Survey des Méthodes de Pré-traitement d'un Corpus

Thomas GIOVANNINI *Mouhamadou Mansour LO*
Karine ZHU *Assoumani Chissi HAMZA*

Avril 2021

Résumé

Aujourd'hui l'exploitation de données est une tâche qui préoccupe presque tous les domaines. La qualité de l'information à explorer dépend d'une première étape qui est le pré-traitement. L'appel aux techniques de la fouille de données textuelles s'avère plus que nécessaire dans la mesure où le volume d'information textuelles disponibles pour les scientifiques augmente continuellement. Dans le domaine de la recherche d'information (RI), de nombreuses techniques permettent aux scientifiques de rechercher et d'obtenir des informations appropriées à leurs recherches dans des corpus textuels. Ainsi l'indexation des documents permet d'appréhender la représentation de l'information et de le retrouver quand c'est nécessaire pour répondre à une requête utilisateur. L'objectif principal de cet article tourne au tour des méthodes et techniques de pré-traitements de données textuelles.

Mots-clés : pré-traitement, stop words, lemmatisation, stemming, TF-IDF, tokenisation

1 Introduction

Avec l'avènement d'internet et grâce au progrès de la science, la numérisation des documents autrefois uniquement disponible sous format papier est de plus en plus récurrente. Les données textuelles comme les données bibliographiques d'articles scientifiques, de livres, thèses ou autres peuvent généralement être stockées de manière assez brute dans des bases de données ou des datasets au vu d'un potentiel traitement informatique ou statistique comme du text mining. Le titre, les auteurs, le résumé du document, les descripteurs du sujet sont des exemples de données bibliographiques souvent présent dans tout articles ou document. Leurs utilisations au vu d'une analyse, visualisation de données ou extraction de connaissances nécessite au préalable un certain nombre de pré-traitements pour rendre ce travail intéressant et exploitable. Notre sujet porte sur le domaine de la recherche d'information (RI),

qui est le domaine qui étudie la manière de retrouver des informations dans un corpus qui est composé de documents d'une ou plusieurs bases de données et qui sont décrits par un contenu ou les méta-données associées. La recherche d'information ou "information retrieval" en anglais est composée de plusieurs phases dont la première est le pré-traitement. Cet article examine les différentes techniques de pré-traitement utilisées en recherche d'information, centrées sur les données textuelles. Pour cela nous avons décidé de tester l'efficacité de certaines méthodes que nous avons trouvé et de les comparer entre elles sur un jeu de données bibliographiques d'articles scientifiques. La suppression du bruit, l'élimination des stop words, la tokenisation, la lemmatisation et enfin l'application d'un TF-IDF nous ont permis de réduire de 99,3% le nombre de mots pour seulement garder ceux qui résument le mieux notre corpus. Enfin, nous avons vectorisé les mots restants pour visualiser les plus proches d'un mot choisi. Cet article est organisé comme suit. La section 2 traite du background. La section 3 donne l'état de l'art. L'étude et les résultats sont discutés dans la section 4. La conclusion est donnée dans la section 5.

2 Background

2.1 La Tokenisation

La Tokenisation qui est une opération de l'analyse lexicale de texte, est le processus qui consiste à briser un flux de texte en mots, phrases, symboles, ou d'autres éléments significatifs appelés jetons (token). Un jeton est souvent une suite de caractère alphabétiques ou alphanumériques. Le but de la Tokenisation est l'exploration des mots dans une phrase. La liste des jetons devient une entrée pour un traitement ultérieur tel que l'analyse ou l'extraction de texte. S. Kannan et V. Gurusamy [1] [P.3] soulignent quelques défis liés pendant cette phase. Ils ont montré les problèmes liés à la suppression des signes de ponctuation, d'autres caractères comme les parenthèses, les traits d'union, etc qui doivent également être traités. L'incohérence peut être due à des nombres et des formats horaires différents. Un autre problème est celui des abréviations et des acronymes qui doivent être transformés en une forme standard. L'algorithme de Tokenisation doit dépendre aussi du langage utilisé dans le texte. Les langues telles que l'anglais et le français sont délimitées par des espaces, car la plupart des mots sont séparés les uns des autres par des espaces blancs. Les langues telles que le chinois et le japonais sont dites non segmentées, car les mots n'ont pas de frontières claires. La Tokenisation des phrases de langues non segmentées nécessite des informations lexicales et morphologiques supplémentaires. Dans la pratique, on utilise nltk (Natural Language ToolKit) qui est la librairie la plus connue pour faire du traitement du langage naturel.

2.2 Le Stop Words Removal

Le Stop Word Removal est un processus permettant de supprimer les mots fréquents dans le texte non significatifs comme les pronoms, les prépositions, les conjonctions, etc. Cette étape est beaucoup plus importante puisqu'il permet de réduire la dimension du texte entre 20 à 30%. **S. Vijayarani, J. Ilamathi et Ms. Nithya [2] [p .10]** présentent dans leur extrait quatre méthodes de Stop Word Removal :

- La méthode classique qui est basée sur la suppression des mots obtenus à partie de listes pré-compilées.
- La méthode fondée sur la loi de Zipf (Z-methods) qui consiste à faire un plus sur la méthode classique en rajoutant trois règles : suppression des mots les plus fréquents (TF-Hight), suppression des mots singletons (TF1) et suppression des mots ayant une faible fréquence de document inverse (IDF).
- La méthode de l'information mutuelle (MI) qui est une méthode supervisée permettant de calculer l'information mutuelle entre un terme donné et une classe de documents. L'objectif de cette méthode est de supprimer les mots avec une faible information mutuelle sur le document.
- La méthode TBRS (Term Based Random Sampling) qui fonctionne en itérant sur des morceaux séparés de données qui sont sélectionnés aléatoirement. La référence de la dernière méthode sur l'article **[2]** donne plus de détails sur son mode de fonctionnement.

2.3 La Racinisation (Stemming) et la Lemmatisation

La Racinisation (Stemming) et la Lemmatisation sont deux notions très proches qui ont le même objectif de regrouper les différentes variantes d'un mot en utilisant des règles de transformation (Stemming) ou de l'analyse grammaticale (Lemmatisation).

2.3.1 La Racinisation

La Racinisation est un procédé de transformation des flexions en leur radical ou racine. Les techniques utilisées pour ce faire reposent généralement sur une liste d'affixes (suffixes, préfixes, postfixes, antéfixes) de la langue considérée et sur un ensemble de règles de Racinisation construites a priori qui permettent, étant donné un mot de trouver sa racine. Toutefois, il existe des inconvénients liés à cette méthode. Par exemple deux mots avec des racines différentes peut ramener la même radicale après Stemming (Over Stemming) et inversement deux mots dont la racine devrait être la même, mais qui ne le sont pas après Stemming (Under Stemming). Plusieurs méthodes existent permettant de faire la racinisation comme on peut les retrouver dans :

- Les méthodes de troncature qui consistent à supprimer les suffixes ou préfixes (qu'on appelle les affixes) d'un mot. L'algorithme le plus connu est l'algorithme de Porter.
- Les méthodes statistiques qui sont basées sur des analyses et des techniques statistiques exemple N-Gram stemmer.
- Les méthodes mixtes exemple Krovetz Stemmer (KSTEM).

2.3.2 La Lemmatisation

La Lemmatisation a pour objectif de retrouver le lemme d'un mot. Elle renvoie l'infinitif pour un verbe et sa forme au masculin singulier pour un nom, adjectif, article, etc. La lemmatisation fait généralement référence à l'utilisation correcte d'un vocabulaire et à l'analyse morphologique des mots, visant normalement à supprimer uniquement les terminaisons flexionnelles et à restituer la forme de base ou de dictionnaire d'un mot, qui est connue sous le nom de lemme [**4**].

2.4 Le TF-IDF (Term Frequency – Inverse Document Frequency)

Pour mesurer l'importance du terme dans un document en fonction d'une requête donnée, la méthode la plus répandue est le TF-IDF. La fréquence des termes de chaque mot dans un document (TF - term frequency) est un poids qui dépend de la distribution de chaque mot dans les documents. Elle exprime l'importance du mot dans le document. Cette mesure est utilisée pour l'amélioration du rappel. Cependant, la fréquence des deux termes seuls ne peut pas assurer une performance optimale, spécifiquement quand les termes à haute fréquence ne sont pas concentrés dans quelques documents particuliers, mais répandus dans la collection entière. Dans ce cas, tous les documents tendent à être récupérés et ceci affecte la précision de la recherche. Par conséquent, un nouveau facteur dépendant de la collection doit être introduit pour favoriser les termes concentrés dans quelques documents de la collection. Il s'agit de la fréquence inverse de document (IDF) [**3**] [**p.42-43**].

La fréquence inverse de chaque mot dans la base de documents (IDF – Inverse Document Frequency) est un poids qui dépend de la distribution de chaque mot dans la base de documents. Il exprime l'importance de chaque mot dans la base de données documentaire (le corpus). Les meilleurs termes pour l'identification du contenu sont ceux qui permettent de distinguer certains documents particuliers du reste de la collection. Ceci implique que les termes importants doivent avoir des fréquences élevées dans le document, mais une fréquence relativement faible dans la collection. Une mesure raisonnable pour l'importance du terme peut alors être obtenue avec le produit de la fréquence de terme et de la fréquence inverse de document, TFxIDF. TF/IDF est une technique qui utilise à la fois TF et IDF pour déterminer le poids d'un terme. En application dans la recherche d'information, on utilise ces mesures soient pour éliminer des mots qui ne sont

pas pertinents dans l'ensemble du corpus ou bien faire un ordre de pertinence à un ensemble de documents potentiels identifiés pouvant répondre à une requête.

3 État de l'art

S. Vijayarani et.al [2] discute de l'objectif du prétraitement des données textuelles, en soulignant les applications de l'extraction de texte et ses diverses éventualités. L'extraction de texte est le processus de recherche ou d'extraction de l'information utile à partir des données textuelles. Il essaie de trouver des modèles intéressants à partir de grandes bases de données. L'extraction d'information permet d'identifier les mots clés et les relations dans le texte, en utilisant des techniques d'appariement de motifs et en les convertissant en une base de données relationnelle.

Le Traitement Automatique du Language Naturel (TAL) (Natural Language Processing – NLP) explique comment les ordinateurs peuvent être utilisés pour comprendre et manipuler le texte du langage naturel. Les applications du TAL comprennent un certain nombre de domaines d'études, tels que la traduction automatique, le traitement et la synthèse de textes en langage naturel, les interfaces utilisateur, Cross Language Information Retrieval (CLIR), la reconnaissance vocale, intelligence artificielle et systèmes experts, etc.

Les étapes de prétraitement sont l'extraction, l'élimination des stop words, la racinisation ainsi que les algorithmes de TF/IDF. L'extraction est utilisée pour transformer le contenu du fichier en éléments individuels (Tokenisation). L'élimination des stop words vise à rendre le texte plus lourd et moins important pour l'analyse en supprimant les stop words pour réduire la dimensionnalité de l'espace de terme. La racinisation (stemming) trouve la racine des mots qui sont phonologiquement liés, c.-à-d., en supprimant les suffixes communs, en réduisant le nombre de mots, pour correspondre précisément aux racines. Enfin, les TF-IDF sert à montrer à quel point un mot est important dans un corpus. Leur article décrit avec détails plusieurs méthodes pour chaque étape du prétraitement.

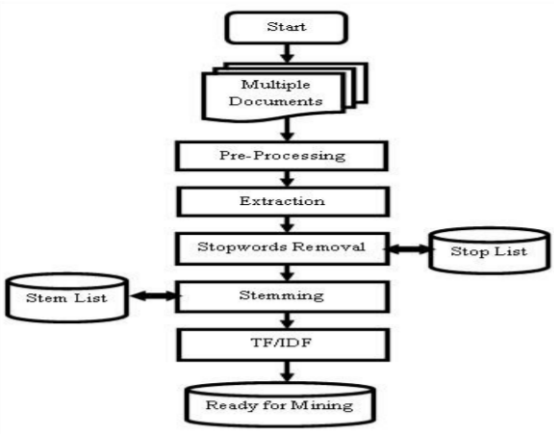

Figure 1: Étapes de pré-traitement du texte [2]

S.Kannan and V. Gurusamy [1] ont analysé l'importance du pré traitement dans le Data Mining, le traitement du langage naturel et la recherche d'information. Ils ont évalué les problèmes liés aux méthodes de pré traitement des archives de textes.

Ils proposent la tokenisation, la suppression des stop words et la racinisation (stemming) comme méthodes de pré traitement.

La tokenisation sonde les phrases et fait une liste de jetons (tokens) qui peuvent être utilisés comme entrée pour d'autres algorithmes. Le piège est que cela comprend la suppression d'éléments tels que les crochets, les tirets et d'autres ponctuations. En plus de cela, utiliser la tokenisation sur d'autres langues peut être laborieux.

Le stemming consiste à trouver la représentation commune des mots. Les lacunes de ce processus sont l'over-stemming et l'under-stemming.

L'étude s'articule autour de techniques de pré-traitement qui éliminent le bruit des données textuelles, racinisent, et coupent la taille des données textuelles.

C. Ramasubramanian et R. Ramya [4] ont travaillé sur les moyens d'améliorer les techniques d'endiguement utilisées dans le pré-traitement. L'article examine les inconvénients d'un des algorithmes de stemming appelé algorithme de Porter. Les inconvénients de l'approche existante sont discutés avec le processus pour les surmonter.

La vérification orthographique est ajoutée pour corriger les erreurs de correspondance et augmenter la précision, ce qui permet d'économiser du temps de traitement pour les mots mal orthographiés. Ils proposent donc une liste de mots intelligents qui supprime les mots d'arrêt efficacement sans supprimer les mots importants.

I. Singh and B. Saini [5] ont travaillé sur des techniques de pré-traitement efficaces pour les systèmes de recherche de l'information. Le document se concentre sur les algorithmes de tokenisation et de racinisation qui peuvent être utilisés lors du pré-traitement des données.

L'article considère la tokenisation comme l'une des étapes cruciales du pré-traitement des données textuelles. Il divise ce "sac de mots" en mots identifiables appelés jetons. La tokenisation fournit également des informations indiquant les fréquences de chaque token, qui peuvent être utilisées dans d'autres étapes de la recherche d'informations. Les algorithmes de suppression de stop words et de racinisation sont ensuite appliqués et la sortie ne contiendra que les jetons jugés valables par les algorithmes de pré-traitement.

Avec certains ensembles de données d'échantillons, une expérience montre une augmentation de l'efficacité pouvant atteindre 68 %, par rapport aux jetons générés après le pré-traitement, par rapport à ceux générés sans pré-traitement.

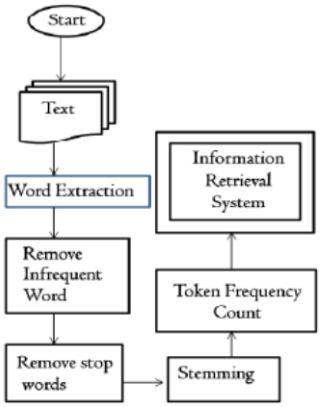

Figure 2: Processus de tokenisation [5]

4 Résultats et discussion

Cette section met un point sur les résultats de l'application des méthodes de pré-traitements. Nous montrons l'impact lié pour chaque méthode dans le processus et discutons également l'intérêt de chacun.

Le Dataset bibliographique utilisé pour nos tests est un corpus généré par le web scraping d'Arxiv. Arxiv est un répertoire libre d'accès d'article scientifique portant sur plusieurs thématiques comme l'astronomie, la physique, l'informatique, etc. Dans notre cas, nous avons choisi comme thème le Big Data. Le corpus obtenu possède 500 documents, contenant des données bibliographiques d'articles comme : le titre, les auteurs, la date de publication, le résumé... Nous avons appliqué nos pré-traitements que sur les résumés

d'articles. Le nombre total de mots sur notre dataset est de 84457.

La première étape de notre pré-traitement est le nettoyage du texte. Elle a pour objective d'éliminer les bruits du texte comme les caractères spéciaux, les balises de tout type de format, les ponctuations, remplacement des abréviations, effaçage des nombres, mise en minuscule du texte. La suppression du bruit est l'une des étapes de pré-traitement de texte les plus essentielles. Comme le montre la figure 1, à l'issue du nettoyage du texte, nous sommes passé de 84457 mots à 77646 mots, soit 6811 mots en moins.

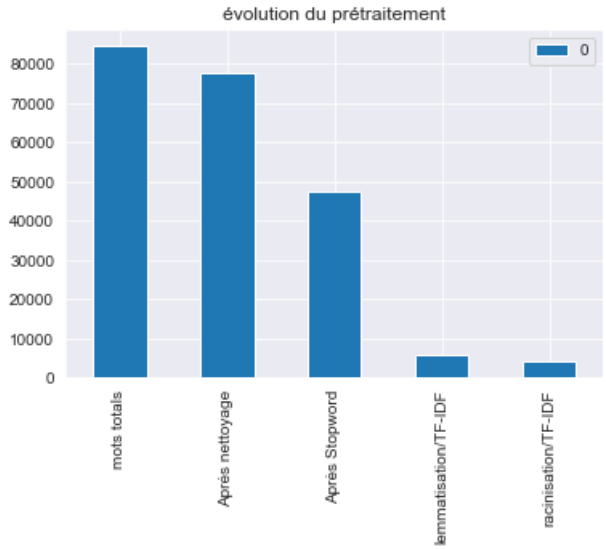

Figure 3: évolution du nombre de mots durant le pré-traitement

La deuxième étape réalisée est l'élimination des Stopwords. Les stopwords sont les mots couramment utilisés dans une langue comme les pronoms, les articles, les prépositions, les conjonctions, etc. Cette étape consiste à garder uniquement les mots les plus importants dans le texte. Nous sommes ainsi passés de 77646 mots à 47305 mots. Soit une perte de 39 % . Ce chiffre montre l'impact majeur des mots communs jugés inintéressants à l'analyse.

La tokenisation suit la deuxième étape.

La lemmatisation et la racinisation (Stemming) sont deux techniques semblables où le but est de supprimer les inflexions et de mapper un mot à sa forme racine. L'inconvénient de la racinisation est qu'elle conduit à des formes qui ne sont pas des mots ayant un sens grammatical. Donc, par soucis de compréhension, nous avons préféré utiliser la lemmatisation plutôt que la racinisation.

Après l'étape de lemmatisation ou racinisation, nous avons appliqué un TF-IDF pour déterminer l'importance de chaque terme dans le corpus. Le résultat de notre expérience présenté sur la figure 1 montre une valeur de 5820 mots obtenus et leurs poids après une lemmatisation contre 4094 mots avec la racinisation.

À l'issue de cette étape, nous avons enlevé les mots peu fréquents c'est-à-dire les mots avec un poids (score) égale à zéro.

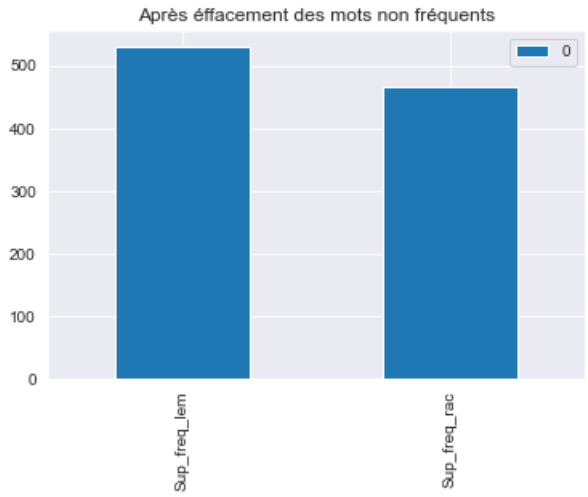

Figure 4: Suppression des mots avec un score TF-IDF nul

Comme le montre la figure 4, l'élimination des mots qui ont un poids nul

nous a permis de passer de 5820 à 529 mots avec la méthode de lemmatisation et 4094 à 465 mots avec la méthode de racinisation. En bref, les pré-traitements utilisés dans notre expérimentation nous ont permis de passer d'un dataset de 84457 mots à 529 mots (lemmatisation). Ce qui nous permet de conclure que ces 529 mots résument le mieux notre corpus.

Ces 529 mots retenus ont été vectorisés grâce à l'outil word2vec. Donc le résultat de notre pré-traitement est une matrice de 529 mots accompagnés de leurs vecteurs.

Cette vectorisation permet de calculer la distance entre deux mots et ainsi par exemple d'afficher sur un plan factoriel les mots les plus similaires à un autre mot après une réduction de dimensions.

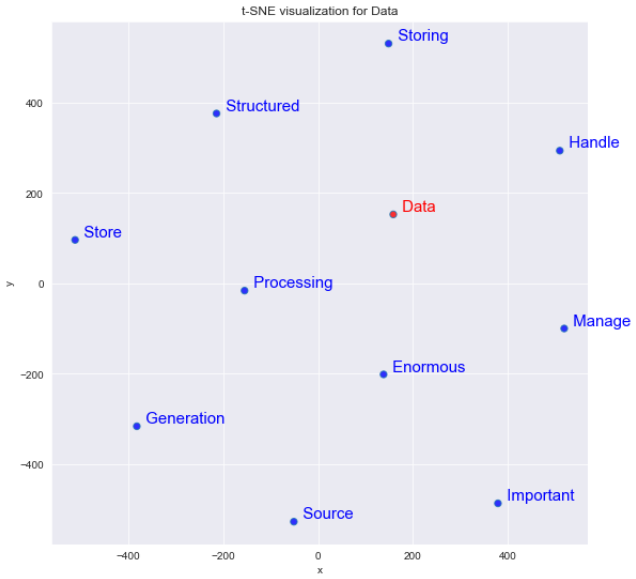

Figure 5: Mots similaires à DATA

Comme le montre la figure 5, en prenant comme exemple le mot Data, on retrouve tous les mots associés avec dans notre corpus final.

5 Conclusion

La recherche d'information étudie la manière de retrouver des informations dans un corpus qui est composé de documents d'une ou plusieurs bases de données. Cet article a permis de montrer l'efficacité des méthodes de prétraitement sur un corpus d'articles scientifiques. La réduction importante du nombre de mots permets de se focaliser sur la recherche d'information au sein du corpus.

Références

[1]. S.Kannan and V. Gurusamy: Preprocessing Technique for Text Mining; International Journal of Computer Science Communication Networks (2014)

[2]. S. Vijayarani, Ms. J. Ilmathi, Ms. Nithya: Preprocessing Technique for Text Mining - An Overview; International Journal of Computer Science Communication Networks (2015)

[3]. Imad Tbahriti : Modèles vectoriels et bibliométrique pour la recherche d'information et la détection de nouveauté appliqués à la protéomique ; thèse de doctorat université de Genève (2009)

[4]. C. Ramasubramanian, R. Ramya: Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm; International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, (2013)

[5]. I. Singh, B. Saini: An Effective Pre-Processing Algorithm for Information Retrieval Systems; International Journal of Database Management Systems, Vol.6, No.6, (2014)