TP3 : Clustering et word2vec

Traitement automatique de corpus

4 novembre 2025

Sur la base des éléments méthodologiques et des enseignements techniques présentés lors du cours théorique, il est demandé dans le cadre de ce TP :

— d'effectuer un clustering des documents CAMille pour une décennie au choix et d'interpréter les résultats
— d'entraîner un modèle word2vec sur l'ensemble du corpus et d'explorer les relations entre vecteurs

Pour ce faire, vous utiliserez différentes librairies Python vues au cours comme scikit-learn et gensim.

Les étapes à mettre en œuvre sont les suivantes :

1. Choisissez une décennie (1890–1899, 1900–1909, 1910–1919, etc.)
2. Effectuez un clustering des documents de cette décennie grâce au notebook `s2_clustering.ipynb`, en adaptant éventuellement le nombre de clusters désirés
3. Tentez d'interpréter les résultats obtenus : les clusters semblent-ils faire sens ? Si oui/non, comment l'expliquez-vous ? Aidez-vous au besoin d'une analyse de keywords et/ou wordcloud (voir TP2).
4. Téléchargez sur l'UV le fichier zippé `sents.txt`, déjà segmenté en phrases, et placez-le dans le dossier data
5. Entraînez un modèle word2vec (word embeddings) sur ces phrases grâce au notebook `s3_word_embeddings.ipynb`, en adaptant éventuellement les paramètres window (taille de la fenêtre) et min_count (nombre minimum d'occurrences d'un mot)
6. Vous pouvez entraîner plusieurs modèles afin de comparer leurs performances, en procédant par essais-erreurs pour choisir le meilleur modèle
7. Explorez le modèle retenu à l'aide des fonctions `similarity` et `most_similar` (choisissez au moins trois exemples pour chaque fonction)
8. Rassemblez votre code dans un nouveau notebook `tp3.ipynb` placé dans le dossier « tp3 » (n'oubliez pas de pusher)
9. Soumettez sur l'UV un court rapport (3-4 pages, format PDF) présentant votre méthodologie et les résultats obtenus

Pour toute question, n'hésitez pas à me contacter via `max.de.wilde@ulb.be` en mettant toujours Denis Lebailly en copie (`denis.lebailly@ulb.be`).