STIC-B545

TRAITEMENT AUTOMATIQUE DE CORPUS (TP2)

Etudiant : Fopa Kemnang Armel Kesnel

Matricule : 000604480

Le TP2 avait pour objectif l'utilisation de 4 notebooks pour effectuer plusieurs analyses automatiques sur des textes extraits du corpus CAMille, dans le but de recueillir des données linguistiques et statistiques pertinentes.

Nous avons décidé de travailler sur l'ensemble du corpus de l'année 1950, lequel comprenais 04 fichiers texte issus des journaux numérises.

Les analyses ont été réalisées à l'aide de quatre notebooks :

- Extraction de mots-clés (s1_keywords.ipynb)
- Génération d'un nuage de mots (s2_wordcloud.ipynb)
- Reconnaissance d'entités nommées (s3_ner.ipynb)
- Analyse de sentiment (s4_sentiment.ipynb)

1. Extraction de mots clés

Nous avons appliqué l'algorithme YAKE (Yet Another Keyword Extractor) au corpus 1950.

Ici, nous avons débuté un premier traitement sur un fichier, puis nous l'avons étendu sur l'ensemble des 4 fichiers.

Le filtrage des bigrammes a permis d'obtenir des termes caractéristiques de la presse de l'époque, tels que : *Van Zeeland roi Léopold, cercle dramatique , Morhet reprendra , pâte dentifrice.*

Ils traduisent le contexte socio-culturel et politique dans les années 50 marqué par des noms propres et des thématiques journalistiques.

2. Nuage de mots

Le nettoyage lexical effectué dans ce cadre a été fait grâce à ntlk et nous avons enrichi la liste de stopwords pour l'adapter au corpus, en supprimant notamment des formes fréquentes ou OCR bruitées telles que Van, bruxelles rossel.

Le nuage de mots généré met en évidence les expressions suivantes : **van**, **avenir**, **bruxelles**, **luxembourg**, **soirée**, **samedi**, **belgique**, **dramatique**, **cercle**, **fanfare**…

Ce graphique illustre la densité thématique du corpus.

3. Reconnaissance d'entités nommées

L'extraction des entités nommées a été réalisée à l'aide du modèle fr_core_news_sm de spacy. Un léger prétraitement a permis de réduire les erreurs dues à l'OCR.

Les résultats sont présentés dans le tableau suivant :

Type d'entités	**Exemple**	**commentaire**
PER	Van Zeeland, Renée Cara, Philippe Gérard, Mme de Sermaize	Noms de personnalités culturelles ou politiques
LOC	Bruxelles, Namur, Charleroi, Paris	Coherent avec le contexte belge
ORG	Sénat, Chambre, Ford, Prodent	Institutions politiques et marques commerciales

Certaines erreurs subsistent (ex. *"Samedi" : PER*), mais elles illustrent bien les **limites d'un modèle générique appliqué à un corpus OCR ancien**.

4. Analyse de sentiments

Nous avons utilisé ici deux approches différentes, et les résultats sont présentés dans le tableau suivant :

Méthode	Outil	Résumé des résultats
TextBlob-FR	Analyse lexicale (PatternAnalyzer)	Polarités très neutres (0.0 à 0.2), subjectivité faible → style journalistique objectif
Transformers (BERT)	Modèle nlptown/bert-base-multilingual-uncased-sentiment	Notes entre **3 et 5 étoiles**, reflétant un ton globalement positif dans les articles culturels

Exemple :

- Jeux de lumière très bien appropriés : 5 stars (0.62)
- Le même spectacle sera rendu ce dimanche : 1 star (0.40)

Globalement, le corpus reflète un **ton neutre à légèrement positif**, typique de la presse descriptive et promotionnelle.

Conclusion

Ce TP a permis de découvrir plusieurs techniques de traitement automatique de corpus :

- Extraction lexicale (YAKE)
- Visualisation (WordCloud)
- Reconnaissance d'entités (NER – spaCy)
- Analyse de sentiments (TextBlob & BERT)

L'ensemble des résultats met en lumière les **spécificités du corpus CAMille 1950** :
des textes culturels, un style journalistique peu subjectif, mais riches en entités géographiques et en noms propres.