Programmation de spécialité (python)
TD 9-10 : mini projet

Julien Velcin

2023-2024

L'objectif de ces deux dernières séances est de travailler sur une petite interface visuelle permettant d'accéder au corpus de documents que vous avez traité tout au long de nos séances. Le sujet n'est pas très précis car il vous permet de proposer vos propres idées et tester les librairies Python que vous voudriez essayer.

Enoncé général

Les documents contiennent énormément d'informations pertinentes pour les chercheurs en SHS (sociologues, linguistes, etc.). Néanmoins, la masse de données rend compliquée leur exploration. Pour cela, on vous propose de créer un outil permettant d'explorer les documents en adoptant une approche comparative. Cet outil sera donc destiné à des utilisateurs qui ne sont pas nécessairement informaticiens.

Quelques pistes

La première piste pour créer ce type d'outil est de permettre une analyse comparative de deux corpus (par ex. comparer les articles issus de Reddit de ceux d'Arxiv).

La seconde piste consiste à permettre d'observer l'évolution temporelle d'un mot (ou d'un groupe de mots) donné. Il faut alors être capable de traiter le champ date pour découper la frise temporelle en périodes. Dans les deux cas, il s'agit d'observer l'importance relative d'un ou plusieurs mots dans un corpus vs. un autre corpus : quels sont les mots communs ? les mots spécifiques ?

Au-delà de la simple fréquence d'un mot (TF), des mesures de l'importance d'un terme dans un corpus sont faciles à coder. La mesure comme TFxIDF (https://fr.wikipedia.org/wiki/TF-IDF) et OKAPI-BM25 (https://fr.wikipedia.org/wiki/Okapi_BM25) constituent une bonne base de départ. Vous pourrez proposer une analyse du vocabulaire en utilisant ces scores. Il semble cependant nécessaire de les adapter un peu, par ex. en considérant un corpus comme un "gros" document.

En terme d'interaction, il semble important de donner la possibilité à l'utilisateur de formuler des requêtes. Ces requêtes peuvent prendre la forme de mots clefs (comme dans le TD sur le moteur de recherche), mais il pourrait aussi intégrer la possibilité de sélectionner un auteur, un type de source ou une date / période.

Toutes les (bonnes) idées sont les bienvenues !