# *STIC-B545*

# *TRAITEMENT AUTOMATIQUE DE CORPUS (TP3)*

***Etudiant : Fopa Kemnang Armel Kesnel***

***Matricule : 000604480***

Le TP3 consiste à étudier différentes approches de représentation et de classification de documents textuels.

Les TP précédents consistaient en une analyse lexicale et les matérialisations des nuages de mots, l'objectif de celui-ci est d'appliquer des méthodes d'apprentissage automatique dans le but de :

- Classifier automatiquement les documents selon leur contenu,
- Regrouper les textes similaires sans catégorie prédéfinie,
- Représenter les mots selon leurs relations sémantiques à l'aide de Word Embeddings.

L'étude a été menée sur un corpus standard de la collection 20 NewsGroups pour la classification supervisée, et sur le corpus CAMille pour le clustering et l'analyse sémantique.

## 1. Méthodologie

### a. Prétraitement du corpus

Les textes ont initialement été nettoyés à l'aide des expressions régulières et des stopwords.

Nous avons :

- Mis en minuscule,
- Supprimé la ponctuation et des chiffres,
- Eliminé des mots très courts et des mots vide.

Pour les expérimentations sur CAMille, nous nous sommes concentrés sur la décennie 1920 afin de garantir la cohérence temporelle du corpus.

### b. Représentation vectorielle (TF-IDF)

Tous les documents sont représentés sous forme de vecteurs TF-IDF, permettant ainsi l'évaluation de l'importance relative des mots dans un document par rapport à l'ensemble du corpus.

Nous avons utilisé les paramètres suivants :

- max_df = 0.8, min_df = 2, permettant de laisser de côté les mots trop fréquents et très rares,
- lowercase = True,
- stop_words = sw.

### c. Classification supervisée

Nous avons réalisé la classification à l'aide du modèle Naïve Bayes multinomial sur un sous-ensemble du corpus 20 Newsgroups.

Un pipeline TfidVectorizer + MultimonialNB a été mis sur place et nous avons obtenu les résultats suivants :

- Précision : 0,906,
- Rapport de classification : F1-score ≈ 0,91 pour l'ensemble des classes.
- Matrice de confusion : Les erreurs sont limitées entre les classes *sci.space* et *comp.graphics*

Cette étape met en évidence la capacité du modèle à discriminer efficacement des thématiques distinctes à partir de simples représentations lexicales.

### d. Clustering non supervisé

Dans cette partie, nous avons utilisé une approche non supervisée à l'aide de l'algorithme K-Means.

Ici, l'objectif était d'identifier automatiquement des regroupements thématiques au sein des articles des années 1920.

- **Taille de la matrice TF-IDF** : (7 × 2747)
- **Résultats du clustering :**
  - Cluster 0 → vocabulaire économique / colonial (*prix, Congo, Belg, industrie…*)
  - Cluster 1 → thématique sociale / culturelle (*rue, film, ouvriers, journal…*)
  - Cluster 2 → documents bruités (résidus OCR)
- **Visualisations**
  - Nuage de mots (WordCloud) par cluster : permet de visualiser les termes dominants dans chaque groupe.

- Réduction de dimension (PCA) : visualisation 2D des documents, où les clusters apparaissent clairement séparés selon leur contenu lexical.

### e. Représentation sémantique (Word Embeddings)

Le modèle Word2Vec a été entraîné sur les textes des années 1920 pour étudier les relations sémantiques entre les mots.

La configuration a été faite comme suit :

- vector_size = 100, window = 5, min_count = 2, sg = 1

**Exemples de similarités:**

| **Mot** | **Mots proches** |
|---|---|
| **Prince** | Rue, deux, place, plus, belge |
| **Princesse** | Belgique, deux, plus, rue. dun |
| **Famille** | Belge, tout, ans, rue, dem |

Ces résultats, bien que bruités à cause du faible volume de texte, montrent que le modèle capture certaines proximités contextuelles (royauté, société, nation).

**Visualisation PCA**

Une projection en deux dimensions des vecteurs de mots révèle plusieurs pôles lexicaux :

- Un pôle “royauté” autour de *prince, princesse, mariage, famille* ;
- Un pôle “social / national” autour de *belgique, ouvriers, congo*, *prix*.

## 2. Discussions et analyse

Les représentations TF-IDF et Word2Vec ont permis d’aborder la structure du corpus sous deux angles complémentaires :

- **TF-IDF + KMeans** : regroupement thématique des textes selon la fréquence des mots ;
- **Word2Vec** : représentation sémantique continue des mots et visualisation de leurs relations de sens

La correspondance entre les clusters et les champs sémantiques met en évidence l’efficacité des approches vectorielles, même lorsqu’elles sont appliquées à un corpus historique restreint.

Nous constatons que la performance du modèle demeure limitée par le bruit issu de l'OCR et par la taille réduite du corpus.
L'intégration de bigrammes (ngram_range = (1,2)) ou l'élargissement du corpus contribuerait à accroître la précision des regroupements et des mesures de similarité.

Enfin, les différentes visualisations (WordClouds, PCA des documents, PCA des embeddings) ont permis d'interpréter graphiquement la structure lexicale et sémantique du corpus, renforçant l'analyse quantitative issue des modèles.

**Conclusion**

Ce troisième TP nous a permis de travailler sur différentes approches du **traitement automatique de corpus** :

- La classification supervisée de documents (Naïve Bayes),
- Le clustering non supervisé (K-Means + PCA + WordClouds),
- Et la représentation sémantique par **Word2Vec**.

Ces trois niveaux d'analyse montrent la complémentarité entre :

- Les approches **statistiques** (TF-IDF, KMeans) et
- Les approches **distributionnelles** (Word2Vec).

Malgré la taille réduite du corpus CAMille, les résultats obtenus confirment la pertinence des représentations vectorielles pour la fouille de textes historiques et ouvrent des perspectives vers des analyses diachroniques (évolution des mots selon les décennies).