,M)\ \log\left(\frac{P(T,M)}{P(T)}\right) - \sum_M \sum_T P(T,M)\ \log(P(M)) \\
&= \sum_T \sum_M P(T)\,P(M \mid T)\ \log(P(M \mid T))\ - \sum_M P(M)\ \log(P(M)) \\
&= \sum_T P(T) \sum_M P(M \mid T = t)\ \log(P(M \mid T = t))\ + H(M) \\
&= \sum_T P(T)\left(-H(M|T = t)\right) + H(M) \\
&= \sum_T P(T)\left(H(M) - \boldsymbol{H}(\boldsymbol{M}|\boldsymbol{T} = \boldsymbol{t})\right)
\end{aligned}
$$

Cette dernière forme introduit l'information mutuelle ponctuelle IMP (ou PMI pour "Ponctual Mutual Information" en anglais) qui nous permet d'estimer IM.

$$IMP(T = t, M) = h(M) - h(M|T = t)$$

où h est la quantité d'information.

$IMP(T = t, M)$ est précisément ce que nous allons utiliser pour le poids adapté d'un terme. Plutôt que d'utiliser la quantité d'information du mot (TF-IDF), nous allons utiliser la quantité d'information mutuelle entre un mot et les métriques. Lorsqu'un mot est ambigu, sa distribution tend à suivre la distribution des métriques (qui devrait être uniforme), sa quantité d'information tend à être similaire à celle des métriques. $IMP(T = t, M)$ tend alors vers 0. Dans le cas contraire, c'est $h(M|T = t)$ qui tend vers 0 et $IMP(T = t, M)$ tend vers $h(M)$.

MOT	INFORMATION MUTUELLE PONCTUELLE
THE	0.033713
WORK	0.2216
COMPANY	0.241618
MANAGER	0.219172
NOTHING	0.327286
THING	0.018078
PRAISE	1.161003
HEALTH	0.975814
CHAT	1.327468
HOME	0.561876
MISSION	1.571635
CULTURE	0.942081
THANK	1.258634
FEEDBACK	0.886052
BENEFIT	0.953637
CHALLENGE	0.888184
TRAINING	1.136337
MENTORSHIP	2.230832
SHE	1.420289
HER	1.4707

Table 3-3- Information Mutuelle Ponctuelle pour une sélection de mots (la même que pour les tests d'indépendance)

Il est intéressant d'observer que l'ordre des mots est presque le même que celui des valeures-p observées précédemment lors des tests d'indépendance khi-deux.

3.5.3 Mesure de distances

Dans le cas d'une vectorisation TF (binaire), la distance cosinus diminue plus l'intersection de mots entre deux documents est grande. Nous avons utilisé la distance cosinus pour les poids TF et TF-IDF. Ci-dessous, la formulation exacte si on s'intéresse uniquement à la présence d'un mot dans un document et non à son occurrence dans le document.

$$cosine(d_1, d_2) = 1 - \frac{d_1 \cdot d_2}{\sqrt{n_1}\sqrt{n_2}}$$

Où d_i est le document i et n_i est le nombre de mots uniques dans d_i.

La distance euclidienne est un autre choix possible. Elle met plus d'accent sur la dissimilarité des documents. Prenons l'exemple de 3 documents où la présence d'un mot est représentée par un 1 si présent et un 0 sinon.

DOCUMENTS	MOT1	MOT2	MOT3	MOT4	MOT5
$\boldsymbol{d1}$	1	0	0	0	1
$\boldsymbol{d2}$	1	1	0	0	1
$\boldsymbol{d3}$	0	0	1	1	1

Ci-dessous les distances calculées entre les documents.

PAIRS DE DOCUMENTS	**DISTANCE EUCLIDIENNE**	**DISTANCE COSINUS**
$(\boldsymbol{d1}, \boldsymbol{d2})$	1	0.184
$(\boldsymbol{d1}, \boldsymbol{d3})$	1.73	0.592
$(\boldsymbol{d2}, \boldsymbol{d3})$	2	0.667

Les écarts de distances entre $(d1, d2)$ et $(d2, d3)$ sont plus importants si on regarde la distance euclidienne. Nous avons préféré la distance euclidienne à la distance cosinus pour notre poids adapté utilisant l'information mutuelle, car les mots fortement ambigus ont un poids très faible, proche de 0, tandis que ceux plus exclusifs aux métriques RH, ont des poids plus fort. Nous supposons que deux documents avec, tous deux, des mots faiblement ambigus mais sans une grande intérsection entre ces mots, devraient correspondre à des thèmes distincts. La distance euclidienne devrait refléter cette divergence.

3.5.4 Réduction de dimensions

Une fois que nous avons des documents dans un espace vectoriel, nous pouvons tenter de réduire cet espace et d'utiliser les mesures de distance sur ces vecteurs réduits.

Latent Sementic Indexing

LSI utilise la décomposition en valeurs singulières de la matrice correspondante au corpus, c'est-à-dire l'ensemble des documents sous forme vectorielle. On peut alors réduire le nombre de dimensions de cette matrice en conservant les dimensions qui minimalisent l'erreur de reconstruction. Plus on conserve de dimensions, plus cette erreur tend vers 0, car on finit par retomber sur la décomposition initiale, qui est sans perte d'information.

Cette technique de compression sémantique prend donc essentiellement un paramètre, le nombre de dimensions à conserver. LSI peut fonctionner à la fois sur TF ou TF-IDF, nous l'avons aussi appliqué à notre modèle adapté. Pour les trois modèles de poids, nous avons choisi trois paramètres différents, 50, 100 et 200 dimensions. Si on compte les vecteurs non compressés, cela fait 12 représentations vectorielles du corpus.

Latent Dirichlet Allocation

LDA est un modèle de référence pour l'extraction de thèmes. Il est populaire, car son objectif est très clair. À partir de l'occurrence des mots des documents, il infère les thèmes du corpus sous une formulation probabiliste. Un sujet sera représenté par une distribution de mot. Cette définition est proche de celle que nous avons donnée dans notre contexte, donc ce modèle est tout à fait pertinent. Il ne nécessite pas de choix particulier sur le poids des mots, mais demande un choix judicieux de paramètres.

Il peut être considéré, tout comme LSI, comme un algorithme de réduction sémantique de dimensions. On dit que les vecteurs cibles sont inscrits dans l'espace latent des sujets. Chaque dimension correspond donc à un sujet.

Les paramètres de LDA sont:

- le nombre de sujets
- α, la concentration des sujets dans chaque document (e.g : un document a plus de chance d'avoir un mixe de 2 sujets plutôt que de 4 sujets)
- β, la concentration des mots de chaque sujet (e.g : un sujet est essentiellement un mixe de 3 mots plutôt que de 5 mots)

Puisque nous sommes en présence de documents courts, nous avons choisi des paramètres favorisant peu de sujets par documents et peu de mots par sujets. Voici nos paramètres:

Nombre de thèmes	50 et 100
α	0.0001
β	0.0001
Nombre max d'itérations	2000

LDA peut être utilisé comme un modèle capable de généralisation, c'est-à-dire pour donner des prévisions sur de futurs documents. Il est fréquent de séparer le jeu de données pour créer un jeu d'entrainement et un jeu de validation pour mesurer à la fois la précision et la généralisation de l'algorithme. Ici, nous l'utilisons uniquement comme un algorithme de réduction de dimensions, c'est-à-dire que nous sommes conscients, mais indifférents à l'effet de surentrainement, car nous voulons avant tout faire émerger des thèmes pertinents dans un corpus de grande taille.

3.6 Evaluation

Nous utiliserons dans un premier temps une approche par exploration qui nous permettra de mieux comprendre le corpus à notre disposition. Ensuite, nous utiliserons une méthode de regroupement, DBSCAN, pour trouver automatiquement des regroupements intéressants dans notre contexte.

3.6.1 Exploration

La méthode par exploration que nous avons utilisée est couplée à l'outil d'analyse que nous avons développé. L'objectif est de pouvoir naviguer dans le corpus par les mots des documents. Concrètement, nous avons créé plusieurs visualisations interactives que nous

présentons ci-dessous. Les captures d'écrans ne font pas justice à l'interactivité de l'outil, des explications se trouvent en Annexe.

Wordcloud par métrique

Le wordcloud est très classique dans le domaine du traitement du langage. Il consiste à créer une image constituée de mots avec des tailles proportionnelles à leur poids dans le corpus. Nous avons utilisé le vocabulaire constitué uniquement des noms et des adjectifs, et l'occurrence comme poids. Nous avons créé un wordcloud par métrique pour donner une idée des mots importants.

Distribution des mots filtrés

Nous avons trouvé intéressant de visualiser la distribution des mots, pouvoir la filtrer par métrique et par mots. Nous pouvons changer l'ordre des poids des mots en changeant le modèle de poids (TF, TF-IDF ou notre modèle adapté) et aussi changer de stratégie de vocabulaire, par exemple choisir seulement les noms et les sujets. Voir Annexe 1.

Distribution d'un sous-ensemble de mots dans les métriques

Il est aussi intéressant de voir la distribution conditionnelle des métriques pour un mot ou plusieurs mots. Ceci permet notamment de voir quels mots sont plus ou moins exclusifs à une métrique ou quels mots se trouvent dans toutes les métriques. Voir Annexe 2.

Évolution des mots dans le temps

Nous avons défini un thème dans notre contexte comme une tendance, il est donc normal de s'intéresser à l'évolution dans le temps d'un ou plusieurs mots. Voir Annexe 3.

Représentation des documents en 2D

Les visualisations précédentes perdent la notion de document, nous voulions aussi développer une visualisation qui permet de voir l'ensemble des documents.

Cette représentation dépend d'une tâche de réduction de dimension, car nous voulons projeter un espace multidimensionnel dans un espace 2D, compréhensible à l'oeil nu. Sous sa forme

vectorielle non compressée, un document peut avoir plus de 2000 dimensions dépendamment du vocabulaire sélectionné.

Nous avons choisi t-SNE, qui permet de réaliser une réduction de dimension en se concentrant sur les distances locales. Nous rappelons que la distance locale est la distance document à document. Par comparaison, PCA va favoriser une conservation de la variance globale (minimisation de l'erreur de reconstruction).

Il est important de noter que l'algorithme ne garantit pas le placement d'un point en matière de distance entre les points. Puisqu'on réalise une optimisation statistique, il se peut qu'on observe des erreurs grossières de placement. La méthode ne garantit pas non plus que deux nuages de points proches dans l'espace 2D soient proches sémantiquement.

Malgré ces derniers points, t-SNE est parfaitement adapté à la visualisation, car ce qui nous intéresse est l'aspect général du graphique. Deux points proches localement dans l'espace d'origine auront une forte probabilité d'être proches dans l'espace 2D, ce qui n'est pas garanti avec une méthode telle PCA. La méthode n'impose pas de mesure de distances, nous avons choisi celles que nous avons présentées dans la section concernée.

t-SNE a essentiellement un paramètre important, la perplexité. Ce paramètre est lié à la notion de voisinage d'un point, il correspond à la prédiction du nombre moyen de voisins d'un document dans le corpus. Plus la valeur est petite, plus l'attention de l'algorithme sera locale. Il est conseillé d'utiliser une valeur entre 5 et 50. Après avoir essayé plusieurs paramétrages, voici les paramètres qui semblent nous donner les meilleurs résultats visuels:

Perplexité	30
Nombre de dimensions de l'espace d'arrivée	2
Nombre d'itérations max	5000
Initialisation des points	Aléatoire

Nous pouvons ensuite colorer le graphique obtenu avec les métriques RH ou autre dimension intéressante. Nous avons aussi développé la possibilité de chercher des mots et de naviguer dans cette espace en déplaçant une lentille sur les points d'intérêt. Voir Annexe 4.

3.6.2 Regroupement non supervisé automatique

t-SNE nous permet de visualiser le corpus suivant une représentation vectorielle choisie (e.g : vecteurs LSI), mais nous devons tout de même naviguer dans un espace 2D et chercher à la main, les groupes de documents qui nous intéressent. Qui plus est, les erreurs de placements ne sont pas acceptables si on tente d'analyser plus finement la distribution des thèmes dans le corpus. C'est la raison pou laquelle l'utilité de t-SNE est essentiellement la visualisation. Nous proposons d'exécuter un algorithme de regroupement, DBSCAN, afin de trouver automatiquement des groupes dans l'espace multidimensionnel des documents et d'offrir une synthèse des résultats sur laquelle nous pourrons nous baser pour tirer certaines conclusions.

DBSCAN est déterministe, ce qui rend ses résultats compréhensibles et cohérents sur plusieurs exécutions. Les implémentations de certaines méthodes probabilistes, telles LDA ou t-SNE, ne garantissent pas le même résultat à chaque exécution. C'est aussi une méthode qui ne tente pas de placer chaque point dans un groupe, ce sont les paramètres qui vont déterminer quel type de groupe l'algorithme va détecter, et donc combien de groupes il est possible de détecter. La forme et la taille des groupes peuvent varier pour un même paramétrage. C'est une différence majeure avec k-means qui tente d'assigner chaque point à un groupe, et dont la forme des groupes est circulaire. Nous désirions une méthode de regroupement nécessitant peu d'hypothèses sur le corpus et ses sous-groupes de documents. C'est également un algorithme très performant qui peut être exécuté pour de nombreuses configurations de paramètres et sur une grande quantité de données.

En fait, nous avons sélectionné les paramètres de la méthode, le nombre de points $minPts$ minimum à proximité d'un point noyau, et le rayon ε définissant le voisinage d'un point noyau, en regardant les distances du K prochain voisin d'un point pour différentes valeurs de ε.

Nous notons que DBSCAN est très sensible à la mesure de distance choisie, surtout dans des espaces à de nombreuses dimensions. Par contre, tout comme t-SNE, aucune mesure n'est

imposée. Nous sommes restés cohérents avec les choix précédents, nous avons utilisé la distance cosinus pour les modèles TF, TF-IDF ainsi que leurs transformations réduites par LSI. Pour le modèle de poids adapté et ses réductions, nous avons utilisé la distance euclidienne.

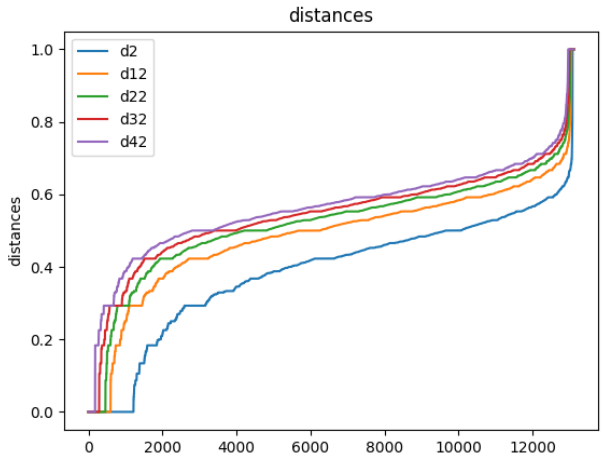


Figure 3-8- distances du K voisin pour chaque document vectorisé TF
d2 correspond à la distance du deuxième voisin
d12 correspond à la distance du douzième voisin

Sur la figure, on observe que la distance séparant un point et son voisin K, pour K compris entre 2 et 42, est au dessous de 0.7 pour la quasi-totalité des points. Ce graphique nous aide à ne pas considérer trop de paramètres. Pour TF, nous avons choisi les valeurs 0.5, 0.6 et 0.7. Nous avons effectué cette analyse pour les 14 types de transformations vectorielles que nous avons obtenues par les modèles présentés, voici l'ensemble des paramètres que nous avons essayé.

MODÈLE	**NBR DE DIMENSIONS**	**MESURE DE DISTANCE**	ε **(0.1 INC)**	$minPts$
TF	2001	Cosinus	[0.5, 0.7]	2, 12, 22, 32, 42
TF-LSI	200	Cosinus	[0.4, 0.6]	2, 12, 22, 32, 42
TF-LSI	100	Cosinus	[0.3, 0.6]	2, 12, 22, 32, 42
TF-LSI	50	Cosinus	[0.2, 0.6]	2, 12, 22, 32, 42
TF-IDF	2001	Cosinus	[0.6, 0.8]	2, 12, 22, 32, 42
TF-IDF LSI	200	Cosinus	[0.4, 0.6]	2, 12, 22, 32, 42

TF-IDF LSI	100	Cosinus	[0.4, 0.5]	2, 12, 22, 32, 42
TF-IDF LSI	50	Cosinus	[0.2, 0.4]	2, 12, 22, 32, 42
MI	2001	Euclidienne	[1.5, 2.9]	2, 12, 22, 32, 42
MI-LSI	200	Euclidienne	[0.4, 2.6]	2, 12, 22, 32, 42
MI-LSI	100	Euclidienne	[0.3, 1.9]	2, 12, 22, 32, 42
MI-LSI	50	Euclidienne	[0.1, 1.5]	2, 12, 22, 32, 42
LDA	100	Euclidienne	[0.3, 0.6]	3, 6, 9, 12
LDA	50	Euclidienne	[0.3, 0.5]	3, 6, 9, 12

Table 3-4 - Ensemble des modèles etparamètres DBSCAN

Ceci correspond à 1415 exécutions. Pour chacune, nous avons construit des statistiques nous permettant de qualifier les regroupements effectués:

- Le nombre de groupes
- La taille moyenne d'un groupe
- Le pourcentage de points appartenant à un groupe
- L'homogénéité des métriques RH sur les groupes

L'homogénéité est la seule qui demande plus d'explications. Elle a été définie par Andrew Rosenberg et Julia Hirschberg (Rosenberg and Hirschberg 2007) comme mesure d'évaluation de regroupement dans le cas où l'on a accès à des documents étiquetés. L'idée est la suivante; après l'étape de regroupement, on regarde si chaque groupe est constitué de points appartenant à une même classe. Si c'est le cas, alors on a une homogénéité parfaite sur l'ensemble des groupes. Bien sûr, ceci n'arrive presque jamais, mais on peut quantifier cette notion par l'entropie conditionnelle de la distribution d'une classe étant donné un regroupement. Dans notre cas, nous considérons qu'une classe est une métrique RH. Ceci rejoint l'hypothèse qu'un groupe sémantique a une plus forte probabilité d'appartenir à une même métrique.

$$h = \begin{cases} 1 & si\ H(M,K) = 0 \\ 1 - \dfrac{H(M \mid K)}{H(M)} & sinon \end{cases}$$

Où h est l'homogénéité d'un regroupement, M la variable aléatoire correspondant à une métrique et K la variable aléatoire correspondant à un groupe du regroupement.

3.7 Développement d'un outil d'analyse

Avant de passer aux résultats, nous voulons donner plus de détails sur l'implémentation de l'outil que nous avons développé pour réaliser notre analyse.

Aucun logiciel "boite noire", tels SAS ou SPSS, n'ont été utilisés. Ces logiciels permettent de rapidement exécuter des algorithmes déjà implémentés et préconfigurés, mais sont moins flexibles quand l'objectif est de développer ses propres méthodes et d'offrir une visualisation plus interactive. Nous ne disons pas qu'il est impossible de réaliser ce projet avec ces logiciels, mais priorisant la flexibilité, nous avons favorisé une approche davantage paramétrable.

Pour la partie algorithmique, nous avons utilisé le langage de programmation Python ainsi qu'un ensemble de librairies afin de ne pas réimplémenter des algorithmes très populaires dans les analyses de textes.

Voici les librairies principales utilisées pour cette partie :

- scikit-learn : une boite à outils d'algorithmes d'apprentissage machine, tels SVM, kNN, PCA...
- gensim : une boite à outils d'algorithmes spécialisés dans le NLP et en particulier la modélisation de sujet. e.g : TF-IDF, LDA
- NLTK : Une librairie facilitant le prétraitement d'un corpus de documents. Par exemple, un de ses modules permet de faire une lemmatisation des documents en utilisant la base de données WordNet.
- Spacy : Une libraire très similaire à NLTK, mais qui offre des algorithmes plus avancés pour certains de ces modules, tel son algorithme de POS (Part of speech).

Toutes ces librairies ont des fonctionnalités en commun. Nous ne justifierons pas toujours l'utilisation d'une librairie plutôt qu'une autre, car cela peut venir de la facilité d'utilisation de la librairie et non de la justesse ou de la performance de son implémentation.

Pour la partie visualisation graphique, nous avons choisi d'implémenter une interface Web afin d'apporter un maximum d'interactivité dans l'exploration de nos données et résultats. Voici les librairies principales utilisées pour la visualisation graphique :

- Javascript/HTML
- React : Un framework facilitant le développement d'applications Web
- D3 : Un framework offrant une grammaire de manipulation d'objets graphique, particulièrement utile pour la manipulation de SVG (Scalable Vector Graphics)

Annexe 5 montre un diagramme (en anglais) du processus de traitement des données jusqu'au moteur graphique de l'application Web.

4 Résultats

Nous avons déjà montré des résultats provenant de notre outil d'analyse en rapport avec la distribution des mots dans le corpus. Ils nous ont permis d'émettre certaines hypothèses et d'argumenter la méthodologie choisie. Pour l'étape d'exploration, nous allons nous concentrer sur les représentations 2D que nous avons obtenues par t-SNE. Ceci nous permettra de comparer les modèles sélectionnés qualitativement. Finalement, nous montrerons les résultats obtenus par DBSCAN, qui tente d'automatiser l'étape d'exploration.

4.1 Comparaison des graphiques t-SNE pour différent modèle de poids

Comme nous l'avons expliqué dans la méthodologie, nous avons utilisé la méthode LSI pour réduire les documents sous forme vectorielle à 200, 100 et 50 dimensions. Pour chaque modèle, nous avons construit une représentation 2D à l'aide de t-SNE . Nous avons regroupé les graphiques par famille de poids (TF, TF-IDF et poids basés sur IM) afin de voir l'impact général de cette technique de réduction sémantique. Nous avons exploré chaque famille, pour comprendre quel type d'agrégats nous pouvons extraire, et analyser si ces agrégats ont un sens sémantique. Nous rappelons qu'un point, dans un de ces espaces 2D, est un feedback. Deux points proches correspondent à une petite distance locale dans l'espace multidimensionnel d'origine, donc une forte probabilité d'avoir des mots en commun. Les métriques RH nous guideront dans notre analyse, chaque couleur correspond à une métrique RH tel qu'indiqué dans la légende.

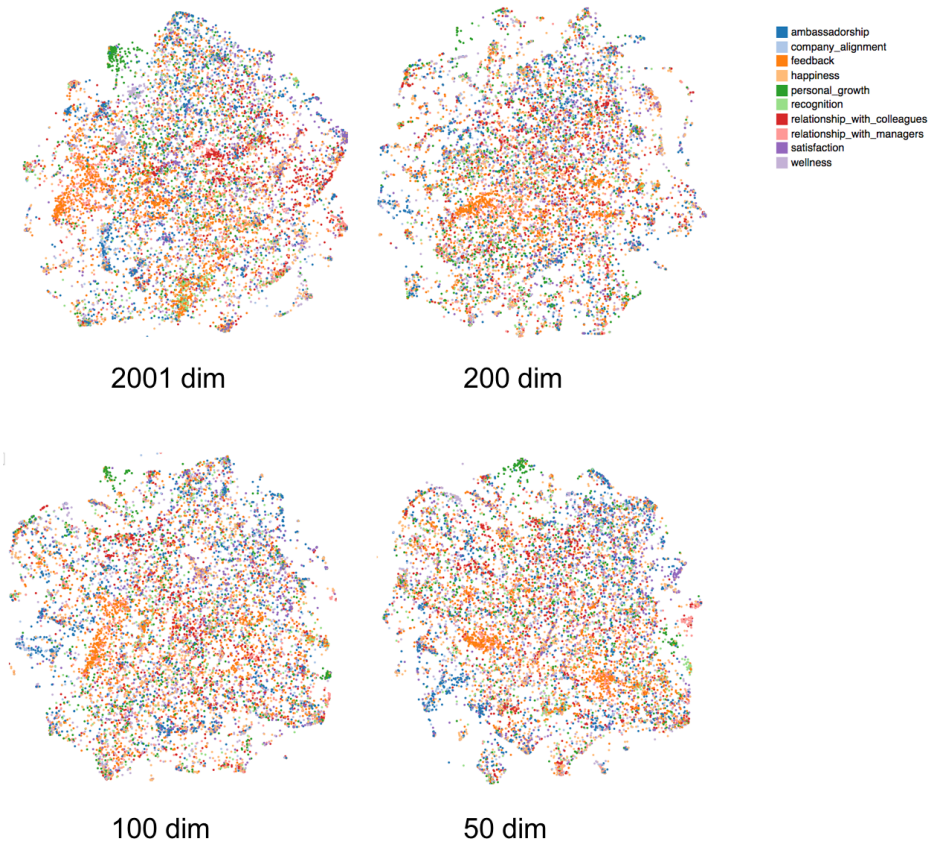


Figure 4-1 - t-SNE appliqué à une vectorisation à poids TF

La vectorisation TF sans réduction (TF 2001 dimensions), nous permet d'observer des agrégats avec une certaine concentration de métriques RH. Si nous zoomons sur un agrégat de ce type comme ci-dessous, nous trouvons des mots qui sont plutôt exclusifs à la métrique RH "Personal Growth".

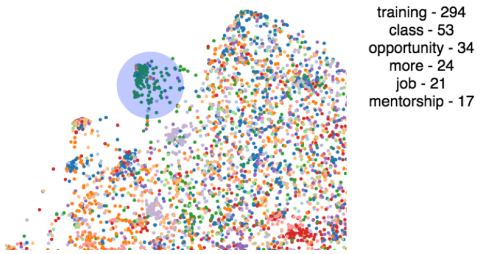


Figure 4-2 - Zoom sur un agrégat TF 2001 dimensions
La lentille bleue permet de calculer des statistiques uniquement pour les documents dans son rayon

D'un simple regard, il est difficile d'extraire ces agrégats du reste des points. Il n'est pas non plus évident de conclure si les vecteurs réduits apportent une meilleure visualisation pour la famille TF.

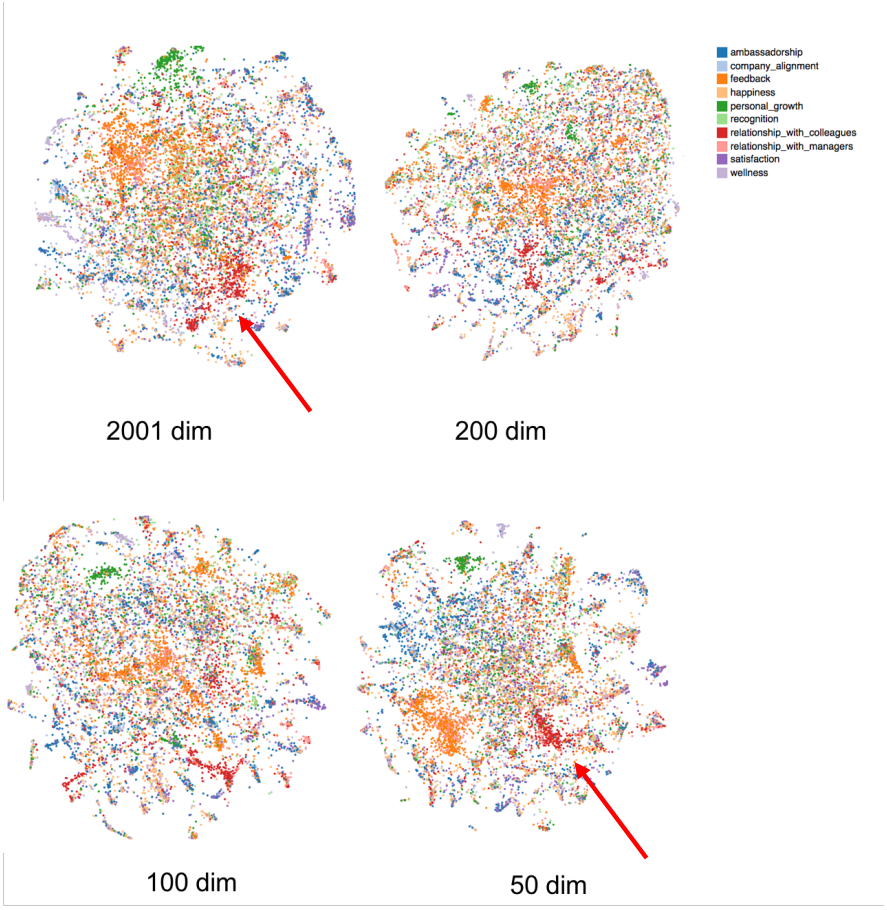


Figure 4-3 - t-SNE appliqué a une vectorisation à poids TF-IDF

Si on regarde maintenant les graphiques pour la famille TF-IDF, nous pouvons plus facilement identifier l'agrégat avec une concentration de la métrique "Relationship with colleagues".

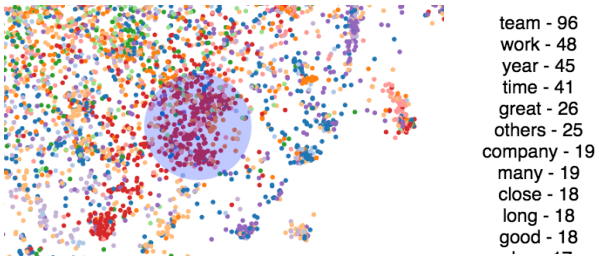


Figure 4-4 - Zoom sur un agrégat où la métrique "Relationshiop with colleagues" domine (TF-IDF LSI 2001)

L'outil d'analyse nous permet de lire le contenu des feedbacks sous la lentille et de réaliser que le thème dominant est; la sensation des employés d'appartenir à une famille et la joie qu'ils ont de travailler en équipe.

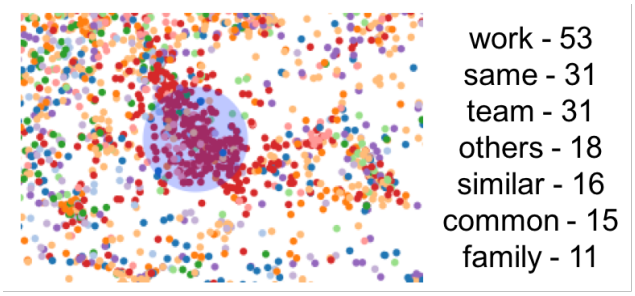


Figure 4-5 - Zoom sur un agrégat où la métrique "Relationship with colleagues" domine (TF-IDF LSI 50)

Une réduction à 50 dimensions accentue la séparation entre les agrégats. Si on recherche l'agrégat contenant le thème décrit plus haut, on observe également un agrégat dans cet espace. Cette fois-ci, les mots principaux sont plus proches du thème émergeant de ces documents. Nous n'avons pas besoin de lire le contenu pour nous en rendre compte, nous avons seulement besoin de regarder la liste des mots les plus fréquents. LSI fonctionne ici comme un filtre, d'ailleurs on ne voit pas d'autres agrégats significativement concentrés autour de la métrique "Relationship with colleagues".

Ci-dessous, un autre exemple, où nous avons sélectionné deux agrégats où la métrique "wellness" domine. Les deux agrégats ont des termes en commun, comme "program", mais contiennent également des mots respectivement dominants tels "health" et "wellness".

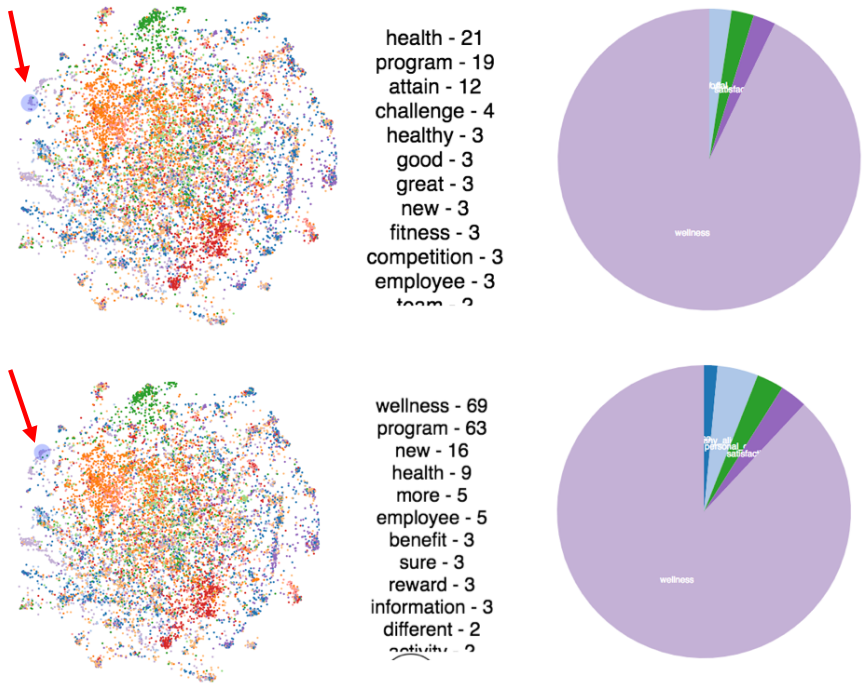


Figure 4-6 - Deux agrégats corrélés où la métrique "Wellness" domine (TF-IDF LSI 2001)

Ces deux agrégats sont proches dans l'espace 2D, car ils sont constitués de documents ayant une cooccurrence commune. Dans l'espace à 50 dimensions, en cherchant les mots qui reviennent le plus souvent dans les deux, on observe qu'ils sont concentrés dans un seul agrégat.

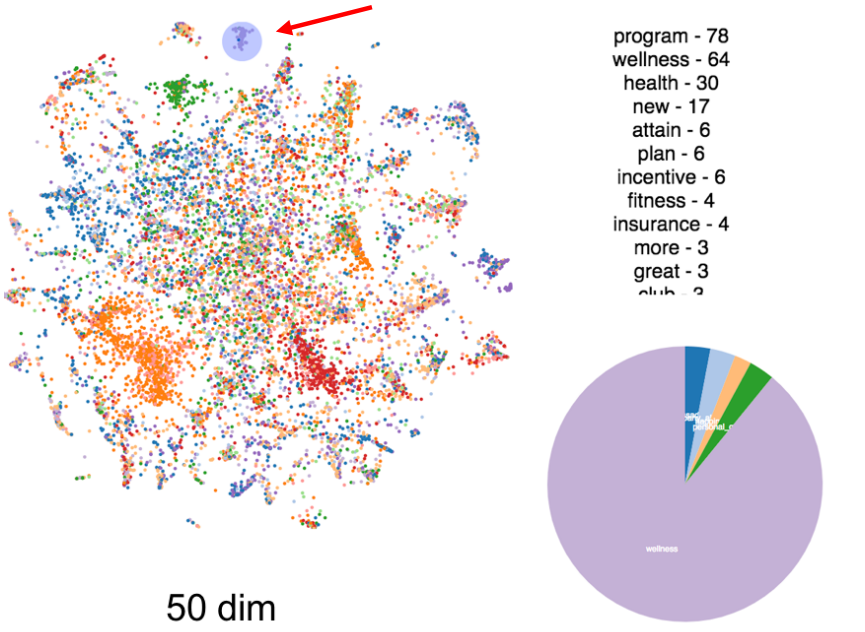


Figure 4-7 - Agrégat où la métrique "Wellness" domine (TF-IDF LSI 50)

Nous faisons une petite parenthèse pour dire qu'il peut être intéressant de définir un thème à partir de ces mots clés et de regarder leur évolution dans le temps. Un gestionnaire pourrait être intéressé par l'impact d'actions prises dans l'entreprise ou tout simplement pour vérifier qu'il s'agit d'un thème récurrent dans l'entreprise.

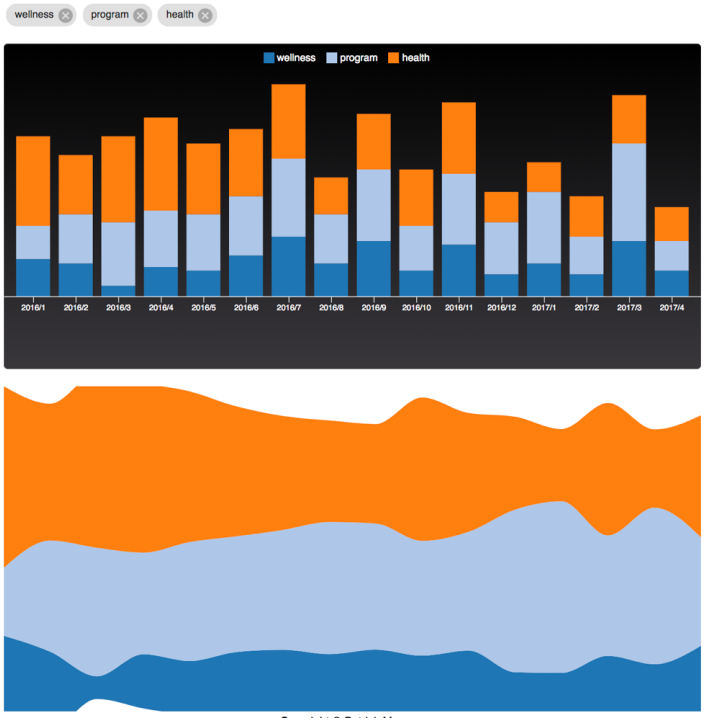


Figure 4-8 - Évolution des mots "wellness", "program" et "health" dans le temps

Nous avons aussi observé des agrégats où plus d'une métrique RH dominaient. Ces cas sont intéressants, car ils permettent de voir que certaines métriques RH ont une intersection sémantique.

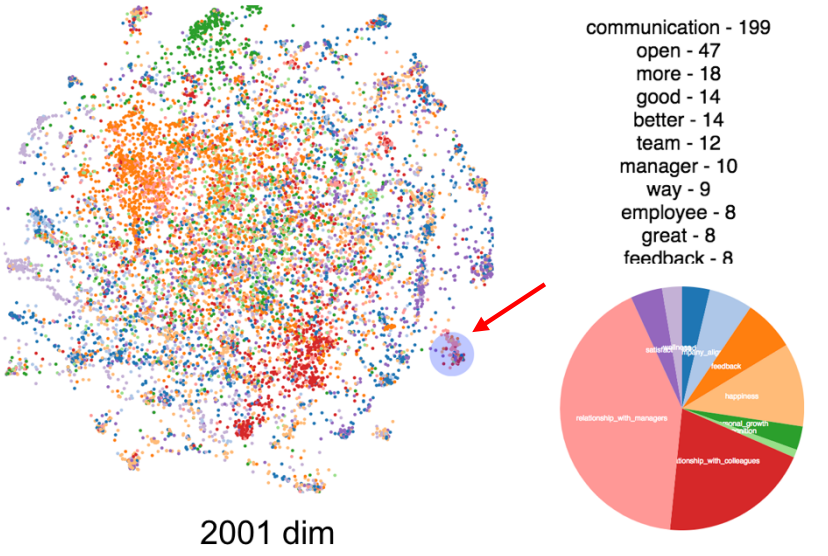


Figure 4-9 - Aggregat avec une concentration de deux métriques RH (TF-IDF LSI 2001)

Ci-dessus, on remarque un agrégat plutôt isolé du reste des points concernant principalement les métriques "Relationship with managers" et "Relationship with colleagues". Le mot qui revient le plus souvent est "communication".

TF-IDF nous offre globalement une visualisation plus intéressante et montre que l'allocation de poids plus élevés aux mots rares nous permet d'isoler plus d'agrégats. LSI va aussi dans ce sens, la réduction à 50 dimensions est celle qui nous offre les résultats les plus convaincants jusqu'à présent. Ces résultats sont bien sûr qualitatifs. On peut difficilement se satisfaire de ces graphiques, car la densité des points semble plutôt homogène. C'est-à-dire qu'on observe peu de concentration de points qu'on peut appeler un agrégat. C'est une des raisons pour lesquelles nous avons cherché d'autres modèles de poids de mots. Nous avons appliqué la même méthodologie à notre poids de mots adapté et basé sur l'information mutuelle entre un terme et une métrique RH.

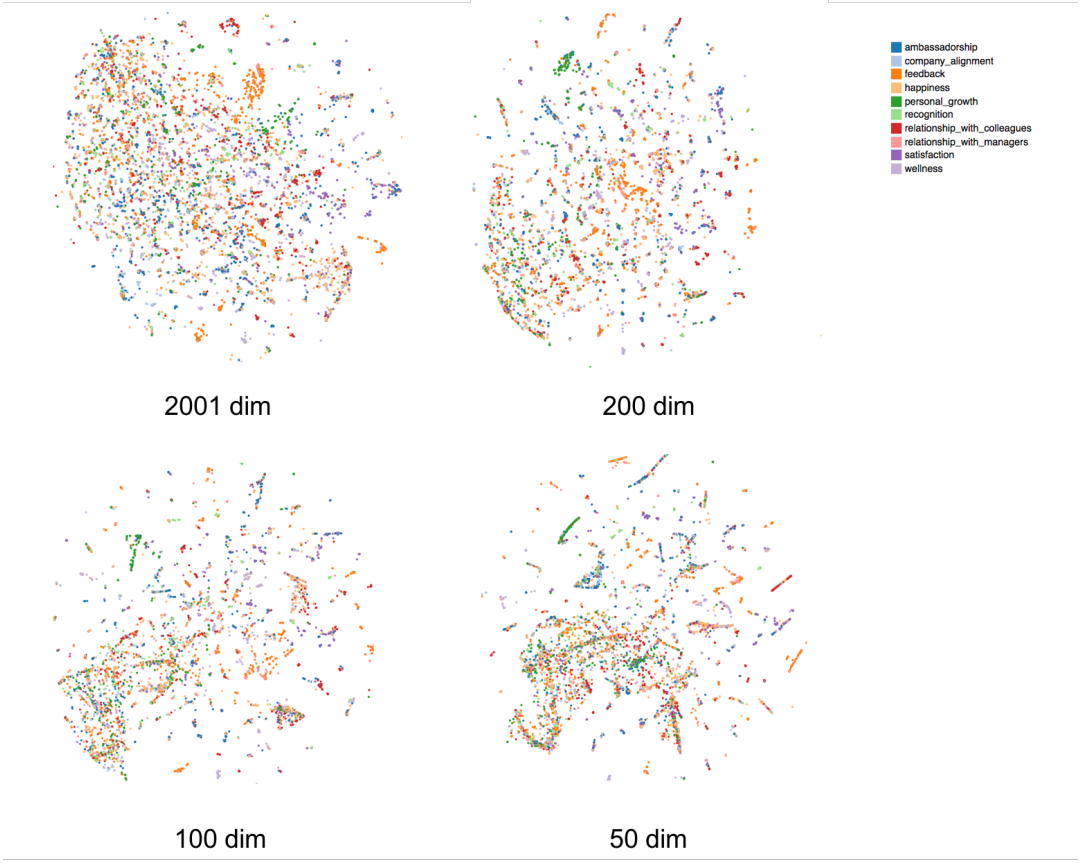


Figure 4-10 - t-SNE appliqué a une vectorisation à poids basé sur l'information mutuelle

La première impression est que le graphique parait beaucoup moins chargé, comparé à TF et TF-IDF. Cette impression est trompeuse, car le nombre de points est exactement le même, les points sont en fait beaucoup moins dispersés et parfois très concentrés jusqu'à avoir une centaine de points superposés dans l'espace 2D.

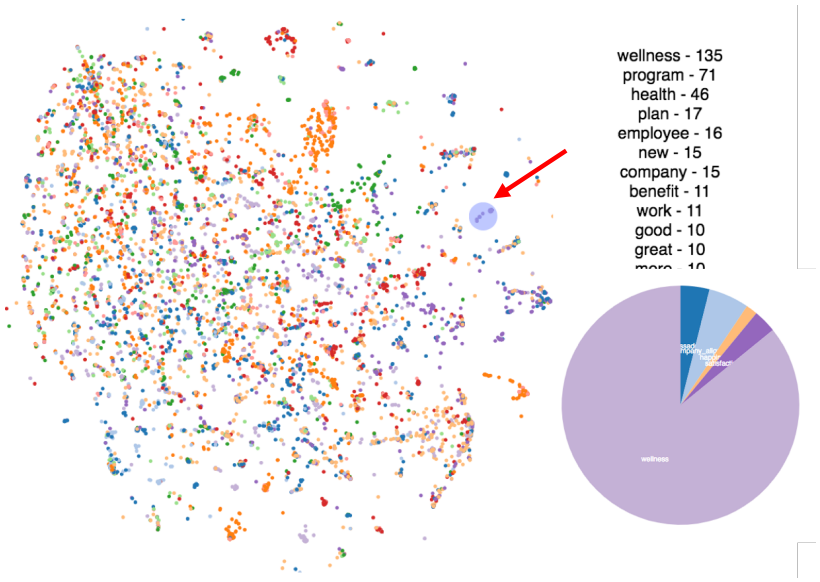


Figure 4-11- Agrégat où la métrique "Wellness" domine (MI)

Nous avons recherché le thème composé des mots "wellness", "program" et "health". Sans même avoir réalisé de réduction de dimension, il semble isolé et concentré dans l'espace 2D.

LSI permet là encore d'accentuer les distances entre certains agrégats. Il est intéressant de voir que le résultat obtenu par une réduction à 50 dimensions avant t-SNE, nous offre un graphique avec une densité très hétérogène.

Il reste difficile de compter ces agrégats, mais nous commençons à avoir une idée du type d'agrégats sémantique que nous cherchons. Les feedbacks liés par un même thème doivent avoir des mots clés définissant ce theme et avoir une forte concentration de métrique RH.

4.2 Évaluation de LDA

Nous prenons maintenant une orientation différente avec LDA, qui permet d'automatiser complètement l'extraction de thèmes, sans avoir à faire un choix de poids de mots. Nous avons sélectionné deux modèles, 50 et 100 thèmes, pour rester pertinents vis-à-vis de l'analyse comparative de poids de mots et LSI. Ci-dessous, les représentations t-SNE utilisant les vecteurs réduits où une dimension correspond à un thème.

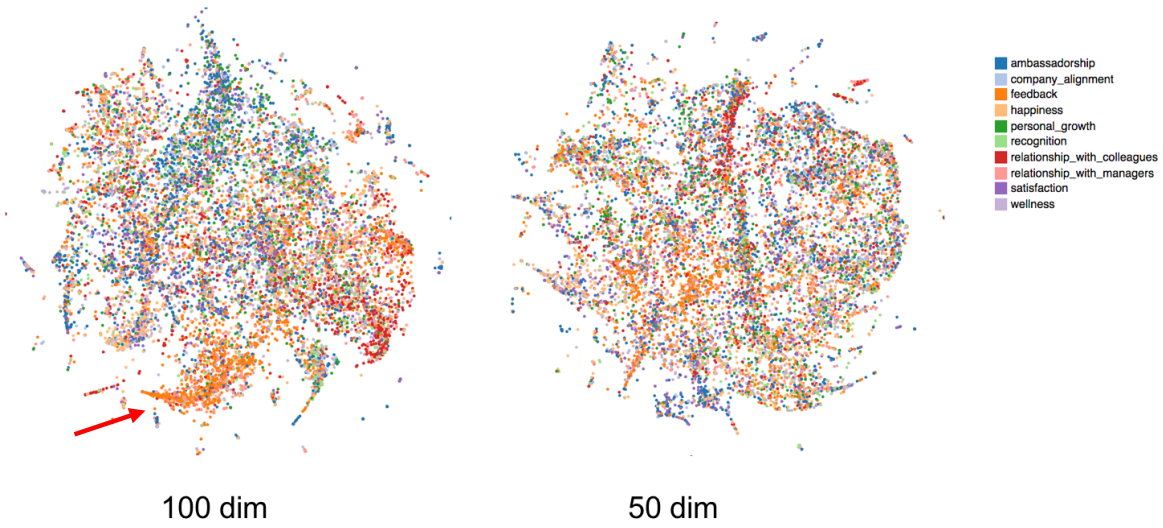


Figure 4-12 - t-SNE appliqué a une vectorisation LDA

On observe là aussi, des concentrations de métriques RH, par exemple pour la métrique "feedback".

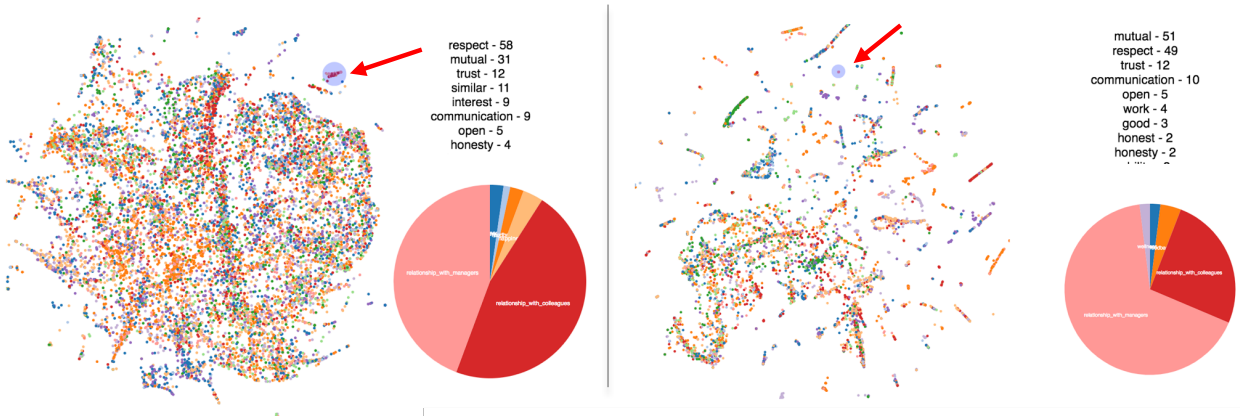


Figure 4-13 - Comparaison t-SNE de LDA 50 thèmes et Vectorisation poids MI 50 dimensions

On va retrouver des agrégats correspondant à des thèmes que l'on peut également identifier avec les modèles de poids. Ci-dessus, l'agrégat pointé dans les deux représentations correspond au thème "mutual respect".

Toutefois, les graphiques t-SNE résultant de LDA, ont une densité très homogène qui ne permet pas de rapidement identifier des concentrations sémantiques de feedbacks. Le modèle MI semble là encore plus approprié pour identifier de tels agrégats.

Contrairement aux autres modèles présentés, nous pouvons grâce à LDA obtenir des thèmes représentés par des distributions de mots. Ci-dessous, cinq thèmes sélectionnés parmi les 50 extraits, donnant une bonne idée du résultat de l'algorithme.

	MOT1	MOT2	MOT3	MOT4	MOT5	MOT6	MOT7	MOT8	MOT9	MOT10
THEME1	customer	service	enjoy	really	while	show	look	overall	effort	appreciation
THEME2	very	communication	open	supportive	cross	young	score	necessarily	minded	punished
THEME3	transparency	joke	testing	paper	admire	trickle	rolled	prompt	conflict	crash
THEME4	is	it	to	the	that	and	but	not	on	know
THEME5	help	support	to	grow	leadership	career	colleague	and	point	system

Table 4-1- Les 10 mots ordonnés par occurrence dans un thème, pour 5 thèmes sélectionnés de LDA à 50 dimensions

Les thèmes choisis sont représentatifs de la qualité des thèmes que l'on peut obtenir avec LDA sur notre corpus. Certains thèmes semblent avoir une tendance sémantique, comme le thème 2 qui tourne autour de la communication, mais si on regarde les mots moins fréquents, il est difficile de trouver un raisonnement valable.

De plus, on obtient des thèmes réunissant tous les mots courants du langage comme le thème 4 où l'on retrouve les mots "the", "it", "to", etc.

4.3 Regroupement automatisé de documents par DBSCAN

4.3.1 Sélection du meilleur modèle suivant nos critères

Nous présentons ici les résultats de DBSCAN. Nous avons 14 modèles différents avec une gamme de paramètres pour chacun, déterminés par une analyse de distance du K-voisin (voire méthodologie).

Nous filtrons aussi les regroupements par les contraintes suivantes:

- Un groupe doit contenir plus de documents que 1% du nombre moyen de documents contenu dans une métrique RH.

- Un groupe ne doit pas contenir plus de documents que 50% du nombre moyen de documents contenu dans une métrique RH.
- Un regroupement doit avoir au moins 20 groupes (deux fois le nombre de métriques RH).

Après filtrage, nous obtenons 364 regroupements. Nos deux critères d'évaluation de qualité sont le pourcentage de points regroupés et l'homogénéité moyenne des groupes de chaque regroupement.

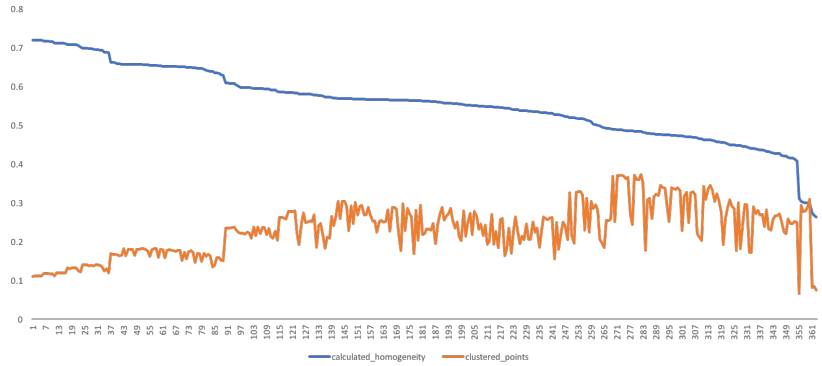


Figure 4-14- Homogénéité (bleu) vs % de points regroupés (orange)

Ci-dessus, nous avons ordonné les regroupements par leur homogénéité moyenne, du plus haut au plus bas. Nous y avons ajouté la courbe représentant le pourcentage de points regroupés. Nous voulons sélectionner le regroupement offrant le meilleur équilibre entre homogénéité et nombre de points regroupés. Ce graphique nous permet d'éliminer certains regroupements. En effet, nous considérons que pour deux regroupements ayant un critère égal, nous regardons celui qui a la plus haute valeur pour l'autre critère.

Nous avons conservé deux candidats qui respectivement étaient les meilleurs pour l'homogénéité et pour le nombre de points regroupés.

MODÈLE	MI LSI 50	MI LSI 200
POINTS REGROUPPÉS	37.11%	10.95%
HOMOGÉNÉITÉ	0.48	0.72
DISTANCE MIN	0.2	1.4
NBR POINTS MIN	3	15
NBR DE GROUPES	74	9

Figure 4-15 - Résultats DBSCAN

Dans les deux cas, ce sont les modèles de poids basé sur l'information mutuelle (MI) qui étaient les meilleurs candidats. Si nous choisissons des paramètres favorisant l'homogénéité, il semble que le nombre de groupes est fortement impacté. Ceci est déterminant dans notre choix final.

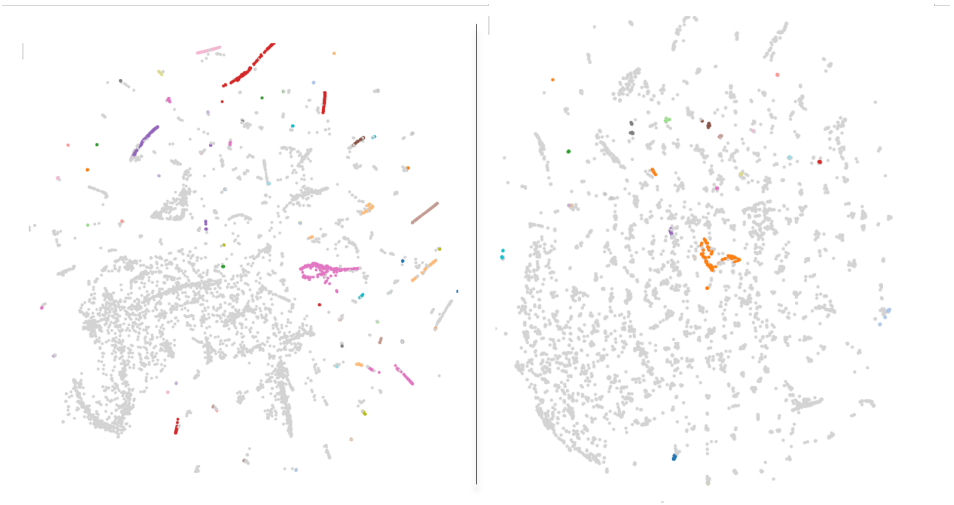


Figure 4-16- MI LSI 50 (gauche) vs MI LSI 200 (droite)
Les couleurs sont seulement utiles à distinguer les groupes, elles n'ont aucune signification
Les points gris sont les feedbacks non regroupés

Nous avons donc choisi le modèle à 50 dimensions pour la suite de notre analyse, plus qualitative.

4.3.2 Étude du regroupement du modèle sélectionné

Nous obtenons avec le modèle retenu (MI LSI 50), 79 groupes. Dans un premier temps, nous nous sommes intéressés aux métriques RH qui revenaient le plus dans chaque groupe. Nous rappelons qu'un des critères de sélection de modèles est l'homogénéité, nous pouvons donc nous attendre à des groupes avec de fortes concentrations de métriques RH.

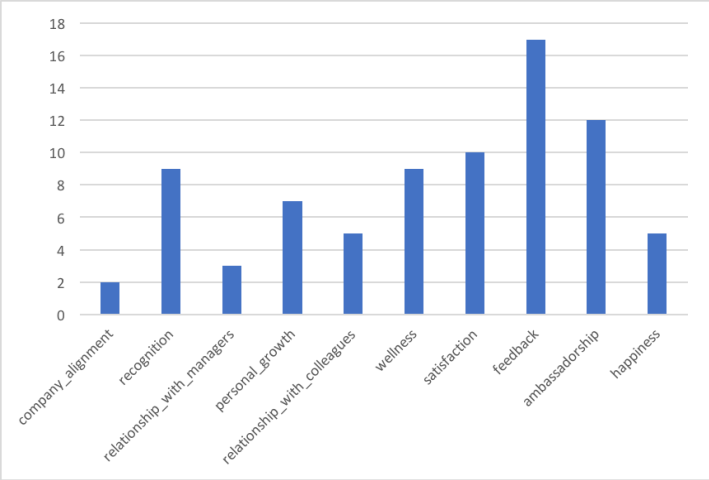


Figure 4-17 - Distribution de la métrique RH avec la plus grande occurrence dans chaque groupe

Nous pouvons observer ci-dessus que le nombre de groupes trouvés n'est pas proportionnel au nombre de feedbacks par métrique RH. En effet, seulement 5 groupes sont trouvés pour la métrique "Hapiness". Ceci confirme que les résultats DBSCAN sont dépendant de la densité des points et non du nombre de points dans un groupe.

MÉTRIQUE RH	MOT 1	MOT 2	MOT 3	BIGRAM 1	BIGRAM 2	TRIGRAM 1
AMBASSADORSHIP	culture	our	of	company_culture	our_culture	the_culture_of
COMPANY_ALIGNMENT	mission	value	company	value_mission	of_company	value_and_mission
FEEDBACK	feedback	my	that	feedback_from	my_manager	feedback_from_my
HAPPINESS	home	from	work	from_home	at_home	work_from_home
PERSONAL_GROWTH	training	on	for	training_on	training_opportunity	up_to_date
RECOGNITION	praise	my	in	praise_from	from_my	recognition_and_praise
RELATIONSHIP_WITH_COLLEAGUES	common	in	goal	common_goal	in_common	thing_in_common
RELATIONSHIP_WITH_MANAGERS	respect	mutual	trust	mutual_respect	respect_trust	mutual_respect_and
SATISFACTION	environment	work	great	work_environment	environment_great	the_work_environment
WELLNESS	wellness	program	with	wellness_program	new_wellness	the_wellness_program

Table 4-2- Sélection d'un groupe par métrique RH
Mot N signifie le mot avec la N-ième plus forte occurrence (e.g : Mot1 = Plus forte occurrence)
Bigram N signifie les deux mots consécutifs avec la N-ième plus forte occurrence
Trigram 1 signifie les trois mots consécutifs avec la plus forte occurrence

Nous avons sélectionné 10 groupes, un par métrique RH. Dans la colonne "Mot 1", nous trouvons les mots avec la plus grande occurrence dans chaque groupe. Il est intéressant d'observer que

plusieurs de ces mots correspondent aux mots que nous avons défini comme faiblement ambigus, tels "culture", "praise" ou "training". Nous avons aussi vérifié que les groupes avaient une forte homogénéité, comme présenté ci-dessous.

Figure 4-18 - Distribution des métriques RH par groupe

Finalement, nous avons regardé plus en détail les groupes de la métrique "Wellness" pour montrer que notre méthode automatisée permet d'extraire des thèmes que nous avons déjà étudié avec l'approche par exploration.

MOT 1	MOT 2	MOT 3	BIGRAM 1	BIGRAM 2	TRIGRAM 1
insurance	health	cost	health_insurance	cost_of	health_insurance_is
health	of	my	cost_of	health_care	the_cost_of
gym	to	access	access_to	gym_at	access_to_gym
wellness	program	with	wellness_program	new_wellness	the_wellness_program
break	time	lunch	break_time	of_our	to_take_break
healthy	of	to	healthy_lifestyle	for_healthy	-- *confidentiel* --
wellness	plan	company	wellness_plan	new_wellness	the_new_wellness
membership	gym	for	gym_membership	for_gym	for_gym_membership
health	wellness	my	health_wellness	for_my	health_and_wellness

Table 4-3- Groupes de la métrique "Wellness"

En fait, nous observons plusieurs groupes intéressants dont les mots phares sont "health", "wellness" et "program". Une analyse de l'occurrence des mots consécutifs (bigrams et trigrams) montre que le mot "health" est souvent composé pour former "health insurance" ou "healthy lifesytle", qui peuvent définir deux thèmes distincts. Nous pouvons observer également que le mot "wellness" est souvent composé avec "plan" ou "program".

En faisant la même analyse par métrique RH, nous avons réalisé que le modèle DBSCAN sélectionné a détecté les agrégats que nous avions remarqués visuellement par t-SNE, et de nouveaux agrégats que nous n'avions pas trouvé. Par exemple, le groupe de la métrique "Wellness" que nous avions exploré dans une des sections précédantes, avait une forte occurrence du bigram "break time".

5 Limites de la méthodologie

Nous pensons que les contraintes que nous avons fixées lorsque nous avons défini un thème dans notre contexte ont eu un impact sur les résultats. Nous tenterons d'expliquer pourquoi. Nous avons fait certains choix techniques, parfois au risque de ne pas approfondir certaines méthodes. Nous avons aussi réalisé que notre exploitation des données n'est pas optimale et nous avons pensé à d'autres possibilités.

5.1 Limites liées aux hypothèses

Nous avons orienté le choix des méthodes en partant du principe que nous cherchions en moyenne un thème par feedback. Nous obtenons des thèmes avec une petite granularité, des thèmes identifiables par un ou deux mots. L'approche que nous avons proposée peut être alors perçue comme l'inverse d'une recherche par mots clés, et non une tâche d'extraction de thèmes complexes.

Ceci nous amène à notre deuxième hypothèse, nous sommes dépendants du vocabulaire employé. Si nous obtenons deux agrégats avec des expressions distinctes désignant le même thème, nos méthodes détecteront plutôt deux thèmes distincts. En effet, l'approche par exploration t-SNE montre que deux agrégats proches dans l'espace 2D, ne signifie pas qu'ils sont proches sémantiquement. La méthode automatisée basée sur DBSCAN que nous proposons ne résout pas ce problème. Nous espérions que les feedbacks plus longs fassent émerger des thèmes plus complexes, mais ceci n'a pas été une évidence nous concernant.

Nos méthodes ne prennent pas en compte la variable temporelle, nous devions faire cette analyse séparément et par mots. Il fut difficile de conclure quoi que ce soit sur l'importance d'une tendance dans l'entreprise. Nous pensons que cette analyse dans le temps serait de toute façon plus pertinente si nous avions une idée des grands événements et changements dans l'entreprise.

Suite à notre approche par exploration, nous avons implicitement ajouté l'hypothèse qu'un thème est probablement contenu dans une ou deux métriques RH. Nous avons développé un modèle de poids de mots prenant en compte cette information. Lorsque nous avons automatisé l'extraction de thèmes par DBSCAN, nous nous sommes basés sur l'homogénéité, qui

est lié à notre modèle par l'entropie conditionnelle d'une métrique étant donné un terme. Les résultats ont montré que c'est ce modèle qui performait le mieux suivant ce critère. Même si nous pensons que ceci n'est pas trivial, ce n'est pas non plus une surprise.

5.2 Limites techniques

Nous n'avons pas réalisé une étude comparative poussée pour sélectionner toutes les méthodes et tous les paramètres. Par exemple, nous aurions pu utiliser d'autres mesures de distance de vecteurs comme l'index de Jaccard. Nous avons fait des choix de paramétrage LSI et LDA (e.g : 50, 100 et 200 dimensions) approximatifs, en suivant uniquement les recommandations de la littérature.

Pour des raisons d'implémentation, nous n'avons pas utilisé des variantes de LDA pour les textes courts, comme le modèle Biterm (Yan, Guo, Lan and Cheng 2013), oubien le modèle LabeledLDA (Ramage, Hall et al. 2009) qui propose de modéliser le problème par des variables explicatives. Nous avons rencontré plusieurs problèmes avec les implémentations de LDA, il nous aurait fallu implémenter nous-mêmes ces algorithmes pour avoir confiance dans les résultats. Ce n'était pas l'objectif de ce mémoire.

Aujourd'hui, il existe une tendance vers les modèles de langages tentant de modéliser la séquence de mots d'un document. Nous avons fait le choix d'utiliser des modèles "sac de mots" plus classiques. Nos modèles ont l'hypothèse forte que l'ordre des mots dans un document importe peu. Nous continuons de penser que c'est une bonne hypothèse pour les textes courts. Nous aurions pu tout de même comparer nos méthodes de vectorisation des documents à des modèles plus sophistiqués, comme des réseaux de neurones artificiels.

5.3 Limites liées aux données

Nous aurions pu exploiter davantage les données fournies, car nous avions des informations additionnelles telles, l'identifiant de l'employé ou la réponse à choix multiples associée au feedback textuel. Toutefois, nous avons remarqué que pour ce dernier, il ne semblait pas toujours cohérent par rapport au contenu du feedback. Puisque nous avons peu d'informations sur l'algorithme générateur de questions incitant l'employé à donner un feedback,

nous nous sommes concentrés sur les métriques RH. D'ailleurs, contrairement à ce qui nous avait été indiqué, la distribution de ces métriques sur l'ensemble du corpus n'était pas uniforme, nous n'avons pas tenté de corriger ceci.

Finalement, nous pensons que plus de questions diversifiées, donc plus de métriques RH, nous auraient permis d'aller plus loin dans notre analyse.

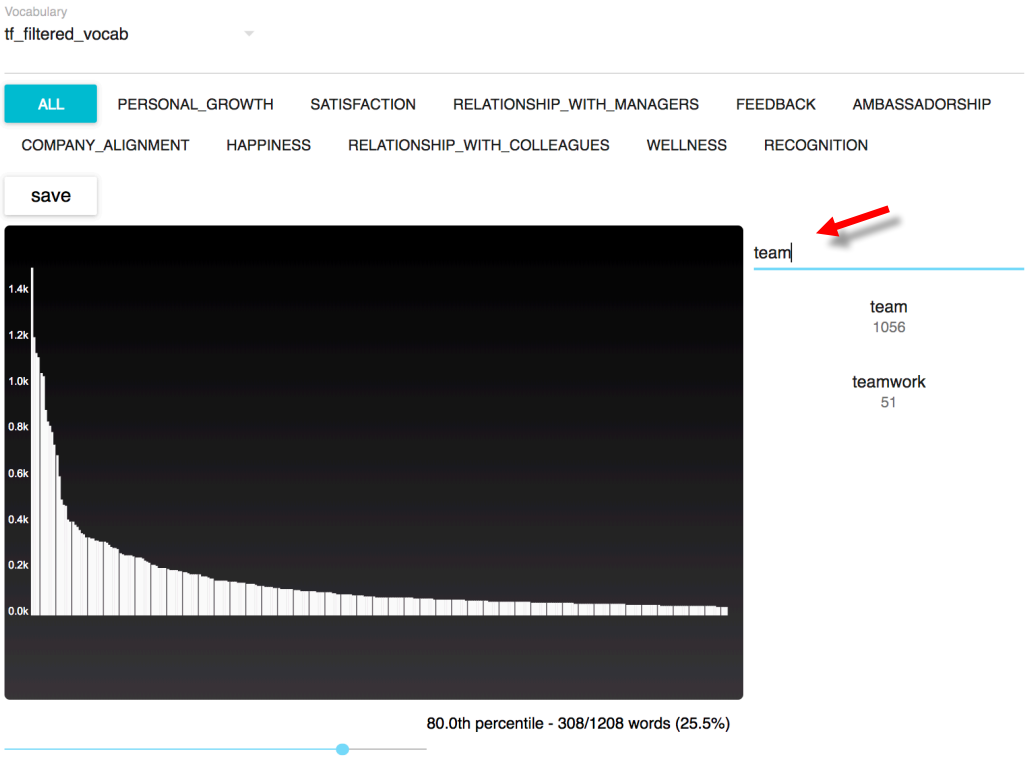
6 Conclusion

Le but de cette étude était de trouver des méthodes afin d'apporter une information nouvelle et utile aux gestionnaires d'une entreprise cherchant à exploiter des milliers de feedbacks. Nous avons considéré une approche par exploration permettant à un analyste d'aller plus loin qu'une simple recherche par mot clé. Nous avons utilisé des méthodes de traitement du langage de la littérature, couplé à un outil de visualisation pour permettre une navigation innovante d'un corpus. Nous avons aussi considéré une méthode automatisée de regroupement, ne nécessitant pas d'expertise en analyse de données, dont l'objectif est de trouver des groupes sémantiques de feedbacks.

Le contexte particulier de nos données nous a poussé à prendre certaines hypothèses sur le type de thèmes que nous pouvions extraire. Nous avons utilisé des métriques issues du domaine de la gestion des ressources humaines, définies par les concepteurs de l'outil de collecte de données, pour guider notre analyse. Nous avons observé que les groupes de feedbacks qui se distinguaient sémantiquement étaient souvent corrélés à une ou deux métriques. Nous avons développé un modèle de poids de mots intégrant ces informations. Il se base sur l'information mutuelle entre un terme et une métrique RH.

Nous avons choisi les paramètres de l'algorithme de regroupement, DBSCAN, en utilisant la même hypothèse. C'est certainement un des biais de notre méthodologie, mais les résultats ont confirmé que le modèle de poids que nous avons développé permettait d'extraire davantage de groupes avec une forte homogénéité de métriques RH. Si l'objectif d'une entreprise est de trouver de tels thèmes, notre approche peut s'avérer être efficace. Finalement, notre analyse fut limitée par le manque d'information temporelle sur l'entreprise. Il fut difficile d'évaluer l'importance d'un thème dans notre corpus sans pouvoir le situer correctement dans le temps. Il serait intéressant de continuer cette étude en intégrant en plus la variable temporelle.

Annexe 1



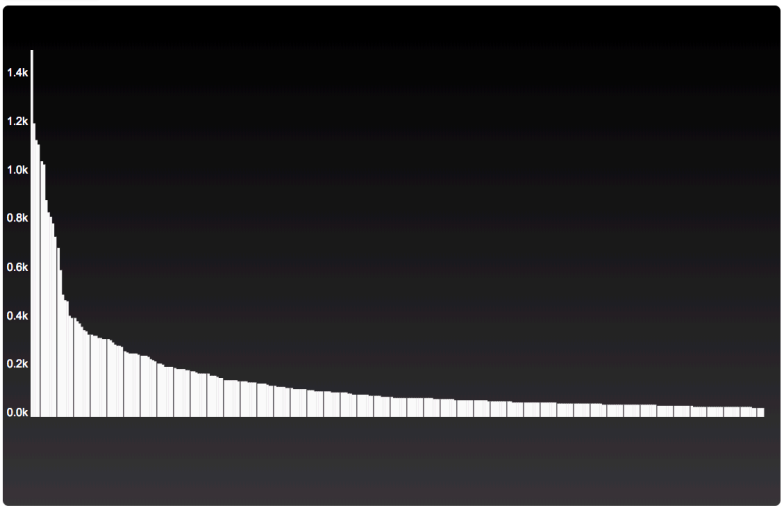
Sur l'interface ci-dessus, le menu déroulant en haut à gauche permet de sélectionner la stratégie de vocabulaire, "tf_filtered_vocab" veut dire que le corpus a été filtré et qu'on utilise une vectorisation TF. On a conservé uniquement les noms et les adjectifs. La liste à droite de graphe correspond aux mots ordonnés par leur poids, ici filtrés par le mot "team". On peut filtrer par occurrence, par exemple on ne voit que 25.5% des mots pour le 80e pourcentile. Finalement, la liste des métriques permet de filtrer la distribution des mots par métrique RH. Ci-dessous, un exemple de filtrage par métrique.

Vocabulary

tf_filtered_vocab

ALL | PERSONAL_GROWTH | SATISFACTION | RELATIONSHIP_WITH_MANAGERS | FEEDBACK | AMBASSADORSHIP | COMPANY_ALIGNMENT | HAPPINESS | RELATIONSHIP_WITH_COLLEAGUES | WELLNESS | RECOGNITION

save



work
1512

company
1211

employee
1143

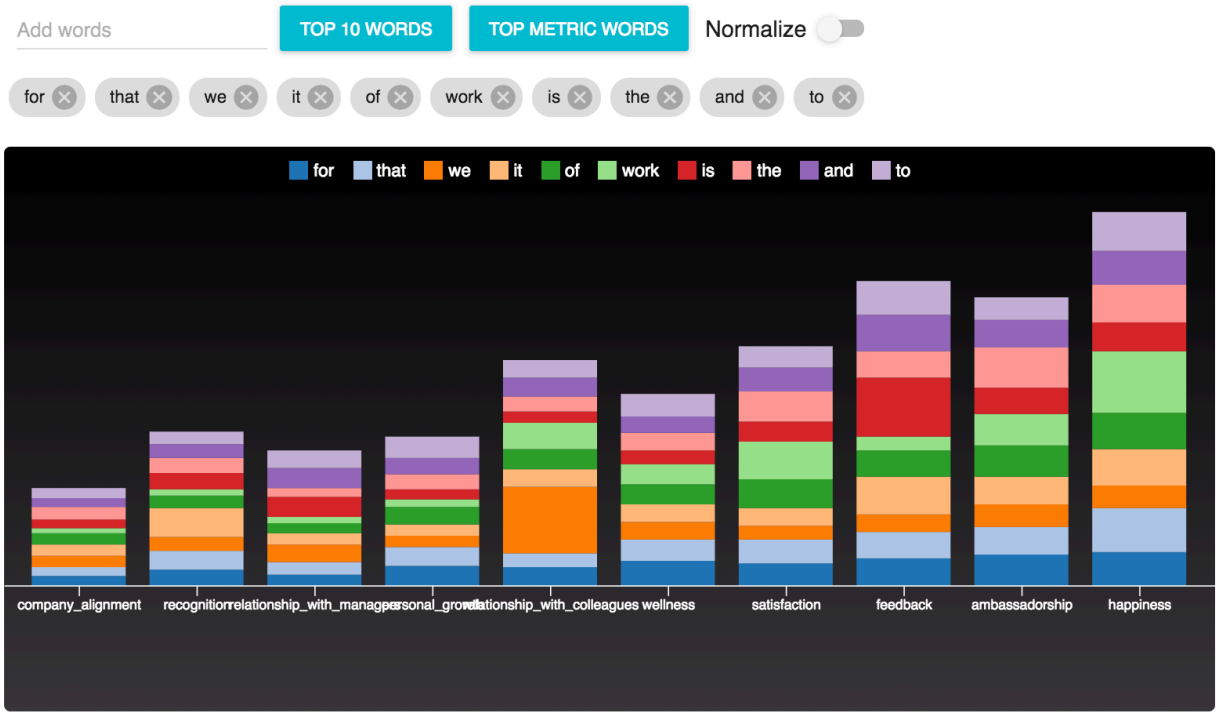
great
1124

team
1056

time

80.0th percentile - 308/1208 words (25.5%)

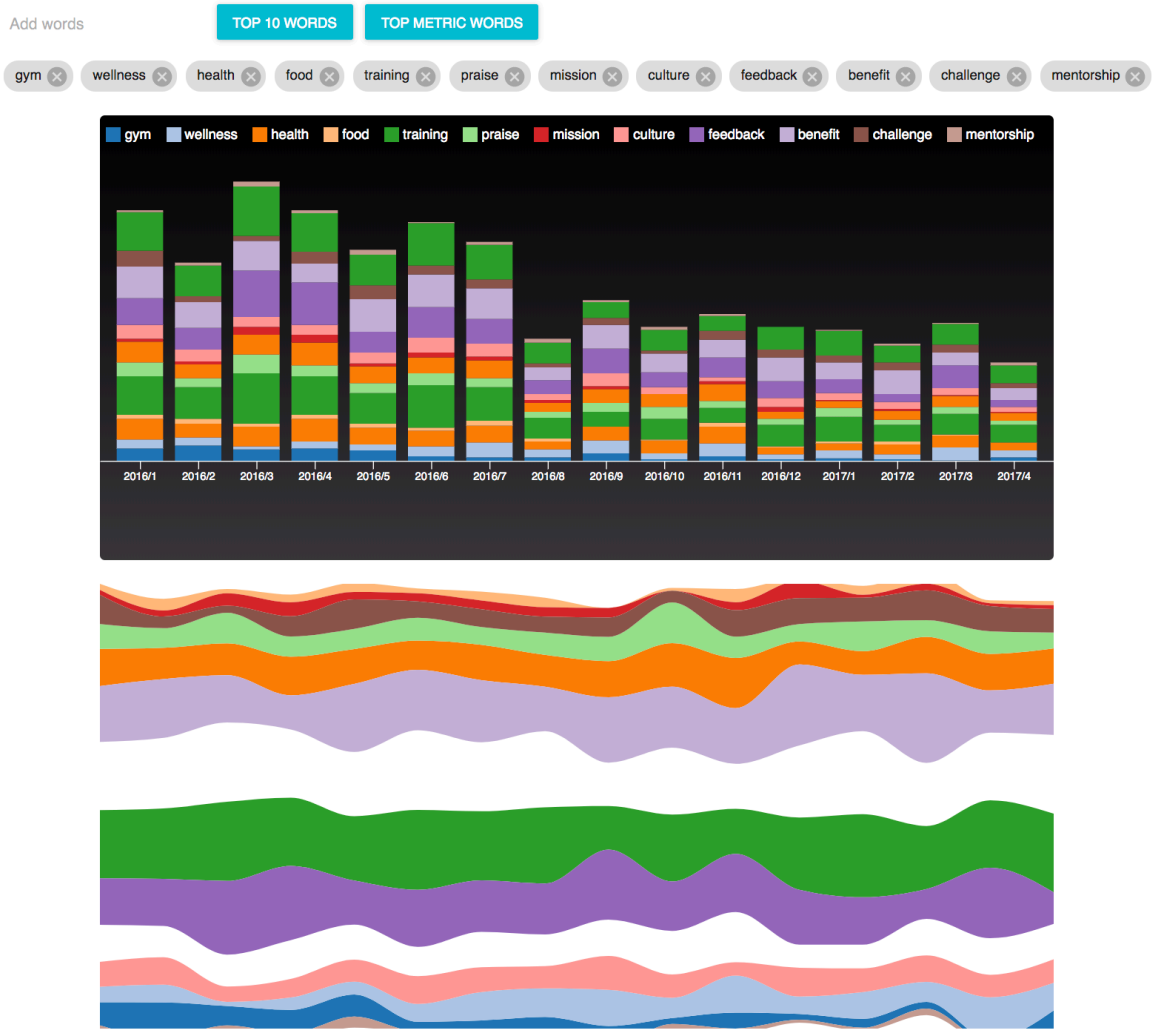
Annexe 2



Sur cette visualisation interactive, on peut ajouter un mot par le champ en haut à gauche ou bien sélectionner les 10 mots revenant le plus souvent, soit dans tout le corpus, soit par métrique RH.

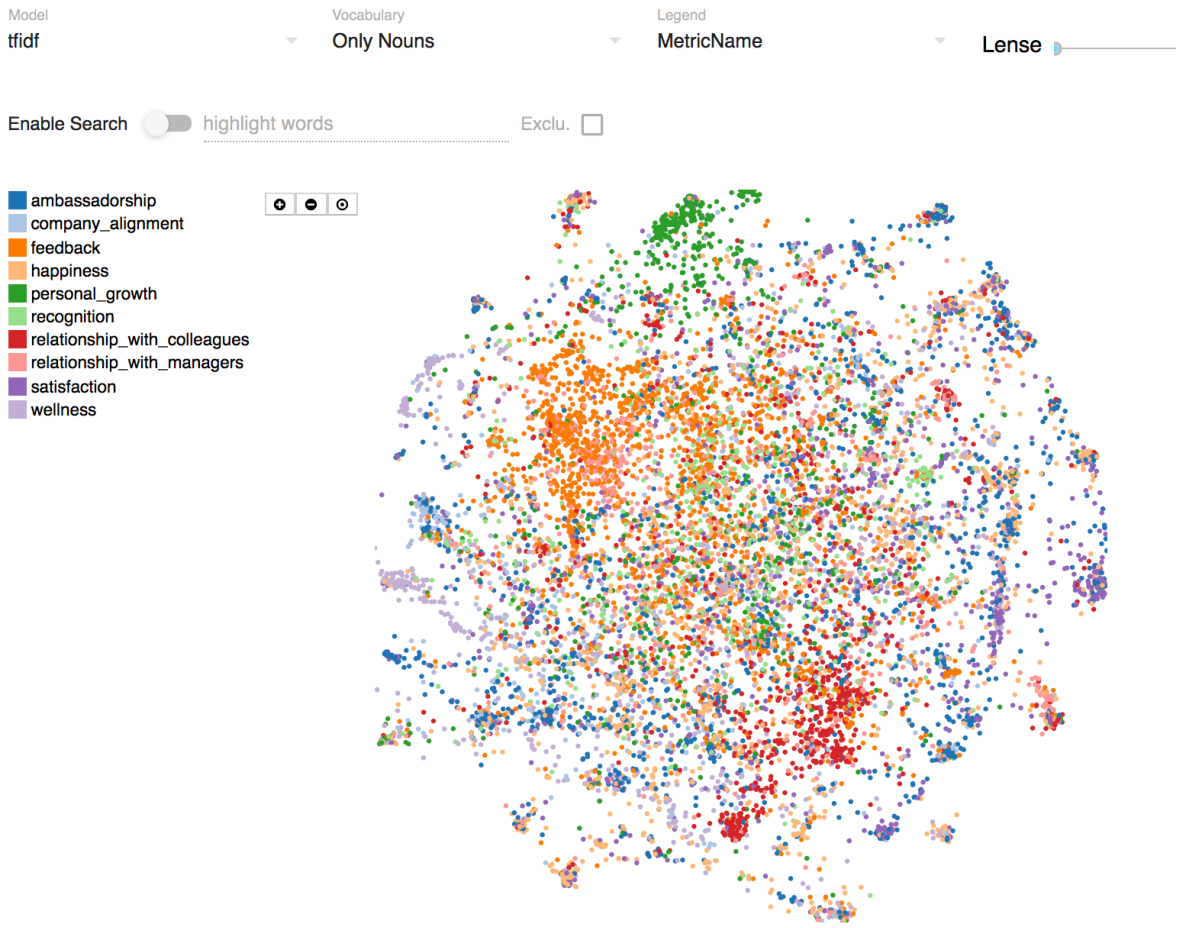
Ici, on peut observer par métrique les 10 mots les plus fréquents dans le corpus lorsqu'il n'est pas filtré par catégorie de mot. Il est intéressant de noter que ces mots cumulés reforment la distribution des métriques dans le corpus, "Happiness" par exemple est la métrique avec le plus d'occurrences. Ceci n'est pas le cas si on insère des mots moins ambigus.

Annexe 3

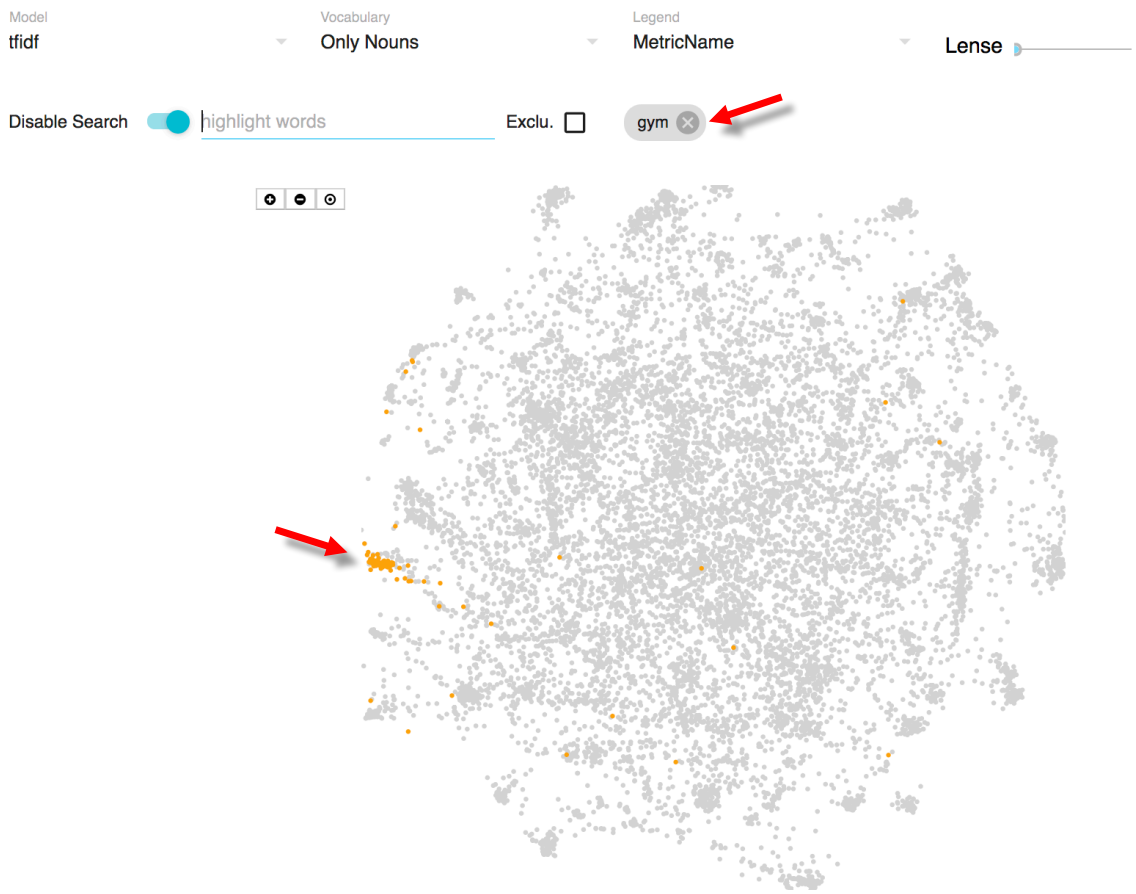


Le premier graphique à barres montre l'évolution de l'occurrence d'une sélection de mots. On observe que globalement l'occurrence baisse dans le temps. Le deuxième graphique permet d'isoler chaque mot visuellement. Lorsqu'une bande correspondant à un mot s'amincit, le mot perd en occurrence proportionnellement aux autres mots à ce moment.

Annexe 4

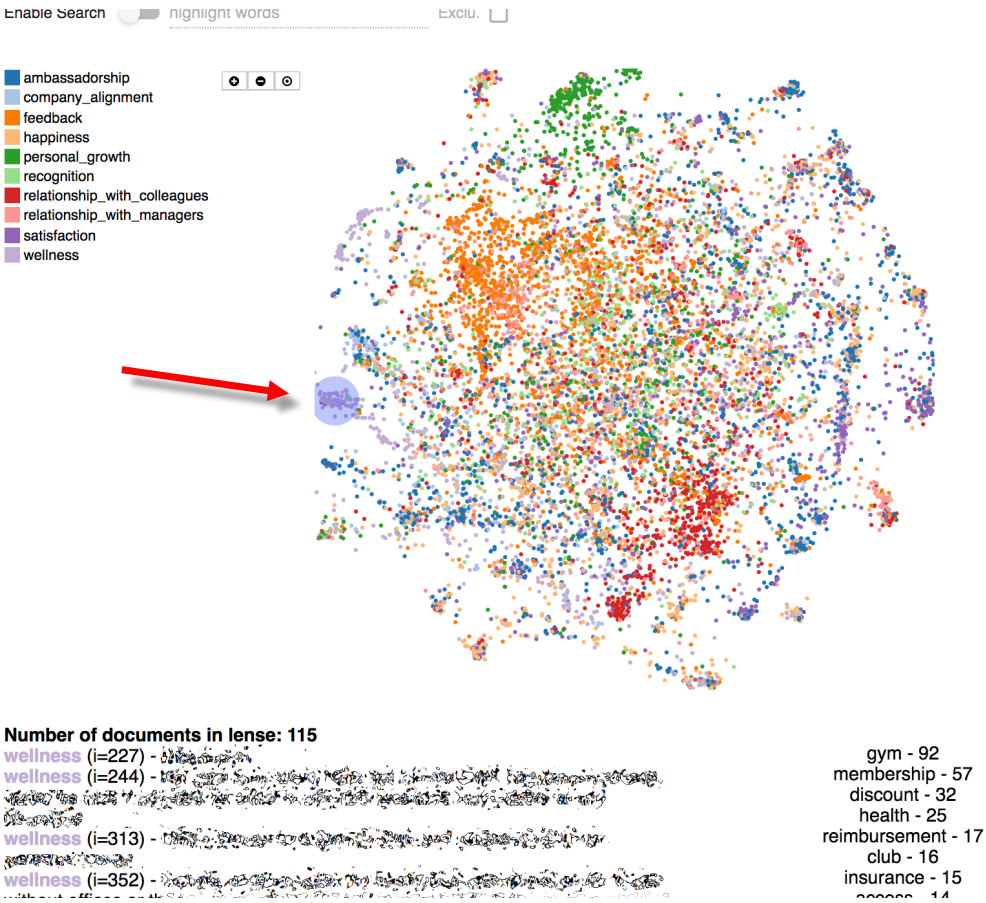


Sur cette visualisation, on observe le nuage de points correspondants aux documents, colorés par les métriques. Chaque point correspond donc à un feedback. On peut, là aussi, changer la stratégie de vocabulaire par le menu déroulant en haut à gauche, ainsi que le modèle de poids de mots par le menu déroulant à sa droite. On peut également zoomer sur le graphique.



Nous avons développé une recherche par mot clé qui permet de voir où se situe un mot dans le nuage de points. Sur la capture ci-dessus, on peut observer où se situe le mot "gym", les feedbacks qui contiennent ce mot sont coloriés en orange.

Finalement, nous avons développé une lentille permettant de voir le contenu des feedbacks se trouvant à l'intérieur du cercle bleu pointé par la flèche rouge ci-dessous.



Exemple d'utilisation de la lentille - Partie1

La deuxième partie montre plus de détails du contenu. Nous avons notamment ajouté la possibilité de voir l'occurrence des mots des feedbacks se trouvant dans la lentille et la distribution des métriques associée à ces feedbacks. Ici, on peut voir que "gym" est le mot qui revient le plus et la métrique RH "wellness" est dominante.

Number of documents in lense: 115

wellness (i=227) -

wellness (i=244) -

wellness (i=313) -

wellness (i=352) - without offices or th

wellness (i=362) -

wellness (i=431) -

wellness (i=444) -

wellness (i=550) -

wellness (i=715) -

wellness (i=717) -

wellness (i=792) -

wellness (i=891) -

wellness (i=921) -

wellness (i=930) -

wellness (i=985) -

wellness (i=1063) -

wellness (i=1073) -

wellness (i=1118) -

wellness (i=1125) -

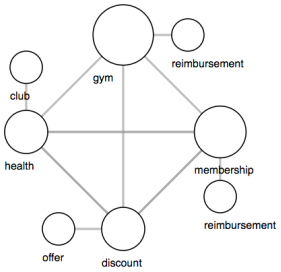
wellness (i=1177) -

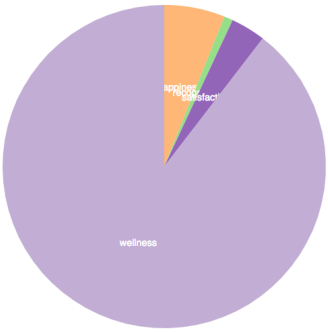
wellness (i=1202) -

wellness (i=1387) -

wellness (i=1560) -

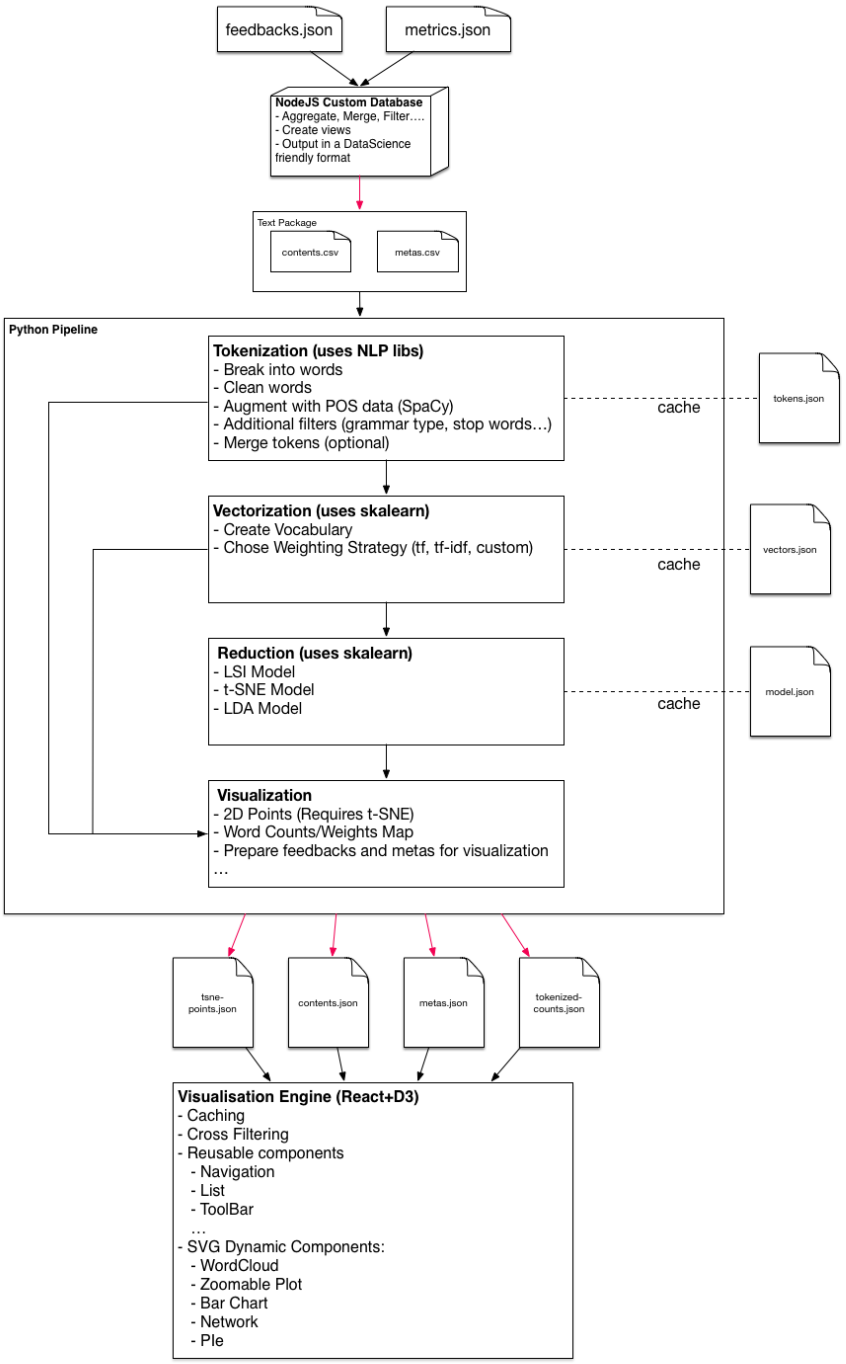
gym - 92
membership - 57
discount - 32
health - 25
reimbursement - 17
club - 16
insurance - 15
access - 14
free - 12
fitness - 12
offer - 11
employee - 11





Exemple d'utilisation de la lentille - Partie2

Annexe 5


feedbacks.json
metrics.json
NodeJS Custom Database
- Aggregate, Merge, Filter….
- Create views
- Output in a DataScience friendly format
Text Package
contents.csv
metas.csv
Python Pipeline
Tokenization (uses NLP libs)
- Break into words
- Clean words
- Augment with POS data (SpaCy)
- Additional filters (grammar type, stop words…)
- Merge tokens (optional)
cache
tokens.json
Vectorization (uses skalearn)
- Create Vocabulary
- Chose Weighting Strategy (tf, tf-idf, custom)
cache
vectors.json
Reduction (uses skalearn)
- LSI Model
- t-SNE Model
- LDA Model
cache
model.json
Visualization
- 2D Points (Requires t-SNE)
- Word Counts/Weights Map
- Prepare feedbacks and metas for visualization
…
tsne-points.json
contents.json
metas.json
tokenized-counts.json
Visualisation Engine (React+D3)
- Caching
- Cross Filtering
- Reusable components
- Navigation
- List
- ToolBar
…
- SVG Dynamic Components:
- WordCloud
- Zoomable Plot
- Bar Chart
- Network
- Ple

Bibliographie

Andor, Daniel, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov and Michael Collins (2016). "Globally normalized transition-based neural networks." arXiv preprint arXiv:1603.06042.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (1999). "Modern information retrieval", ACM press New York, **463**.

Birant, Derya and Alp Kut (2007). "ST-DBSCAN: An algorithm for clustering spatial–temporal data." Data & Knowledge Engineering **60**(1): 208-221.

Bishop, Christopher M (2006). "Pattern recognition and machine learning", Book published by springer.

Blei, David M, Andrew Y Ng and Michael I Jordan (2003). "Latent dirichlet allocation." Journal of machine Learning research **3**(Jan): 993-1022.

Cui, Weiwei, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu and Xin Tong (2011). "Textflow: Towards better understanding of evolving topics in text." IEEE transactions on visualization and computer graphics **17**(12): 2412-2421.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer and Richard Harshman (1990). "Indexing by latent semantic analysis." Journal of the American society for information science **41**(6): 391.

Grimmer, Justin and Brandon M Stewart (2013). "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". Political analysis **21**(3): 267-297.

Kusner, Matt, Yu Sun, Nicholas Kolkin and Kilian Weinberger (2015). "From word embeddings to document distances". International Conference on Machine Learning, pages 957-966.

Le, Quoc and Tomas Mikolov (2014). "Distributed representations of sentences and documents". Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1188--1196.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE." Journal of machine Learning research **9**(Nov): 2579-2605.

MacQueen, James (1967). "Some methods for classification and analysis of multivariate observations". Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA.

Manning, Christopher D and Hinrich Schütze (1999). "Foundations of statistical natural language processing", MIT Press.

McCandless, Michael, Erik Hatcher and Otis Gospodnetic (2010). "Lucene in Action: Covers Apache Lucene 3.0", Manning Publications Co.

Miller, George A (1995). "WordNet: a lexical database for English." Communications of the ACM **38**(11): 39-41.

Powell, Thomas C and Anne Dent-Micallef (1997). "Information technology as competitive advantage: The role of human, business, and technology resources." Strategic management journal: 375-405.

Ramage, Daniel, David Hall, Ramesh Nallapati and Christopher D Manning (2009). "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora". Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics.

Ramlall, Sunil (2004). "A review of employee motivation theories and their implications for employee retention within organizations." Journal of American Academy of Business **5**(1/2): 52-63.

Rosenberg, Andrew and Julia Hirschberg (2007). "V-measure: A conditional entropy-based external cluster evaluation measure". Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL).

Rust, Roland T, Greg L Stewart, Heather Miller and Debbie Pielack (1996). "The satisfaction and retention of frontline employees: A customer satisfaction measurement approach." International Journal of Service Industry Management **7**(5): 62-80.

Salton, Gerard and Christopher Buckley (1988). "Term-weighting approaches in automatic text retrieval." Information processing & management **24**(5): 513-523.

Sanchez, Paul M (2007). "The employee survey: more than asking questions." Journal of Business Strategy **28**(2): 48-56.

Steinbach, Michael, George Karypis and Vipin Kumar (2000). "A comparison of document clustering techniques". KDD workshop on text mining, Boston. **400**(1): 525-526.

Yan, Xiaohui, Jiafeng Guo, Yanyan Lan and Xueqi Cheng (2013). "A biterm topic model for short texts". Proceedings of the 22nd international conference on World Wide Web, ACM. 1445-1456.

Zipf, George K (1935). "The psychology of language." NY Houghton-Mifflin.