Fouille de données lexicale

Exercices

Tous ces exercices utilisent les fichiers de données du Grand Débat National (GDN).

Essayez vos implémentations sur l'échantillon (*sample*) avant de les tester sur le corpus complet.

Segmentation

Implémentez un segmenteur simple qui sépare les tokens selon les espaces et les ponctuations.

Réalisez des segmentations d'un extrait du corpus avec :

- votre segmenteur
- NLTK
- spacy
- stanza
- BERT

Comparez le nombre de tokens et la liste des tokens.

Promptez un LLM et comparez le résultat produit.

Exécutez cela sur le corpus et comparez le nombre de tokens, la taille du vocabulaire, le temps d'excécution.

Recherche de motifs séquentiels

Recherchez le nombre d'occurrences du mot « citoyen » dans le corpus avec :

- une expression régulière
- spacy
- Unitex

Faites de même avec le verbe « croire » et comparez qualitativement et quantitativement les résultats.

Modélisez par des graphes Unitex :

- les expressions de croyance
- les ministères français réels, puis ceux qui sont imaginés ou proposés,
- les personnalités publiques (présidents, ministres, secrétaires, etc.).

Recherche de motifs de graphe

À l'aide du `DependencyMatcher` de spaCy, recherchez :

- les sujets des verbes « dire » et « déclarer »
- les paires de sujets et d'objets du verbe « croire »