M2SOL034 Corpus, ressources et linguistique outillée

TD1 : Loi de Zipf et pré-traitement du texte

Ljudmila PETKOVIĆ

Sorbonne Université
Master « Langue et Informatique » (M1 ScLan)
UFR Sociologie et Informatique pour les Sciences Humaines
Semestre 2, 2024-2025, le 31 janvier 2025

Table des matières

1 Pré-traitement du texte **1**

2 Loi de Zipf **2**

Les ressources pour le TD1 sont disponibles sur le dépôt GitHub : `https://github.com/ljpetkovic/M2SOL034/tree/main/_TD/TD1`.

Vous pouvez utiliser Jupyter Notebook ou Google Colab (si vous optez pour la dernière méthode, les liens dédiés sont fournis pour chaque exercice).

1 Pré-traitement du texte

L'objectif de cet exercice est de mettre en pratique quelques concepts fondamentaux du TAL : tokenisation, lemmatization, racinisation et la segmentation de phrases.

Suivez le tutoriel `1_nlp_basics_tokenization_segmentation.ipynb` de Saravia (*s.d.*) et complétez quatre exercices proposés.

1. copiez le code indiqué dans le tutoriel et ajoutez des espaces supplémentaires à la valeur de chaîne attribuée à la variable `doc` et identifiez le problème avec le code. Essayez ensuite de résoudre le problème. Astuce : utilisez `text.strip()` pour résoudre le problème ;

2. essayez le code indiqué avec différentes phrases et voyez si vous obtenez des résultats inattendus. Essayez également d'ajouter des ponctuations et des espaces supplémentaires, plus courants dans le langage naturel. Que se passe-t-il ? ;
3. essayez d'utiliser différentes phrases dans le code indiqué et observez l'effet du racinisateur ;
4. créez votre propre algorithme de segmentation de phrases en utilisant `spaCy`.

2 Loi de Zipf

L'objectif de cet exercice est d'implémenter la loi de Zipf en Python.

À partir du script `Zipf_exo.ipynb`, réaliser les étapes suivantes :

1. Charger un texte depuis le fichier `zipf.txt`
2. Pré-traiter le texte : convertir le texte en minuscules
3. Compter la fréquence des mots
4. Trier les mots par fréquence
5. Extraire les fréquences et les rangs
6. Afficher les mots avec leurs fréquences
7. Tracer la loi de Zipf

Références

Saravia, E. (*s.d.*). *Fundamentals of NLP (Chapter 1) : Tokenization, Lemmatization, Stemming, and Sentence Segmentation*. GitHub https://github.com/dair-ai/nlp_fundamentals/blob/master/1_nlp_basics_tokenization_segmentation.ipynb. Consulté le 30 janvier 2025 (cf. p. 1).