HEC MONTRÉAL

Compétition étude de cas :

Outils de classification sémantique en recherche documentaire

Corpus de publications scientifiques portant sur la Covid-19

Par

Gabriel Boulanger Théberge

Simon Tye-Giguère

15 janvier 2021

Introduction

Le contexte de pandémie actuel a pour conséquence la production d'un volume colossal de travail par la communauté scientifique et ce, dans un large éventail de domaines du savoir. La quantité de publications scientifiques qui en résulte est tout aussi importante et la consultation et le suivi de ces travaux par les experts et les décideurs publics représente en soi un défi.

Les moteurs de recherche traditionnels employés en recherche documentaire fonctionnent généralement par mots clés et ne sont par le fait même pas adapté à la recherche par thème. Ce genre de moteur identifie les ouvrages contenant les mots de recherche indiqués par l'utilisateur et ordonne les résultats en fonction de la fréquence des mots de recherche contenus dans les ouvrages et ce, sans tenir compte du sens de ces mots. Ainsi, les ouvrages portant sur un thème similaire à celui recherché par l'utilisateur, mais ne contenant pas les mots exacts de l'utilisateur ne seront pas retournés par le moteur de recherche. Ceci a pour conséquence de limiter la possibilité de découvrir des ouvrages sur un thème donné qui emploieraient un vocabulaire différent. Notre ambition dans ce projet fut de construire un outil de recherche exploitant davantage la sémantique pour permettre à un utilisateur d'identifier, parmi un large corpus scientifique, les articles correspondants à un thème spécifique. À partir de notre outil, un utilisateur peut entrer une ou plusieurs phrases résumant le thème de sa recherche et l'outil lui retournera une liste d'articles classés par proximité sémantique avec la ou les phrases qu'il aura fournies. Pour ce faire, notre outil exploite des algorithmes fréquemment rencontrés dans le domaine du *topic modeling*, une branche de la famille des techniques de traitement du langage naturel.

Notre outil a été construit et entrainé à partir de la base de données CORD-19 (Covid-19 Open Research Dataset). Cette base de données assemblée par *Allen Institute for AI* contient plus de 350 000 articles scientifiques portant sur la recherche liée à la Covid-19 et est mise à jour sur une base régulière pour intégrer les nouveaux travaux sur le sujet. Notre travail de modélisation porte plus spécifiquement sur les résumés (*abstracts*) des articles contenus dans cette base.

Le compte rendu de notre projet va comme suit. La prochaine section présente une revue des concepts employés dans ce projet pour introduire les lecteurs non familiers à ces concepts. La section 3 porte sur la méthodologie élaborée pour la conception de l'outil alors que les performances de même que les avantages et les limites de celui-ci sont discutés en détail dans la section 4. La section 5 conclue sur notre projet.

Revue des concepts

Clustering de documents

Le clustering de documents est une méthode de partitionnement non supervisée utilisée notamment pour la recherche de documents similaires, la réorganisation d'un gros volume de texte et la détection de contenu redondant. Les méthodes de clustering de documents ont commencé à être utilisées dans les années 1970. Les auteurs (Jardin & Van Risjbergen, 1971) déclarent que des documents similaires tendent à être dans le même groupe. Le clustering de document opère sur la notion d'intersimilarité. Initialement, cette technique était utilisée pour trouver rapidement des documents similaires (Buckley & Lewitt, 1985).

Les auteurs (Singh, Tiwari, & Garg, 2011) utilisent la méthode K-Means pour effectuer un clustering de document. L'algorithme K-Means a pour objectif de partitionner un ensemble de données en k cluster (grappes) contenant des observations partageant des caractéristiques semblables. La ressemblance entre les observations est basée sur une mesure de distance entre chaque observation et le centre de son cluster d'appartenance. L'algorithme cherche itérativement à identifier les coordonnées de k centre de clusters tel que la somme des distances de toutes les observations avec leur centre soit minimisée. Les mêmes auteurs rapportent que l'usage de la stemmatisation de même que l'implémentation de l'algorithme k-means sur une matrice TF-IDF plutôt qu'une matrice « bag of words » améliore les résultats.

Matrices TF et TF-IDF

Pour qu'un corpus de textes rédigés en langage naturel puisse être interprété par les algorithmes de topic modeling, celui-ci doit être mis sous la forme d'une représentation vectorielle. Deux des représentations vectorielles les plus communes sont la matrice de compte (TF) et la matrice TF-IDF. La première, parfois appelée *bags of words,* est une matrice dans laquelle chaque rangée représente un document du corpus, et chaque colonne l'ensemble des mots unique de ce corpus. Les valeurs de cette matrice sont simplement les fréquences des mots dans chaque document. La matrice TF-IDF est une matrice de dimensions identiques et pour laquelle chaque cellule représente le score TF-IDF d'un mot dans un document. Ce score est obtenu à partir du produit de la fréquence du mot dans le texte (TF) et de l'inverse du log du nombre de documents contenant ce mot (IDF).

Topic Modelling

La structuration d'un corpus de documents peut aussi être faite en analysant la distribution de thèmes à l'intérieur de ce dernier. La différence principale entre le *topic modeling* et le *document clustering* est que le premier effectue un soft clustering, déterminant à travers le corpus entier des groupes de mots pour chaque topic. Chaque mot a sa probabilité d'occurrence dans un thème/topic, et chaque document se voit attribuer une pondération pour chaque thème. Ceci distingue les méthodes de topic modeling des méthodes de clustering dans lesquelles les documents sont regroupés dans des clusters distincts via une métrique de distance tel que Jacquard, Minkowski ou Euclide.

Le topic modeling commence dans les années 1980 avec le *Latent Semantic Analysis (*LSA*)* introduit par (Deerwester & al., 1990). Le LSA est capable de grouper des documents ensemble en posant l'hypothèse que des mots similaires d'un point de vue sémantique tendent à apparaître ensemble. Les auteurs (Blei & al., 2003) développent, à partir du LSA, une méthode intitulée *Latent Dirichlet Allocation* (**LDA**). Cette méthode fréquemment employée en topic modeling fonctionne de la manière suivante. Après avoir choisi un nombre de topic, chaque mot du corpus se voit attribuer un de ces topics de manière aléatoire. Par la suite, deux mesures seront calculées, et ce pour chaque mot et pour chaque document. On calcule d'une part la probabilité que le document appartienne à un certain topic. Cette mesure est basée sur le nombre de mots de ce document qui appartiennent au topic du mot évalué. Plus ce nombre est élevé, plus le topic est important pour ce document. On répète pour tous les mots du document. Ensuite, on calcule la proportion des documents qui sont assignés au topic du mot évalué en raison de ce mot. On obtient ainsi l'importance de ce mot pour ce topic. On répète ces deux opérations sur tous les mots de tous les documents jusqu'à convergence, c'est-à-dire jusqu'à ce que les mots formant un topic et les topics assigné aux documents ne changent plus. On obtient alors un vecteur de pondération de topics pour chaque article.

Le topic modeling est utilisé à diverses fins. Il est employé par (Yau & al., 2014) pour décomposer un large corpus d'articles scientifiques en différentes disciplines académiques. Il est utilisé également à des fins de classifications (Zheng & al., 2006) (Suominen & Toivanen, 2016), ainsi que pour cartographier les thèmes récurrents dans un ensemble de documents (Székely & Brocke, 2017).

Il est avéré que le prétraitement permet d'améliorer les performances du topic modeling. Les auteurs (Székely & Brocke, 2017) en détaillent les 7 principales étapes telles que la lemmatisation, la stemmatisation et le retrait des mots non significatifs, appelés *stop words* en anglais. La lemmatisation consiste à ramener un mot à sa forme canonique également appelé lemme. Ceci permet de ramener à une même forme les mots d'une même famille partageant une sémantique commune. La stemmatisation pour sa part consiste à retrancher un préfixe et/ou un suffixe à un mot pour le ramener à une forme plus simple ou à son *stem* (tige). La stemmatisation ne considère pas le sens du mot, mais uniquement sa forme. Il est donc moins précis que la lemmatisation, mais a une plus grande portée que ce dernier (Manning, Raghavan, & Schütze, 2008). La stemmatisation peut retourner un mot qui n'existe pas, alors que la lemmatisation retournera toujours un mot existant.

Les méthodes de topic modeling requièrent le choix d'un nombre de thèmes, ce qui est considéré comme une des limitations de ces modèles par (Yau & al., 2014). Même si des métriques de cohérence existent pour identifier le nombre de topics optimal telles que la métrique c_v, (Carter & al, 2016) en concluent que ce nombre de topic est davantage fonction du contexte. Les auteurs (Suominen & Toivanen, 2016) soutiennent qu'en plus des scores de cohérence, une analyse qualitative des résultats demeure nécessaire.

En somme, la structuration d'un large corpus de documents peut se faire soit en regroupant les documents par similarité ou par niveau d'appartenance à des thèmes sous-jacents au corpus.

Méthodologie

Notre objectif dans ce projet est de construire un outil permettant à un utilisateur d'entrer une portion de texte tel qu'une phrase ou un petit paragraphe, et d'obtenir une liste de n articles scientifiques dont le thème se rapproche de celui de la portion de texte entré par l'utilisateur. Pour y arriver, nous avons eu recours à deux approches différentes, soit une basée sur l'algorithme LDA et une basée sur l'algorithme K-Means. Pour rappel, pour fin de simplification, notre modélisation porte uniquement sur les abstracts et non sur les articles en entier. Ce choix s'est imposé principalement parce que pour la majorité des articles contenus dans la base de données, seuls les abstracts sont disponibles en accès libre.

Dans cette section, nous présentons d'abord le prétraitement de la base de données et des textes analysés. Nous présentons ensuite la conception des outils de recherche à partir respectivement du LDA et du K-Means. Finalement, nous expliquons de quelle manière cet outil peut être employé comme système de recommandation pour les chercheurs et les décideurs publics.

Prétraitement des données

Les étapes suivantes sont réalisées sur les résumés de notre base de données avant le déploiement des méthodes de topic modeling. La première étape retient uniquement les articles contenant un résumé de plus de 100 caractères et de langue anglaise, et retire les doublons. Ceci réduit le nombre d'articles du corpus à environs 237 000. Ensuite sont retirés les caractères spéciaux, les nombres, la ponctuation et les caractères uniques. L'étape qui suit consiste à *tokeniser* les textes, c'est-à-dire à décomposer les textes en une série de mots. Les listes de *tokens* peuvent être soit lemmatisées ou stemmatisées. Dans le cadre de notre projet, nous avons opté pour la stemmatisation puisqu'elle permet de réduire de manière simple et efficace la quantité de mots utilisés dans les matrices *Bag of Words*. Ensuite, les *stopwords* sont retirés du texte selon deux méthodes. La première consiste à retirer les *stopwords* contenus dans une liste prédéfinie. La seconde consiste à y ajouter les mots qui se retrouvent dans 90% ou plus des abstracts de notre corpus. Après ces étapes, nous obtenons les listes de mots prêtes à être vectorisées.

Représentation vectorielle des textes

Deux représentations vectorielles des textes sont employées pour alimenter les différentes méthodes de topic modeling. La première est la matrice de décompte de mots et la seconde la matrice TF-IDF. Durant la modélisation, la méthode K-Means prend en entrée la matrice TF-IDF alors que la méthode LDA prend en entrée la matrice de fréquence.

Sélection d'un échantillon d'abstracts

Pour réduire le temps nécessaire à l'entraînement des modèles, nous avons utilisé un sous-ensemble de 50 000 abstracts sélectionnés aléatoirement, ainsi qu'un nombre de topics équivalent à 1500 pour les deux méthodes. Évidemment, le modèle final utilisera l'ensemble des articles et permettra une recherche à travers l'ensemble du corpus.

Nombre de topics optimal

L'identification du nombre de topic optimal est généralement une étape importante en modélisation par regroupement et des techniques comme le *elbow method* et la *silhouette method* peuvent notamment permettre d'identifier ce nombre optimal. Cependant, le contexte de notre projet ne se prêtait pas à cette approche, et ce pour deux raisons. D'une part, le nombre très élevé de features (plus de 50 000 termes) de nos modèles fait en sorte que la courbe de distances intra-cluster est une droite à pente négative relativement constante. Ainsi, la *elbow method* n'affiche aucun « coudes » pour toutes valeurs de k testées entre 0 et 500. D'autre part, plus le nombre de topics augmente, plus les topics se précisent. Ainsi, le nombre de topic optimal dépend principalement du niveau de précision de thème que cherche à obtenir l'utilisateur de la base de données. Celui-ci dépend aussi du nombre de résultats moyen qu'il désire obtenir dans chaque cluster et donc du nombre d'abstracts moyens générés par une recherche lorsqu'il emploie un modèle K-Means.

Conception de l'outil de recherche documentaire à partir du LDA

La première étape consiste à entrainer un modèle LDA sur les 50 000 abstracts qui ont été retenus pour la phase de test. La performance optimale du LDA est obtenue en entrainant celui-ci sur la matrice de fréquence des mots du corpus et non sur la matrice TF-IDF comme c'est le cas pour l'algorithme K-Means. Imaginons pour la démonstration que la matrice de fréquence compte 17 000 mots uniques.

Une fois le modèle entrainé, nous obtenons deux matrices, soit une matrice A de dimensions 1500 x 17 000, contenant les topics en rangées et les mots uniques du corpus en colonnes, et une matrice B de dimensions 50 000 x 1500 contenant les abstracts en rangées et les topics en colonnes. Pour rappel, la matrice A représente l'apport (la pondération) de chaque mot à un topic donné. Cette matrice est ensuite normalisée pour que la somme des rangées donne 1. La matrice B pour sa part représente la pondération de chaque topic dans un abstract donné et est également normalisée de sorte que chaque rangée donne 1.

L'étape qui suit consiste à prendre le texte de recherche entré par l'utilisateur et à appliquer le même prétraitement sur celui-ci que celui effectué sur la base de données et présenté à la section précédente. Ce prétraitement retourne donc un vecteur de mots nettoyés et ayant subi une stemmatisation. Il faut alors identifier à quelle colonne de la matrice A correspond chacun de ces mots, et récupérer ces colonnes. On calcule ensuite la somme de ces vecteurs pour obtenir un seul vecteur de dimensions 1500 x 1, et on normalise ce vecteur pour qu'il somme à 1. Ce vecteur représente ainsi la pondération moyenne du groupe de mots entré par l'utilisateur pour chacun des 1500 topics. Appelons ce vecteur, « vecteur-recherche ». Notons ici que le vecteur-recherche est de mêmes dimensions que les rangées de la matrice B. L'idée est alors de trouver quels sont les abstracts qui sont les plus similaires au vecteur-recherche. La proximité est ici établie à partir de la distance euclidienne entre tous les vecteurs rangée de la matrice B et le vecteur-recherche. Après avoir calculé cette distance pour les 50 000 abstracts, il est possible de les ordonner en ordre croissant de proximité avec le vecteur-recherche. En retenant les abstracts correspondants aux n premières distances de cette liste, on obtient alors les n abstracts ayant les pondérations de chaque topic qui sont les plus rapprochées de celle du vecteur-recherche. Si le modèle est bien optimisé, on peut s'attendre à ce que ces articles soient ceux qui partagent la sémantique la plus proche de celle du groupe de mots entré par l'utilisateur.

Conception de l'outil de recherche avec K-Means

L'objectif de l'outil de recherche basée sur le modèle K-Means est de retourner un ensemble de documents appartenant au même groupe et représentant le mieux possible la thématique correspondant à la recherche. À l'aide des 50 000 abstracts sélectionnés aléatoirement, nous avons ajusté, à l'aide de la matrice TF-IDF, un modèle K-Means comportant 1500 clusters. Cela nous permet d'obtenir en moyenne une trentaine

d'articles par groupe. L'étape suivante consiste à entrer une phrase ou un groupe de mots-clés correspondant à une recherche spécifique. Cet ensemble de mots subit le même prétraitement que les articles (retrait des caractères spéciaux et des nombres, tockenisation, stemmatisation, etc.). L'idée est donc de le traiter comme un nouvel abstract et d'utiliser le modèle K-Means pour prédire à quel cluster il devrait appartenir. Le résultat de la recherche correspond à l'ensemble des abstracts se trouvant à l'intérieur du cluster d'appartenance des mots de recherche. Évidemment, il est possible d'ajuster l'hyperparamètre du nombre de clusters afin d'obtenir une quantité différente de résultats.

Les abstracts à l'intérieur du cluster peuvent ensuite être classés de plusieurs manières, telles que par occurrence des termes utilisés dans la recherche, par date de parution ou encore par ordre alphabétique. Pour classer les abstracts selon l'occurrence des termes, nous effectuons la procédure suivante :

Pour chaque abstract du cluster, on somme les valeurs TF-IDF correspondant aux mots utilisés dans la recherche. La valeur trouvée correspond en quelque sorte à un score d'occurrence des termes pour chaque abstract. Les documents sont ensuite classés en ordre décroissant à l'aide de ce score.

L'intuition derrière cette méthode est ainsi de favoriser les articles dont les mots sont similaires à ceux utilisés dans la recherche. Évidemment, le classement par occurrence n'offre aucune garantie de similarité sémantique ou de pertinence par rapport à la recherche. Le premier résultat n'est donc pas nécessairement le plus pertinent même si plusieurs mots sont similaires à ceux utilisés dans la recherche. Inversement, le dernier résultat qui contient peu (ou pas) de mots relatifs à la recherche peut être plus pertinent que plusieurs autres résultats classés plus haut en termes d'occurrence.

Résultats

Difficultés d'évaluation

Les méthodes non supervisées représentent un défi au niveau de l'évaluation de la performance. Ce qui rend d'autant plus complexe l'évaluation de la performance dans notre cas est que d'une part le nombre d'abstracts est très élevé et d'autre part, que chacun de ceux-ci renferme un contenu relativement technique dont le thème n'est pas simple à interpréter. Étant donné que l'objectif de notre outil est de retourner des abstracts partageant le même thème qu'un groupe de mots de recherche, la seule manière de mesurer la performance du modèle est de lire et d'analyser manuellement les articles retournés par l'outil pour une recherche donnée. En effet, les critères de performances quantitatifs ne peuvent que très difficilement être appliqués dans ce contexte. Une mesure de performance qui serait par exemple basée sur le nombre des mots de recherche qui sont présents dans les abstracts retournés en viendrait à éliminer la notion de thème. En effet, le modèle LDA pourrait retourner un article dont la thématique s'apparente beaucoup à celle de la recherche, mais qui ne partage pas précisément les mots de cette recherche. Dans pareil cas, le critère de performance mentionné plus haut pénaliserait à tort un tel résultat.

Ainsi, considérant que notre outil cherche à exploiter la notion de proximité sémantique, et considérant que la sémantique est quelque chose de très difficilement quantifiable, la seule méthode envisageable pour l'évaluation de la performance de nos outils est l'évaluation qualitative manuelle par un lecteur. Ceci a notamment pour conséquence de rendre beaucoup plus difficile l'optimisation des hyperparamètres des modèles. Pour cette raison, le seul hyperparamètre que nous avons cherché à optimiser dans ce projet est celui correspondant au nombre de topic dans le LDA ou au nombre de cluster dans le K-Means.

Analyse qualitative du modèle LDA

Afin d'évaluer la capacité du modèle LDA à générer des résultats de recherches cohérents, nous avons effectué une analyse qualitative pour comprendre son comportement en fonction du nombre de topics qu'il

contient. Cette partie analyse donc la cohérence et l'interopérabilité des différents topics, ainsi que la capacité, pour le modèle LDA, de classer correctement les abstracts selon ses topics respectifs. Le tableau 1 présente les mots représentant le mieux leurs topics pour deux modèles LDA pour lesquels le nombre de topics a été fixé à 5 et 48 respectivement. Pour rappel, l'importance d'un mot dans un topic est déterminée comme étant sa probabilité à appartenir à un topic. Pour un topic donné le mot avec la plus grande probabilité d'appartenance sera considéré comme le plus important.

Tableau 1 : Meilleurs mots par topic

LDA avec 5 topics			
Topic 0	Topic 1	Topic 2	Topic 3
patient	covid	patient	virus
group	health	covid	detect
use	use	diseas	use
studi	pandem	infect	infect
result	Studi	sever	sampl
method	Model	case	test
LDA avec 48 topics			
Topic 0	Topic 1	Topic 28	Topic 31
respiratori	model	pressur	children
infect	use	ml	women
virus	predict	flow	age
viral	estim	tube	pediatr
pneumonia	differ	fluid	infant
tract	result	cm	adult

En comparant les principaux mots composant les topics d'un modèle LDA à 5 topics à ceux d'un LDA à 48 topics, on remarque qu'avec 5 topics seulement le modèle affiche des thèmes assez généraux. Avec 48 topics, les thèmes semblent beaucoup plus raffinés et précis. On peut aussi observer que l'ensemble des topics semblent bien regrouper des thématiques uniques contenant des mots différents les uns des autres. La cohérence entre les mots d'un même topic étant plus difficile à analyser quantitativement, il est toutefois possible de l'observer à l'aide du tableau 1. Prenons l'exemple du topic 28, où les principaux mots sont « pressure, ml, flow, tube ». Ces derniers sont donc cohérents entre eux et définissent bien des termes reliés à des « expériences en laboratoire ». Il est évidemment possible d'observer cette cohérence dans les différents topics.

Il est également pertinent de vérifier la capacité du modèle LDA à bien classer les abstracts selon ses topics respectifs. Voici un extrait d'abstract ayant été pondéré majoritairement dans le topic 1 et ce, pour les deux modèles LDA du tableau 1.

"The main goal of this paper is to develop the forward and inverse modeling of the Coronavirus (COVID-19) pandemic using novel computational methodologies in order to accurately estimate and predict the pandemic."

Bien que les mots décrivant le topic 1 sont plus précis dans le LDA avec 48 topics, l'abstract semble être classé de la bonne manière avec les deux modèles. L'expérience a été répétée avec différents abstract et les deux modèles semblent constamment effectuer une bonne pondération des topics. Évidemment, plus

le modèle contient de topics, plus la quantité de topics auxquels un abstract peut appartenir augmente. Les proportions deviennent alors plus difficiles à interpréter avec exactitude, mais ces dernières semblent cohérentes avec les différents sujets abordés par un abstract.

Recherche avec l'approche LDA

L'avantage de la méthode LDA est qu'à la différence d'une approche par regroupement, celle-ci exploite davantage la notion de topic dans son classement des articles. Cela signifie qu'un abstract appartiendra en général à plusieurs topics. Ainsi, un modèle LDA permet assez naturellement d'associer une combinaison de topics résultats d'un ensemble de mots de recherche, et donc par la suite de trouver à quels articles ceux-ci s'apparentent le plus sur une base sémantique.

Nous avons malheureusement constaté qu'en dépit d'un support théorique approprié à notre approche, celle-ci n'a pas de donné les résultats escomptés. En effet, les performances de ce modèle se sont avérées décevantes. Les articles suggérés par l'outil avaient souvent peu de lien sémantique avec les mots de recherche entrés par l'utilisateur.

Considérant la robustesse théorique de cette approche, il serait très intéressant de consacrer davantage de temps à l'optimisation des hyperparamètres tels que le alpha et le bêta afin de voir si la performance décevante obtenue est causée par une paramétrisation sous-optimale. Nous demeurons confiants que les fondements de cette approche sont cohérents et que des performances satisfaisantes voir intéressantes pourraient être obtenues moyennant des efforts de modélisation supplémentaires. En raison du constat qui vient d'être présenté, nos efforts ont par la suite davantage été orientés vers l'approche K-Means présentée dans la prochaine sous-section.

Recherche avec l'approche K-Means

Bien que LDA présente un bon potentiel, le modèle K-Means a néanmoins produit les résultats les plus adaptés aux besoins de notre outil. Premièrement, en utilisant les documents présents dans un cluster particulier comme résultats d'une recherche, la thématique semble être bien représentée par les articles et ces derniers sont cohérents avec la recherche effectuée, et ce, pour les différents cas testés. Nous avons effectué différentes recherches afin de vérifier la robustesse de l'outil. Voici quelques exemples de recherches testées:

- How to use machine learning algorithms to help with COVID-19?
- What is the best treatment or vaccine against COVID-19?
- How to prevent the transmission of COVID-19?
- Risks factors, severity and mortality of COVID-19
- Decision making in a pandemic

Pour chacune des recherches effectuées, nous obtenons d'abord le cluster d'appartenance des mots de recherche. Ensuite, nous ordonnons les abstracts contenus dans ce cluster à partir du critère d'occurrence des termes de recherche.

Voici une partie des résultats correspondant à la recherche « **Decision making in a pandemic** ».

Résultat	**Abstract**
1/38	The author offers insights on the importance of critical thinking skills in the midst of the coronavirus disease 2019 (COVID-19) **pandemic** Topics discussed include the clinical **decision making** responsibilities of individuals in medical professions with regard to the **pandemic**, the challenges brought by the influx of misinformation about COVID-19 to healthcare providers, and the legal and ethical consequences of misleading opinions about the **pandemic**

2/38	The COVID-19 **pandemic** and its sequelae have created scenarios of scarce medical resources, leading to the prospect that health care systems have faced or will face difficult **decisions** about triage, allocation, and reallocation These **decisions** should be guided by ethical principles and values, should not be made before crisis standards have been declared by authorities, and, in most cases, will not be made by bedside clinicians Do not attempt resuscitation and withholding and withdrawing **decisions** should be made according to standard determination of medical appropriateness and futility, but there are unique considerations during a **pandemic** Transparent and clear communication is crucial, coupled with dedication to provide the best possible care to patients, including palliative care As medical knowledge about COVID-19 grows, more will be known about prognostic factors that can guide these difficult **decisions**.
…	…
38/38	COVID-19-related controversies concerning the allocation of scarce resources, travel restrictions, and physical distancing norms each raise a foundational question: How should authority, and thus responsibility, over healthcare and public health law and policy be allocated? Each controversy raises principles that support claims by traditional wielders of authority in "federal" countries, like federal and state governments, and less traditional entities, like cities and sub-state nations. No existing principle divides "healthcare and public law and policy" into units that can be allocated in intuitively compelling ways. This leads to puzzles concerning (a) the principles for justifiably allocating "powers" in these domains and (b) whether and how they change during "emergencies." This work motivates the puzzles, explains why resolving them should be part of long-term responses to COVID-19, and outlines some initial COVID-19-related findings that shed light on justifiable authority allocation, emergencies, emergency powers, and the relationships between them.

Les résultats démontrent la pertinence entre la thématique de la recherche et celle des abstracts retournés, et ce, pour chacun des 3 abstracts. Bien que le dernier abstract représente celui dont l'occurrence des termes de recherche est la moins grande, il traite de l'allocation et de la gestion des ressources par les autorités en contexte de COVID-19, ce qui est étroitement relié au thème de la prise de décision pendant une pandémie. Il est intéressant de constater l'absence des mots « decision », « making » et « pandemic » dans le dernier article. En effet, nous avons ici un exemple d'un résultat qui tout en étant pertinent en termes de thème, ne contient aucun mot clé employé dans la recherche. Un tel abstract ne serait notamment pas apparu dans les résultats d'un moteur de recherche documentaire conventionnel ce qui démontre la valeur ajoutée de notre approche.

Système de recommandation basé sur le regroupement

Tel que mentionné plus tôt, la quantité de nouveaux articles augmente considérablement de jour en jour, et il est parfois difficile de rester informé et de se tenir à jour quant aux dernières percées scientifiques. En plus de permettre la recherche d'abstracts en fonction d'une thématique particulière, notre outil prend en considération cette problématique en vue d'y apporter une solution. Puisqu'il est impensable de lire l'ensemble des nouveaux documents à chaque mise à jour de la base de données, une sélection des articles pertinents doit être effectuée afin de cibler ceux correspondant aux champs d'intérêt de l'utilisateur. À l'aide du modèle de recherche K-Means, les clusters correspondant aux intérêts de l'utilisateur peuvent être enregistrés dans une liste de « préférences ». À chaque mise à jour de la base de données, plutôt que de réentraîner un modèle sur l'ensemble des documents, les nouveaux articles se voient attribués une classe à l'aide du modèle déjà entraîné et ce, après avoir subi le prétraitement commun aux autres articles. Si un

ou plusieurs abstracts sont classés dans une des classes identifiées dans la liste de préférence, ils sont alors automatiquement recommandés à l'utilisateur. Cette liste peut être manuellement éditée et il est possible d'ajouter ou de retirer les clusters correspondant à des ensembles d'articles d'intérêt.

Conclusion

Ce projet avait pour objectif de tester le potentiel des algorithmes LDA et K-means dans le développement d'un outil de recommandation d'ouvrage scientifique sur une base sémantique à partir d'une recherche par texte ou par groupe de mots. Deux outils ont été développés et testés. L'un basé sur un algorithme LDA (*Latent Dirichlet Allocation*) et l'autre basé sur un algorithme K-means. Notre analyse a démontré que malgré le potentiel offert par l'algorithme LDA en *topic modeling*, nous ne sommes pas parvenus pour le moment à obtenir un outil performant à partir ce celui-ci. Davantage de travail devrait être investi pour identifier les problèmes de modélisation qui peuvent expliquer cette performance. En contrepartie, le modèle K-means s'est avéré un outil pertinent et relativement performant pour relever le défi que nous nous étions donné. En effet, celui-ci parvient, à partir d'un texte rédigé par un utilisateur, à retourner un groupe d'articles dont le thème s'apparente à celui exprimé par les mots ou les phrases de l'utilisateur. Le nombre d'articles retourné de même que la précision du thème de ces articles est fonction du ratio entre le nombre de clusters employés pour entrainer le modèle et le nombre d'articles sélectionnés. Plus ce ratio est élevé, plus les clusters regroupent un plus petit nombre d'articles ayant des thèmes spécifiques.

La principale difficulté rencontrée au cours de ce projet a été le choix et l'optimisation des hyperparamètres des modèles. Nos deux approches de modélisation contiennent des hyperparamètres à la fois au niveau de l'algorithme de base de même qu'au niveau de l'implémentation de celui-ci. Plusieurs combinaisons de ces paramètres sont donc possibles, mais les résultats qui en découlent sont très difficiles à comparer. À chaque modification, une analyse manuelle des résultats s'impose et nous oblige à passer en revue les abstracts retournés pour tenter d'évaluer leur pertinence par rapport aux mots de recherche. Comment savoir dans ce contexte si une modification d'un hyperparamètre a conduit à une amélioration ou à une détérioration générale de la performance de l'outil? Cette approche manuelle laborieuse et peu précise limite considérablement la capacité à développer et améliorer les performances de notre outil. Pour parvenir à faire un tel travail, il faudrait au préalable prendre une sélection d'abstracts et leur attribuer un thème de manière à pouvoir ensuite adopter une approche de type supervisé. Bien qu'attribuer un thème à un groupe d'article demeure quelque chose de laborieux et délicat, cette alternative serait probablement plus efficace pour l'optimisation des hyperparamètres.

Quoi qu'il en soit, l'outil développé dans ce projet offre des performances tout à fait raisonnables et est par le fait même bien employable. Un peu de travail supplémentaire et l'exploration d'autres approches apprises au cours de ce projet permettraient sans doute d'accroitre ses performances et son utilisabilité, mais l'outil actuel offre selon nous une bonne base pour l'élaboration d'un outil a fort potentiel en recherche et en classification documentaire sur une base sémantique.

Bibliographie

Blei, & al. (2003). Latent Dirichlet Allocation. Journal Of Machine Learning Research. *Journal Of Machine Learning Research, 3(4/5)*, 993-1022.

Buckley, C., & Lewitt. (1985). Optimization of inverted vector searches. *Proceedings of the ACM Special Interest Group on Information Retrieval Conference*, (pp. 97-110). Montreal.

Carter, & al. (2016). Reading the high court at a distance: Topic modelling the legal subject matter and judicial activity of the high court of Australia. *University Of New South Wales Law Journal, 39(4)*, 1903-2015.

Deerwester, & al. (1990). Indexing by Latent Semantic Analysis. Journal Of The American Society For Information Science. *Journal Of The American Society For Information Science, 41*, 391-407.

Jardin, N., & Van Risjbergen, C. (1971). *The use of hierarchical clustering in information retrieval.* Int. J. Inform. Storage Retrieval.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Singh, V. K., Tiwari, N., & Garg, S. (2011). Document Clustering Using K-Means, Heuristic K-Means and Fuzzy C-Means. *International Conference on Computational Intelligence and Communication Networks.* Gwalior.

Suominen, & Toivanen. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal Of The Association For Information Science & Technology, 67(10)*, 2464-2476. doi:10.1002/asi.235

Székely, & Brocke, v. (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling tech. *Plos ONE, 12(4)*, 1-27. doi:10.1371/journal.pone.0174807

Yau, & al. (2014). Clustering scientific documents with topic modeling. *Scientometrics* , 100,767.

Zheng, & al. (2006). Identifying biological concepts from a protein-related corpus with a probabilistic model. BMC Bioinformatics. *BMC Bioinformatics*, 58. doi:10.1186/1471-2105-7-58