Traitement Automatique des Langues

Rapport du projet

« Étude d'outils de détection de topics : TF-IDF et LDA»

Zhongjie LI

Zhiya MA

Perrine QUENNEHEN

Extraction d'informations

dispensé par Madame Monnin

Année universitaire 2023-2024

Master 1 / Semestre 2

Sommaire

Introduction

Le traitement automatique du langage naturel (NLP) est un domaine qui traite de l'interaction entre les ordinateurs et les humains à travers le langage naturel. L'une des applications pratiques les plus importantes du NLP est le "topic modeling" ou modélisation de sujets, une technique statistique qui permet d'extraire les thèmes principaux d'un grand corpus de textes. Cette technique est essentielle pour comprendre et organiser de grandes quantités de données textuelles, et elle trouve des applications dans divers domaines tels que le résumé automatique de textes, la recommandation d'articles, la surveillance des médias et bien d'autres.

La modélisation de sujets a évolué au fil du temps avec le développement de divers modèles algorithmiques, chacun ayant ses propres mécanismes et utilisations optimales. Parmi ces modèles, le TF-IDF (Term Frequency-Inverse Document Frequency) et le LDA (Latent Dirichlet Allocation) sont particulièrement notables. Bien que ces deux techniques servent à analyser des textes, elles se distinguent par leurs approches et leurs objectifs.

Le TF-IDF est une mesure statistique qui évalue l'importance d'un mot dans un document par rapport à une collection de documents ou un corpus. Le poids TF-IDF augmente proportionnellement au nombre de fois qu'un mot apparaît dans le document, mais est compensé par le nombre de documents qui contiennent le mot dans le corpus, ce qui aide à ajuster le fait que certains mots apparaissent plus fréquemment en général. Cette méthode est souvent utilisée pour la sélection de caractéristiques et le prétraitement des données en vue de modélisations plus complexes.

De son côté, le LDA est un modèle probabiliste génératif qui permet de découvrir les sujets cachés dans un ensemble de documents. Il attribue des distributions de probabilités à la fois aux mots et aux documents pour différents sujets, ce qui permet d'inférer la structure des sujets dans les données. Contrairement au TF-IDF, qui se concentre sur la fréquence des mots, le LDA cherche à comprendre le contexte et les motifs sous-jacents qui caractérisent un corpus.

Ces différences fondamentales soulèvent la question centrale de ce projet : en quoi le modèle du LDA se distingue-t-il de celui du TF-IDF dans le cadre de la détection de mots-clés ? Pour y répondre, nous explorerons les principes de fonctionnement de chaque modèle, examinerons leurs performances et discuterons de leurs applications pratiques et de leurs limites respectives.

États de l’art : Modèles TF-IDF et LDA

Introduction

Dans le cadre de notre projet, nous avons choisi de nous concentrer sur deux modèles couramment utilisés dans le domaine de l'extraction de sujets : le TF-IDF et le LDA. Cette décision est motivée par les objectifs spécifiques de notre étude et les caractéristiques distinctes de ces modèles. Le TF-IDF (Term Frequency-Inverse Document Frequency) est une méthode statistique simple et efficace pour évaluer l'importance des mots dans un document par rapport à un corpus, tandis que le LDA (Latent Dirichlet Allocation) est un modèle probabiliste génératif conçu pour découvrir les structures thématiques sous-jacentes dans un ensemble de documents.

Méthode : Extraction de topics à l’aide du LDA

L'article "Latent Dirichlet Allocation" d'Alberto Bietti (2012) présente en détail le modèle LDA. Ce modèle part du principe que chaque document est un mélange de plusieurs topics et que chaque topic est une distribution de mots. Voici les principales composantes du LDA :

- Document (d) : Une suite de mots servant d'unité d'analyse.
- Mot (w) : Unité basique au sein d'un document, représentée par un indice dans un vocabulaire fixe.
- Topic** : Un groupe de mots fréquemment cooccurrents dans les documents.
- Distribution de topics (θd) : Attribuée à chaque document, basée sur une distribution de Dirichlet.
- Distribution de mots (βk) : Attribuée à chaque topic, également basée sur une distribution de Dirichlet.

Le LDA est un modèle probabiliste génératif qui suppose que chaque document du corpus est un mélange de topics qui s'enchaînent, se combinent entre eux. Les topics sont liés, ou autrement dit, contribuent à la génération des mots observés dans un document.

Le processus de génération du modèle LDA commence par l'initialisation des paramètres α et η, qui déterminent les distributions de topics et de mots. Ensuite, une distribution de topics est échantillonnée pour chaque document, et pour chaque mot dans un document, un topic est sélectionné pour générer le mot observé. La pertinence des topics est évaluée en comparant les mots générés avec les mots observés.

Les méthodes principales pour l'estimation des paramètres incluent le Gibbs sampling et la méthode variationnelle. L'article présente plusieurs expériences, dont une analyse de 2246 articles de presse et 17000 articles scientifiques, démontrant la capacité du LDA à identifier des thèmes cohérents au sein des documents.

En vue de tester le modèle LDA, l'auteur a présenté quelques expériences faites par autrui sur différentes données. La première porte sur un corpus de 2246 articles de Associated Press. La deuxième met en pratique le modèle LDA à un corpus 8 fois plus large que la première dont 17000 articles scientifiques du magazine Science. Quant à la dernière, elle élargit la taille du corpus à 100000 articles Wikipedia.

Comme les liens fournis vers les expériences citées sont tous expirés, la présentation des résultats d'ici s'orientera selon ceux qui sont illustrés dans l'article.

Pour l'expérience des articles de presse, le résultat est évoqué sous forme de tableau dont chaque colonne représente un topic et les 20 mots de la même colonne sont classés par probabilité décroissante (figure 1). Nous pouvons constater que les mots d'un topic partagent tous au moins un trait sémantique commun tels que la politique (1ère colonne), le fait policier (3ème colonne), le marché financier (7ème colonne), etc. Ce phénomène réside également dans le résultat des articles scientifiques (figure 2) où la première colonne est axée sur la biologie génétique humaine, la deuxième sur les espèces, la troisième sur le micro-organisme et la quatrième sur les modèles informatiques de données.

L'expérience sur les articles Wikipedia est une extension du modèle LDA. Elle montre la possibilité de regrouper les documents selon topics, de faire écho aux autres topics et documents étroitement liés.

Méthode : Extraction de topics à l’aide du TF-IDF

Le TF-IDF est une méthode statistique utilisée pour évaluer l'importance d'un mot dans un document par rapport à l'ensemble du corpus. L'article "Extraction automatique de termes-clés : Comparaison de méthodes non supervisées" de MA Zhiya compare diverses méthodes d'extraction de mots-clés, y compris le TF-IDF.

- TF (Term Frequency) : Fréquence d'un mot dans un document.
- IDF (Inverse Document Frequency) : Mesure de l'importance d'un mot à travers l'ensemble des documents.

Le TF-IDF est calculé en multipliant la fréquence d'un mot dans un document par l'inverse de sa fréquence dans l'ensemble du corpus. Cette méthode permet de pondérer les mots en fonction de leur importance relative, en réduisant l'impact des termes courants.

L'extraction automatisée de mots-clés est cruciale dans le domaine du traitement automatique de la langue (TAL), facilitant la réduction et l'accès aux documents. Les chercheurs se concentrent sur des méthodes efficaces pour extraire automatiquement des mots-clés pertinents sans nécessiter une supervision humaine intensive. L'article se focalise particulièrement sur les méthodes non supervisées, permettant le traitement d'un large volume de documents sans annotation humaine. En examinant et comparant différentes approches, l'article évalue leur efficacité et leur pertinence dans divers contextes, enrichissant ainsi les connaissances sur l'extraction automatique de mots-clés et guidant le développement de nouvelles techniques.

Les expérimentations menées dans l'article portent sur deux collections de documents académiques, l'une provenant du journal Information Processing and Management (IPM) et l'autre du journal Information Retrieval Journal (IRJ). Les auteurs ont construit ces collections en interrogeant le Web of Sciences.

Pour évaluer les méthodes d'extraction de mots-clés, les auteurs comparent les termes-clés extraits automatiquement avec ceux associés manuellement par les auteurs des documents. Ils mesurent l'efficacité des méthodes d'extraction en utilisant le rappel, qui est défini comme le nombre de termes-clés générés identiques à ceux formulés par les auteurs, divisé par le nombre total de termes-clés générés automatiquement.

Les méthodes étudiées incluent TF-IDF, TextRank, SingleRank, TopicRank, Kcore et WKcore. Les expérimentations ont également varié les paramètres spécifiques à chaque méthode, tels que le nombre de mots dans la fenêtre de cooccurrences pour TextRank, SingleRank, Kcore et WKcore, et le seuil de similarité pour TopicRank.

Les résultats de l'article montrent que la méthode TF-IDF produit les termes-clés les plus similaires à ceux sélectionnés manuellement par les auteurs. Cela peut être attribué au fait que TF-IDF utilise l'ensemble du corpus de documents pour pondérer les termes. Cependant, parmi les méthodes basées sur les graphes, SingleRank se démarque en fournissant des performances proches de celles de TF-IDF. Cette méthode, bien qu'utilisant uniquement le document analysé plutôt que l'ensemble du corpus, parvient à extraire des termes-clés pertinents.

En ajustant les paramètres des méthodes, des variations significatives dans les résultats sont observées. Par exemple, en variant la fenêtre de cooccurrence pour TextRank, SingleRank, Kcore et WKcore, il est constaté que SingleRank maintient des performances relativement stables, tandis que Kcore et WKcore montrent des fluctuations dans leurs performances en fonction de la taille de la fenêtre de cooccurrence. De plus, en modifiant le seuil de similarité pour TopicRank, les résultats varient considérablement selon la stratégie de regroupement et de sélection des termes-candidats.

Comparaison entre TF-IDF et LDA

Le TF-IDF et le LDA diffèrent fondamentalement dans leur approche de l'extraction de sujets :

- TF-IDF : Évalue l'importance d'un mot dans un document en se basant sur sa fréquence relative. C'est une méthode directe et efficace pour extraire des mots-clés, idéale pour des analyses de documents individuels. TF-IDF permet d'isoler rapidement les termes les plus significatifs d'un document donné, en tenant compte de leur rareté dans l'ensemble du corpus.
- LDA : Modélise les documents comme des mélanges de topics et découvre les structures thématiques sous-jacentes. C'est une approche plus complexe mais puissante pour analyser de grands ensembles de données textuelles et identifier des thèmes récurrents. Le LDA excelle dans la capture des relations contextuelles entre les mots et dans l'identification des structures thématiques qui peuvent ne pas être immédiatement apparentes.

Pour notre projet, la méthode TF-IDF est privilégiée pour sa simplicité et son efficacité dans l'extraction directe de mots-clés. Elle permet de mettre en lumière les termes les plus significatifs avec une approche relativement simple et rapide à mettre en œuvre. Cependant, l'utilisation du LDA pourrait être envisagée pour des analyses complémentaires lorsque la compréhension des structures de topics est nécessaire. Le LDA pourrait offrir des insights supplémentaires en révélant des patterns thématiques plus complexes et en fournissant une vision plus profonde des relations entre les mots au sein du corpus.

Conclusion

Les états de l'art examinés fournissent une compréhension approfondie des forces et des limites de chaque méthode. Le TF-IDF, par sa simplicité et son efficacité, semble s'aligner avec notre objectif de détection rapide et précise des mots-clés. En offrant une méthode directe pour évaluer l'importance des mots dans des documents spécifiques, le TF-IDF permet de répondre efficacement à des besoins d'extraction de mots-clés dans un cadre temporel restreint.

Le LDA, en offrant une analyse thématique plus profonde, peut servir à enrichir notre compréhension des documents dans des contextes plus larges. Son aptitude à modéliser les documents comme des mélanges de topics et à découvrir les structures thématiques sous-jacentes le rend particulièrement adapté aux analyses où la compréhension des relations contextuelles et des patterns thématiques est cruciale.

Pour répondre à notre problématique, nous envisagerons d'abord l'application de la méthode TF-IDF pour son efficacité immédiate dans l'extraction de mots-clés pertinents. Ensuite, en fonction des besoins spécifiques de notre projet et des insights obtenus, nous pourrions intégrer le LDA pour des analyses thématiques complémentaires, permettant ainsi une compréhension plus complète et nuancée des documents analysés.

En comparant les différents résultats que nous obtiendrons avec ces deux modèles, nous pourrons valider notre choix méthodologique et ajuster notre approche pour maximiser l'efficacité et la pertinence de l'extraction de sujets et de mots-clés dans notre projet. Cette double approche nous permettra d'exploiter les avantages de chaque méthode et d'optimiser notre analyse en fonction des caractéristiques spécifiques de notre corpus de documents.

I. Mécanisme des modèles

A. Le modèle TF-IDF (Term Frequency-Inverse Document Frequency)

Le modèle TF-IDF est une méthode de pondération très utilisée dans le traitement de texte pour l'extraction de caractéristiques et la mesure de l'importance d'un terme dans un document par rapport à un corpus. Il combine deux mesures : la fréquence des termes (*TF*) et la fréquence inverse des documents (*IDF*).

Mécanisme:

- **Term Frequency (*TF*)**
  - **Définition** : Le *TF* mesure la fréquence d'un terme dans un document spécifique.
  - **Calcul** : $TF(t,\ d)\ =\ \frac{Nombre\ de\ fois\ que\ le\ terme\ t\ apparaît\ dans\ le\ document\ d}{Nombre\ total\ de\ termes\ dans\ le\ document\ d}$
  - **Exemple** : Si le mot "information" apparaît 3 fois dans un document de 100 mot, alors : $TF\ =\ \frac{3}{100}=\ 0.03$
- **Inverse Document Frequency (*IDF*)**
  - **Définition** : L'*IDF* mesure l'importance d'un terme en tenant compte de sa fréquence dans le corpus. Plus un terme est rare, plus son *IDF* est élevé.
  - **Calcul** : $IDF(t,\ D)\ =\ log(\frac{Nombre\ total\ de\ document\ dans\ le\ corpus\ D}{Nombre\ de\ documents\ contenant\ le\ terme\ t})$
  - Exemple : Si le mot "rare" apparaît dans 5 documents sur un corpus de 1000 documents, alors : $IDF\ =\ log(\frac{1000}{5})\ =\ log(200)$
- **TF-IDF**
  - **Définition** : La valeur de TD-IDF d'un terme est le produit de *TF* et d'*IDF*.
  - **Calcul** : $TFIDF(t,\ d,\ D)\ =\ TF(t,d)\ \times\ IDF(t,\ D)$
  - **Exemple** : si TF = 0.03 et IDF = 2.3, alors TFIDF = 0.03 x 2.3 = 0.069

Ce calcul permet de donner un poids à chaque terme de chaque document, en mettant en avant les termes qui sont à la fois fréquents dans un document particulier et rares dans le reste du corpus.

B. Le modèle LDA (Latent Dirichlet Allocation)

Le LDA est un modèle probabiliste génératif qui permet de découvrir les sujets présents dans un ensemble de documents. Il suppose que chaque document est une combinaison de plusieurs sujets et que chaque sujet est caractérisé par une distribution de mots.

Mécanisme:

1. **Initialisation**:
    - **Choix des sujets** : Déterminer le nombre de sujets *K* à extraire du corpus.
    - **Attribution initiale** : Assigner aléatoirement un sujet à chaque mot dans chaque document.
2. **Assignation de sujets**:
    - **Probabilité de sujet** : Pour chaque mot dans un document, recalculer la probabilité qu'il appartienne à un sujet donné. Ce calcul prend en compte la fréquence des mots dans les sujets et la répartition actuelle des sujets dans le document.
    - **Calcul** : Utilisation de la méthode d'échantillonnage de Gibbs Sampling pour ajuster l'attribution des mots aux sujets.
3. **Convergence**:
    - **Itération** : Répéter le processus d'assignation jusqu'à ce que la répartition des mots dans les sujets se stabilise.
    - **Objectif** : Maximiser la vraisemblance des données observées avec les distributions estimées, ce qui rend explicite la structure cachée des sujets dans un corpus textuel.

Le modèle LDA produit ainsi des distributions de probabilités pour chaque terme par sujet et chaque sujet par document, révélant les thèmes principaux du corpus.

II. Comparaison des modèles

Malgré le même but final des deux modèles, la détection de topics, ils le réalisent à travers deux voies qui s'écartent complètement.

Le modèle TF-IDF s'appuie principalement sur la fréquence des mots du corpus. Plus un mot se répète dans un document, plus il sera attribué une importance significative. Pourtant, cela ne garantit pas qu'un mot ayant une fréquence élevée sera toujours pris en compte comme topic. Par exemple, le pronom impersonnel "il", certains verbes courants comme "a", "est" sont normalement omniprésents dans les discours de tout genre, mais ces mots sont loin d'être topics pour l'impertinence. Le modèle TF-IDF, quant à lui, permet de contourner les mots trop fréquents dans tous les documents en utilisant l'inverse de la répartition du terme. Cette méthode assure également d'augmenter l'importance des mots moins courants.

Basé sur la fréquence pondérée des mots, le TF-IDF conclut les résultats de la détection de topics en analysant le corpus entier. Quant au modèle LDA, il porte aussi sur l'ensemble du corpus, mais elle abandonne la stratégie fréquentielle et adopte une méthodologie probabiliste.

Le LDA suppose que chaque document du corpus possède une distribution de topics, que chaque distribution est représentée par une distribution de topics et que chaque mot issu du corpus est généré par l'un des topics. Le modèle cherche à analyser le contenu de chaque document du corpus afin de trouver les unités sémantiques importantes. Pour cela, il réalise de nombreuses itérations sur l'ensemble du corpus en vue d'ajuster les paramètres, qui sont mis en place par la suite pour l'estimation de topics. L'ajustement des paramètres se réalise souvent avec le Markov chain Monte Carlo et Gibbs sampling.

En bref, le TF-IDF étudie l'importance des mots selon leurs fréquences pondérées dans tout le corpus, le LDA cherche plus profondément les connexions sémantiques en effectuant un grand nombre d'itérations sur le corpus, ce qui consomme donc beaucoup plus de temps que le TF-IDF.

III. Présentation des expériences

Les expériences sont réalisées sur deux corpus différents. Le premier est tiré du forum Stackoverflow, qui recueille les questions posées sur le forum ainsi que les étiquettes publiées au-dessous de chaque question. Les étiquettes servent plus tard de groupe référent pour l'évaluation des résultats obtenus avec les modèles. Dans ce corpus on a 20 000 lignes. L'autre corpus collectionne les phrases issues de Wikipédia accompagnées d'une liste de mots-clés pour chaque phrase, il est composé de 13 014 lignes.

Les deux corpus sont tous divisés en 3 sous-corpus, l'un pour l'entraînement (60% du corpus original), l'un comme corpus de tests (20% du corpus original) et le dernier sert de corpus de développement (20% du corpus original). Puis, un prétraitement sera amené sur les 3 sous-corpus. Cela consiste à lemmatiser les tokens, à retirer les balises de HTML, les caractères spéciaux, les caractères numéraux et les mots-vide. Pour le modèle TF-IDF, le prétraitement supprime les mots qui apparaissent dans moins de 20 documents du corpus. Cette étape est ignorée dans le prétraitement pour le modèle LDA, car elle sera effectuée directement lors de la modélisation en utilisant le paramètre "no_below" du LDA.

Une fois que le corpus est bien nettoyé, il sera mis en analyse pour trouver les bigrammes et les trigrammes, qui sont les termes hautement cooccurrents dans chaque document. Ces termes rejoindront et enrichiront les topics détectés un peu plus tard avec leurs informations mutuelles.

Ensuite, les modèles TF-IDF et LDA seront paramétrés. Ils excluent les mots trop fréquents qui apparaissent dans 85 % du corpus et prennent en compte au maximum 1 000 mots comme mots-clés. Pour le modèle LDA, il itérera 200 fois sur le corpus à analyser. Étant donné que LDA extrait des mots appartenant à des sujets et que ces sujets sont ensuite distribués en fonction de leur probabilité d'appartenance à un document, certains mots dans les sujets ne sont pas présents dans le document en question. Il a donc fallu faire une petite modification pour ne récupérer que les mots présents dans le document et ainsi obtenir la liste de

mots-clés. Les modèles paramétrés sont d'abord entraînés à l'aide du corpus d'entraînement avant de passer à l'estimation des mots-clés.

Les résultats de la détection de mots-clés sont présentés sous forme de tableaux, avec le texte du chaque document analysé, le groupe référent de mots-clés, le groupe estimé de topics avec les scores TF-IDF ou LDA, le nombre de mots-clés vrais positifs, le pourcentage de mots-clés vrais positifs par rapport au groupe référent, les scores de précision, de rappel et de F-mesure.

Finalement, les résultats du TF-IDF et du LDA seront comparés respectivement pour le corpus de Stackoverflow et de Wikipedia afin de calculer la différence du nombre de mots-clés vrais positifs. En ce qui concerne les scores de précision, de rappel et de F-mesure, bien que les résultats obtenus ne soient pas très élevés, nous avons quand même décidé.

Difficultés rencontrées

La première difficulté rencontrée a été de trouver au moins un corpus annoté avec des mots-clés, car la majorité de ceux que nous avons trouvés ont les phrases clés annotées mais pas les mots-clés. Finalement, nous en avons trouvé deux : un dont les données proviennent de Stackoverflow, où les tags mentionnés par les utilisateurs ont été utilisés comme mots-clés, et un second, dont les données proviennent de Wikipédia, mais pour lequel nous n'avons trouvé aucune information concernant les mots-clés annotés, notamment s'ils ont été annotés manuellement ou non.

Ensuite, l'une des difficultés rencontrées lors des expériences a été d'harmoniser le prétraitement et les paramètres pour les modèles TF-IDF et LDA. Comme le LDA exige le processus de lemmatisation, nous avons décidé de l'implémenter dans le prétraitement du TF-IDF. Par ailleurs, le TF-IDF ne permet pas de filtrer les mots apparaissant dans moins de 20 documents avec un simple paramètre, ce qui nous a poussés à concevoir une fonction pour le réaliser.

De plus, nous avons observé que les tags annotés contenaient souvent des bigrammes et des trigrammes, tandis que les modèles TF-IDF et LDA ne traitent que

les unigrammes. Pour pallier cela, nous avons développé deux fonctions distinctes : une pour extraire les bigrammes et trigrammes les plus pertinents dans le corpus en fonction de leur score PMI, et une autre pour générer toutes les combinaisons possibles à partir des unigrammes obtenus par TF-IDF et LDA. Ensuite, nous comparons ces combinaisons avec les bigrammes et trigrammes, et si une combinaison correspond, nous l'ajoutons à la liste des unigrammes. Cependant, cette approche ne nous permet pas d'obtenir les scores TF-IDF et LDA associés.

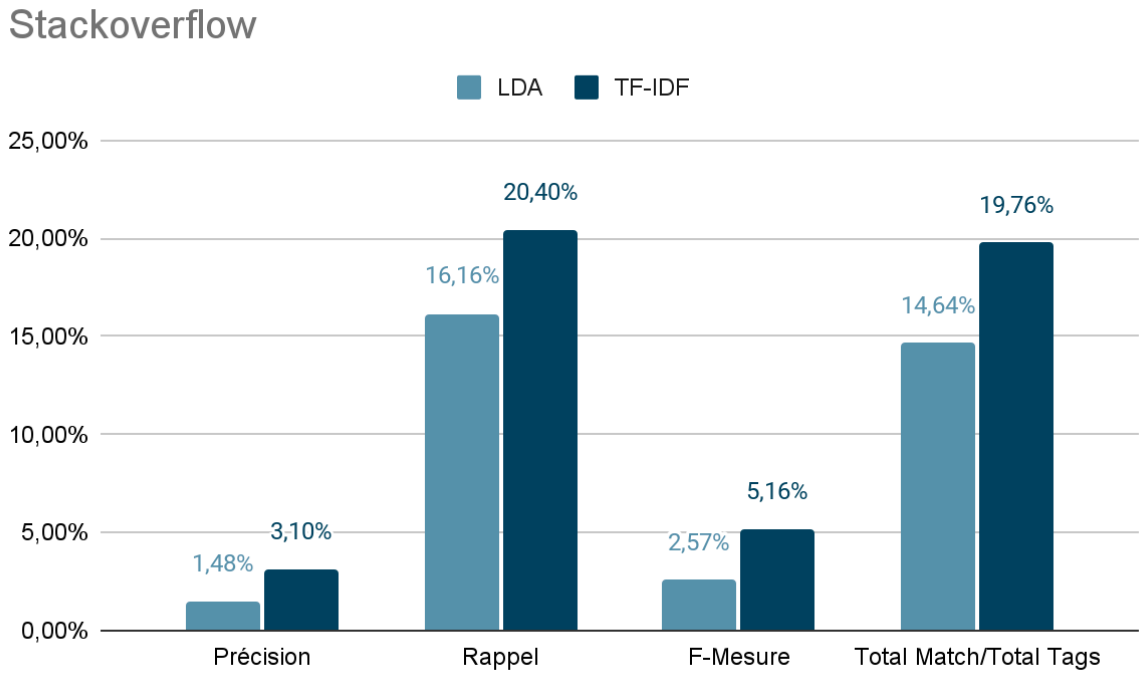
IV. Présentation des résultats

Les résultats obtenus à partir de l'analyse comparative entre les modèles LDA (Latent Dirichlet Allocation) et TF-IDF (Term Frequency-Inverse Document Frequency), sur le corpus de test, pour la génération de mots-clés sur les ensembles de données StackOverflow et Wikipedia fournissent des informations précieuses sur la performance et l'efficacité de chaque méthode.

Pour l'ensemble de données StackOverflow, TF-IDF se distingue par des performances supérieures à celles de LDA dans la plupart des mesures évaluatives.

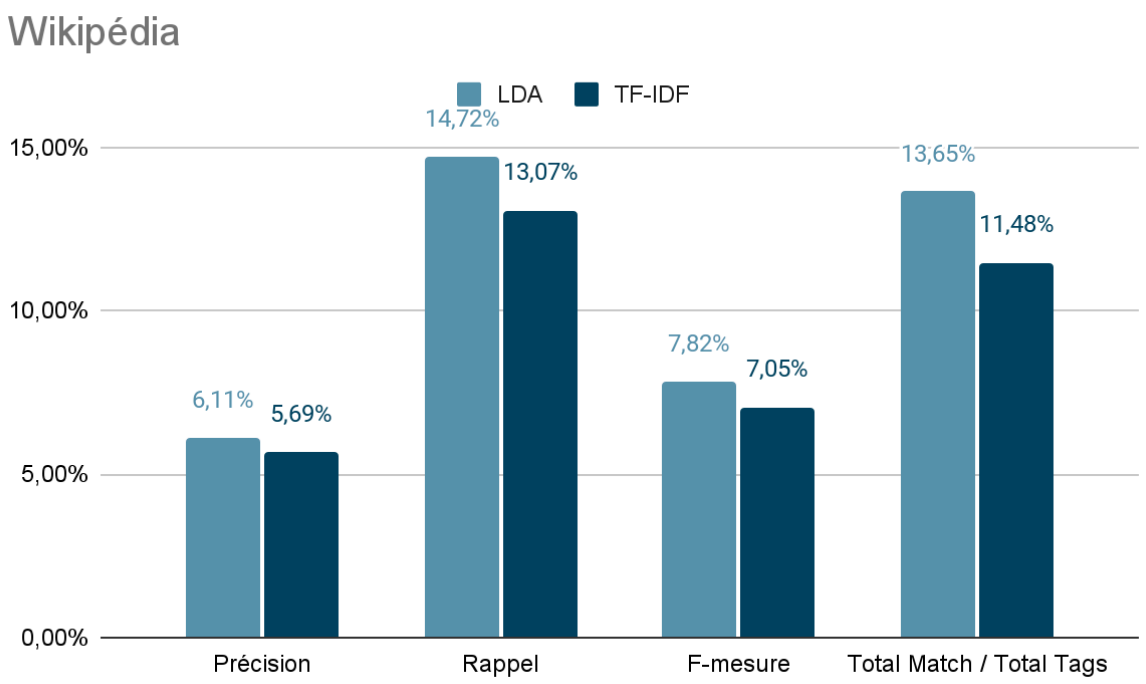
Scores pour StackOverflow (LDA vs TF-IDF) :

- **Précision** : LDA obtient une précision de 1,48 % tandis que TF-IDF atteint 3,1 %. Cela indique que, en moyenne, les mots-clés extraits par TF-IDF sont plus pertinents que ceux extraits par LDA.
- **Rappel** : Le rappel de LDA est de 16,16 % tandis que celui de TF-IDF est de 20,4 %. Cela signifie que TF-IDF réussit à extraire un pourcentage plus élevé de mots-clés pertinents présents dans les données.
- **F-mesure** : LDA obtient une F-mesure de 2,57 % tandis que TF-IDF atteint 5,16 %. Cela montre que TF-IDF produit une meilleure balance entre précision et rappel par rapport à LDA.



Scores pour Wikipedia (LDA vs TF-IDF) :

- **Précision** : LDA a une précision de 6,11 % tandis que TF-IDF obtient 5,69 %. Les deux méthodes ont des performances assez similaires en termes de précision.
- **Rappel** : Le rappel de LDA est de 14,72 % tandis que celui de TF-IDF est de 13,07 %. Encore une fois, les deux méthodes sont assez proches en termes de rappel.
- **F-mesure** : LDA obtient une F-mesure de 7,82 % tandis que TF-IDF obtient 7,05 %. Les deux méthodes sont à peu près équivalentes en termes de F-mesure.



Comparaison des ratios total Match/Total Tags de référence (LDA vs TF-IDF) :

Les ratios total Match/Total Tags de référence pour les modèles LDA et TF-IDF montrent la capacité de ces modèles à extraire des mots-clés pertinents parmi tous les mots-clés de référence présents dans les données.

Plus précisément, ces ratios indiquent quelle proportion des mots-clés extraits par chaque modèle correspond aux mots-clés de référence, c'est-à-dire aux mots-clés véritablement pertinents dans les données. Un ratio plus élevé signifie que le modèle a extrait un pourcentage plus important de mots-clés pertinents parmi tous les mots-clés de référence, ce qui est généralement considéré comme un indicateur de performance positif.

- Pour StackOverflow, le pourcentage de correspondance total entre les mots-clés extraits par les modèles et les mots-clés de référence est de 14,64 % pour LDA et de 19,76 % pour TF-IDF.
- Pour Wikipedia, ces pourcentages sont de 13,65 % pour LDA et de 11,48 % pour TF-IDF.

Ces résultats indiquent que TF-IDF tend à obtenir une meilleure correspondance avec les mots-clés de référence dans l'ensemble de données StackOverflow, suggérant ainsi une meilleure performance de TF-IDF dans l'extraction de mots-clés pertinents par rapport à LDA. Cependant, sur l'ensemble de données Wikipedia, LDA montre des performances légèrement supérieures en termes de précision et de rappel.

Dans l'ensemble, les résultats montrent que TF-IDF surpasse LDA en termes de précision, de rappel et de F-mesure sur l'ensemble de données StackOverflow. TF-IDF extrait des mots-clés plus pertinents et atteint un meilleur équilibre entre la précision et le rappel par rapport à LDA. Cela se reflète également dans le pourcentage de correspondance totale entre les mots-clés extraits et les mots-clés de référence, où TF-IDF montre une performance supérieure avec 19,76 % contre 14,64 % pour LDA.

En revanche, sur l'ensemble de données Wikipedia, les performances des deux modèles sont plus comparables. LDA montre une légère supériorité en termes de précision (6,11 % contre 5,69 %) et de rappel (14,72 % contre 13,07 %), ainsi qu'une meilleure F-mesure (7,82 % contre 7,05 %). Cependant, TF-IDF reste compétitif et ses résultats sont proches de ceux de LDA. En termes de correspondance totale des mots-clés de référence, LDA obtient un pourcentage de 13,65 % contre 11,48 % pour TF-IDF.

Ces résultats suggèrent que le choix entre LDA et TF-IDF peut dépendre de la nature de l'ensemble de données. Pour des données similaires à celles de StackOverflow, TF-IDF est recommandé pour une extraction de mots-clés plus précise et équilibrée. Cependant, pour des données comme celles de Wikipedia, LDA peut offrir une légère avantage en termes de précision et de rappel, bien que la différence soit marginale.

Explication de la précision, du rappel et de la F-mesure :

La précision, le rappel et la F-mesure sont des métriques couramment utilisées pour évaluer la performance des modèles de classification, de régression ou dans le cas présent, de génération de mots-clés.

- **Précision** : est calculée en divisant le nombre de mots-clés pertinents correctement extraits par le nombre total de mots-clés extraits. Une précision plus élevée indique que le modèle extrait moins de mots-clés incorrects.
- **Rappel** : est calculé en divisant le nombre de mots-clés pertinents correctement extraits par le nombre total de mots-clés pertinents dans les données. Un rappel plus élevé indique que le modèle extrait un plus grand nombre de mots-clés pertinents.
- **F-mesure** : La F-mesure est une mesure qui combine à la fois la précision et le rappel en une seule métrique. Elle est calculée comme la moyenne pondérée harmonique de la précision et du rappel. Une F-mesure plus élevée indique un meilleur équilibre entre précision et rappel. Elle est particulièrement utile lorsque les classes (dans ce cas, les mots-clés) sont déséquilibrées, ce qui signifie qu'il y a beaucoup plus de non-mots-clés que de mots-clés dans les données.

V. Perspectives d'améliorations

La première perspective d'amélioration serait d'améliorer l'intégration des n-grammes en affinant les techniques de leur extraction et en les intégrant dans le processus de génération de mots-clés. Cela permettrait d'obtenir des scores pour ces mots-clés prédits, afin de comparer ces scores et de vérifier si les mots-clés prédits corrects ont un score plus élevé que les autres mots-clés prédits.

On pourrait également expérimenter avec des approches hybrides combinant les avantages de TF-IDF et LDA, par exemple utiliser TF-IDF pour une première sélection de mots-clés, suivie de l'affinage avec LDA pour mieux capter les relations sémantiques.

Enfin, on pourrait intégrer l'utilisation de modèles plus récents comme BERT ou GPT, qui peuvent également prendre en compte le contexte dans lequel un mot est

employé, ainsi que les relations sémantiques, contrairement à TF-IDF, et ainsi améliorer la précision et la pertinence de la génération des mots-clés.

VI. Conclusion

Cette étude comparative entre les modèles LDA (Latent Dirichlet Allocation) et TF-IDF (Term Frequency-Inverse Document Frequency) pour la génération de mots-clés a permis d'analyser en profondeur les performances et les caractéristiques de chaque méthode dans deux contextes différents, celui des données de StackOverflow et celui des données de Wikipedia.

Les résultats obtenus montrent que le modèle TF-IDF présente généralement des performances supérieures à celles du modèle LDA en termes de précision, de rappel et de F-mesure pour la génération de mots-clés sur l'ensemble de données StackOverflow. TF-IDF extrait des mots-clés plus pertinents et atteint un meilleur équilibre entre précision et rappel par rapport à LDA. Cela se reflète également dans le pourcentage de correspondance total entre les mots-clés extraits et les mots-clés de référence, où TF-IDF montre une performance supérieure avec 19,76 % contre 14,64 % pour LDA.

En revanche, sur l'ensemble de données Wikipedia, les performances des deux modèles sont plus comparables. LDA montre une légère supériorité en termes de précision (6,11 % contre 5,69 %) et de rappel (14,72 % contre 13,07 %), ainsi qu'une meilleure F-mesure (7,82 % contre 7,05 %). Cependant, TF-IDF reste compétitif et ses résultats sont proches de ceux de LDA. En termes de correspondance totale des mots-clés de référence, LDA obtient un pourcentage de 13,65 % contre 11,48 % pour TF-IDF.

Ces résultats suggèrent que le choix entre LDA et TF-IDF peut dépendre de la nature de l'ensemble de données. Pour des données similaires à celles de

StackOverflow, TF-IDF est recommandé pour une extraction de mots-clés plus précise et équilibrée. Cependant, pour des données comme celles de Wikipedia, LDA peut offrir un léger avantage en termes de précision et de rappel, bien que la différence soit marginale.

Il est important de noter que le choix entre le TF-IDF et le LDA dépend du contexte spécifique de l'application et des objectifs de l'analyse. Dans certains cas, notamment lorsque l'on cherche à identifier les sujets principaux abordés dans un corpus de documents, le LDA peut être plus approprié en fournissant une vue d'ensemble des thèmes dominants. En revanche, si l'objectif est de mettre en évidence les mots-clés les plus saillants dans un texte individuel, le TF-IDF peut être une option plus efficace.

En conclusion, cette étude souligne l'importance de choisir le modèle de génération de mots-clés le plus approprié en fonction des exigences spécifiques de la tâche et du corpus de données. Les modèles TF-IDF et LDA offrent chacun des avantages uniques et complémentaires, et leur sélection dépendra des compromis entre précision, complexité et interprétabilité.

Bibliographie

Bietti, A. (2012, May). Latent dirichlet allocation. (Working Paper).

https://alberto.bietti.me/files/rapport-lda.pdf

Firoozeh, N., Nazarenko, A., Alizon, F., & Daille, B. (2019). Keyword extraction : Issues and methods. *Natural Language Engineering*, *26*(3), 259-291. https://doi.org/10.1017/s1351324919000457

Mothe, J., & Ramiandrisoa, F. (2016). Extraction automatique de termes-clés : Comparaison des méthodes non supervisées de la littérature. *Conférence En Recherche D'Informations et Applications - Rencontres Jeunes Chercheurs En Recherche D'Information (RJCRI CORIA 2016)*, 315-324. https://doi.org/10.24348/sdnri.2016.rjc7

Siddiqi, S., & Sharan, A. (2015). Keyword and Keyphrase Extraction Techniques : A Literature Review. *International Journal Of Computer Applications*, *109*(2), 18-23. https://doi.org/10.5120/19161-0607