Traitement automatique des langues
TD 1 (Analyse d'un corpus)

Youna Froger, Alice Gyde et Mathis Quais

8 octobre 2024

1 Réponses au questionnaire

1.1 What is Project Gutenberg, and what are its objectives and relevance?

Le projet Gutenberg est une bibliothèque virtuelle composée, en date de février 2024, de plus de 70 000 e-books. Il a été fondé en 1971 par Michael Hart. Le projet se donne comme objectif d' "encourager la distribution et la création d'e-books" (Hart, 2007). Il fonctionne de manière décentralisée et n'importe qui peut ajouter les livres qu'il souhaite, qui sont dans le domaine public. Les textes sont disponibles dans différents formats, comme en epub, html ou en UTF-8. Ce dernier a remplacé le format ASCII qui était utilisé au départ mais qui est problématique pour numériser d'autres langues que l'anglais. Le projet Gutenberg est utile dans le cadre du TAL car il offre, par la diversité des textes et des langues, un corpus particulièrement précieux.

1.2 What does Zipf's Law refer to, and is the same behavior expected across different languages?

La loi de Zipf stipule que lorsque l'on classe l'ensemble des mots d'un texte, leur fréquence d'apparition est inversement proportionnelle au rang et suit ainsi une échelle logarithmique (Piantadosi, 2014). Par exemple, si le premier mot revient 100 fois, ce sera 50 fois pour le second, 33 pour le troisième etc... Il s'agit d'une loi empirique, qui dans les faits ressemble davantage à une approximation. Elle a été formulée par George Kingsley Zipf durant la première moitié du XXe siècle. Son aspect scientiste est assez représentatif du contexte intellectuel de l'époque (Bertin & Lafouge, 2020). On considère généralement que cette loi peut s'appliquer à des textes de n'importe quelle langue, même si la variation peut évoluer selon le fonctionnement du langage en question. La loi de Zipf s'applique également dans d'autres domaines que la linguistique, comme en géographie.

1.3 What are stop words, and how can they be removed from a corpus?

Les stop words sont des mots qui n'influencent pas ou peu le sens d'un texte. Par exemple, ils peuvent être des prépositions, des articles, ou bien encore des pronoms. Il peut parfois être nécessaire de les exclure ou de les filtrer lors des analyses de texte. On peut les enlever en établissant une liste des mots que l'on veut retirer.

1.4 What are the advantages and disadvantages of removing stop words from a corpus? Which tasks would you not recommend it for?

Retirer les stop words permet d'enlever du bruit à l'analyse du texte, de se concentrer sur son sens et ses particularités. Néanmoins, en les retirant on peut se priver de certains éléments essentiels à l'analyse du texte, notamment le contexte de certaines portions. En retirant certains mots qui peuvent être utiles, cela peut même en détruire le sens. De manière générale, nous pouvons dire que le choix de les supprimer ou non dépend du travail que l'on veut effectuer. Par exemple, retirer les mots vides n'a pas beaucoup de sens dans le cadre d'une recherche par citation. De même, on peut imaginer que pour certaines formes d'écriture, comme la poésie, retirer les mots vides pourrait dénaturer le poème, en lui faisant perdre son rythme. Plus généralement, si les chercheurs avaient tendance à beaucoup retirer les stop words il y a quelques années, ce n'est plus le cas aujourd'hui. En effet, cette pratique répondait alors à des nécessités d'optimisation. Aujourd'hui, avec la grande puissance des machines cela n'est plus nécessaire et la tendance s'est renversée. On travaille en prenant les données comme telles, sans chercher à retirer du bruit.

1.5 How can the NLTK Python package be used to access and analyze texts from Project Gutenberg?

NLTK est une librairie Python dédiée au traitement automatique des langues (Bird et al., 2024). Elle est notamment utilisée pour la linguistique, mais aussi dans d'autres domaines, comme le journalisme. Ce package nous permet de réaliser différentes opérations spécifiques au TAL, comme la tokenization, la recherche de similitudes entre les mots d'un texte, la production de certaines informations statistiques... Il est possible d'importer directement des corpus depuis cette librairie, dont notamment des textes issus du projet Gutenberg. On peut noter qu'il en existe une autre librairie, CLTK qui fonctionne de la même manière et qui est dédiée aux langues anciennes comme le latin ou le grec.

1.6 How does NLTK handle stop words, and how can you customize the stop word list for specific tasks?

La librairie NLTK nous offre différentes possibilités pour gérer les mots vides. De base, elle nous propose des listes de stop words dans différentes langues que l'on peut importer pour les retirer directement. Mais il est également possible de personnaliser une liste pour un usage spécifique.

2 Exercices du TD

Voir le lien GitHub : https://github.com/younafroger/TAL$_H NM$2.
Sur le GoogleColab, les fichiers se suppriment automatiquement. Il est important de rajouter les fichiers dont nous avons besoin (accessible sur Github dans TP1).

3 Conclusion

Au travers de ce travail, nous avons pu constater que même sans utiliser une librairie destinée au traitement automatique des langues, quelques lignes de code en Python suffisent à obtenir des résultats assez pertinents sur le fonctionnement des langues et leurs particularités. Même si certaines connaissances de base en programmation sont nécessaires, une connaissance poussée en informatique n'est pas requise pour avoir accès aux possibilités offertes par la TAL. De notre côté, nous n'avons pas rencontré de difficultés particulières dans le cadre de cet exercice.

En ce qui concerne l'exercice en lui-même, le graphique montre que dans toutes les langues étudiées, la taille des mots est généralement comprise entre 1 et 10 lettres, à l'exception du mandarin. Il s'agit à chaque fois de courbes suivant une structure logarithmique, ce qui semble valider l'application de la loi de Zipf, même si c'est d'une manière plus ou moins marquée selon les langues. Certaines ont des courbes très similaires. Assez logiquement, ce sont ici des langues appartenant à la même famille. Des langues romanes, comme le français et l'espagnol, ont par exemple des courbes qui se chevauchent. A l'inverse, le finnois et l'arabe ont une courbe beaucoup plus lissée, montrant une distribution plus équilibrée. Le mandarin et la japonais se distinguent avec des courbes vraiment différentes. Pour le mandarin, la courbe se concentre autour de 15 lettres. Ce fait peut être causé par la structure de la langue, ou bien une utilisation réduite des mots vides.

En ce qui concerne nos domaines personnels, nous pouvons imaginer différentes applications de ce genre de traitement du langage. En histoire, en reprenant les morceaux de code produits lors du TP, cela pourrait nous permettre de comparer différentes sources textuelles de notre corpus. Par exemple, nous pourrions ainsi probablement obtenir des informations sur le niveau de langue des différents auteurs, surtout si l'on travaille sur un corpus avec des personnes issues de classes

sociales différentes et ayant donc un rapport différent à l'écrit. De même, en modifiant notre code nous pourrions voir aussi la variété des mots utilisés dans un texte. Peut-être que des individus disposant d'un capital culturel moins élevé, avec ainsi un autre accès à la culture et l'éducation, auraient un vocabulaire moins varié. De ce fait, la loi de Zipf serait ainsi plus accentué car la distribution des mots serait moins réparti ? De cette manière, nous pourrions établir des catégories initiales pour les différents auteurs de nos sources.

References

Bertin, M., & Lafouge, T. (2020). La loi de zipf 70 ans après : Pluridisciplinarité, modèles et controverses. *Communication & langages*, *206*(4), 111–134.

Bird, S., Klein, E., & Loper, E. (2024). *Nltk documentation* [Consulté le : 06/10/2024]. https://www.nltk.org/

Hart, M. (2007). The project gutenberg mission statement [Consulté le : 06/10/2024]. https://www.gutenberg.org/about/background/mission_statement.html

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130.