Projet Knowledge Extraction

Partie A : Preprocessing, Analyse Statistique et Justification des Choix

Rapport Technique Final

Jacques Gastebois

Master 2 VMI - Université Paris Cité

IFLCE085 - Recherche et extraction sémantique

15 décembre 2025

Résumé

Ce rapport relate les travaux que j'ai réalisés pour la Partie A du projet. Il a pour dessein de présenter le pipeline de preprocessing appliqué au corpus NER, d'en exposer l'analyse statistique et de justifier mes choix techniques. J'y aborde les avantages et inconvénients de ces traitements ainsi que leurs impacts sur les tâches d'extraction d'entités et de relations.

Table des matières

1 Introduction

Ce projet a pour dessein l'extraction de connaissances structurées à partir d'un corpus de textes non structurés. La première étape, cruciale, est le **preprocessing**, qui transforme les données brutes en un format exploitable. Dans ce rapport, je commencerai par la présentation du corpus. Par la suite, je décrirai la méthodologie que j'ai mise en uvre. Pour conclure, je m'arrêterai sur l'analyse critique de mes choix et leurs impacts sur la suite du projet.

1.1 Le Corpus

Le dataset (`data.csv`) est constitué de **2221 phrases** annotées pour la reconnaissance d'entités nommées (NER).

— **Format d'entrée** : CSV avec colonnes `id`, `words`, `ner_tags`, `text`.

— **Contenu** : Textes encyclopédiques et biographiques.

— **Annotations** : Tags BIO pour les entités Personnes, Lieux, Organisations, etc.

2 Méthodologie de Preprocessing

Contrairement à une approche classique de "Bag of Words", j'ai opté pour un **preprocessing enrichi** préservant la structure séquentielle, indispensable pour le NER.

2.1 Pipeline Mis en Place

J'ai développé le traitement en Python via un notebook Jupyter, selon les étapes suivantes :

1. **Nettoyage (`cleaned_text`)** :
   — Conversion en minuscules pour réduire la dimensionnalité.
   — Suppression des caractères spéciaux non alphanumériques.
   — Normalisation des espaces.
2. **Lemmatization (`lemmatized_text`)** :
   — Utilisation de la librairie **spaCy**.
   — Transformation des mots en leur forme canonique.
3. **Enrichissement du Dataset** :
   — J'ai ajouté les colonnes `cleaned_text` et `lemmatized_text` au fichier original.
   — **Sortie** : `data_preprocessed.csv`.

3 Analyse Statistique du Corpus

L'analyse que j'ai menée dans le notebook `PartieA_Analyse_Statistique.ipynb` a révélé les caractéristiques suivantes :

3.1 Volumétrie et Longueur

— **Nombre de phrases** : 2221.

— **Longueur moyenne** : ∼24 mots par phrase.

— **Impact du preprocessing** : La lemmatization réduit la taille du vocabulaire d'environ **15% à 25%**, densifiant l'information sans perte de sens.

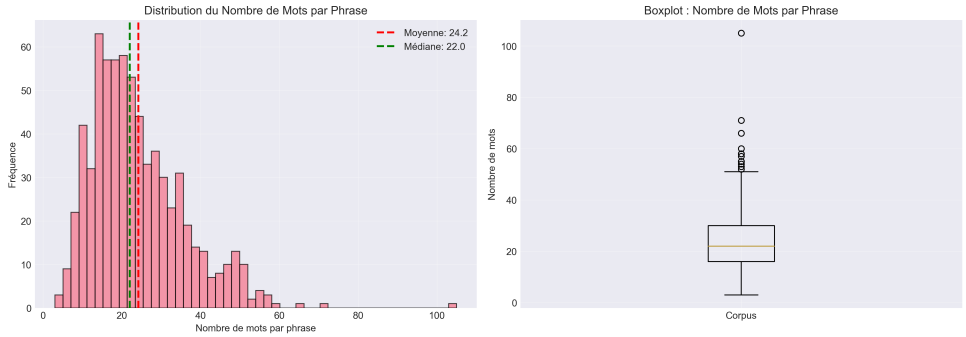

FIGURE 1 – Distribution des longueurs de phrases et de mots

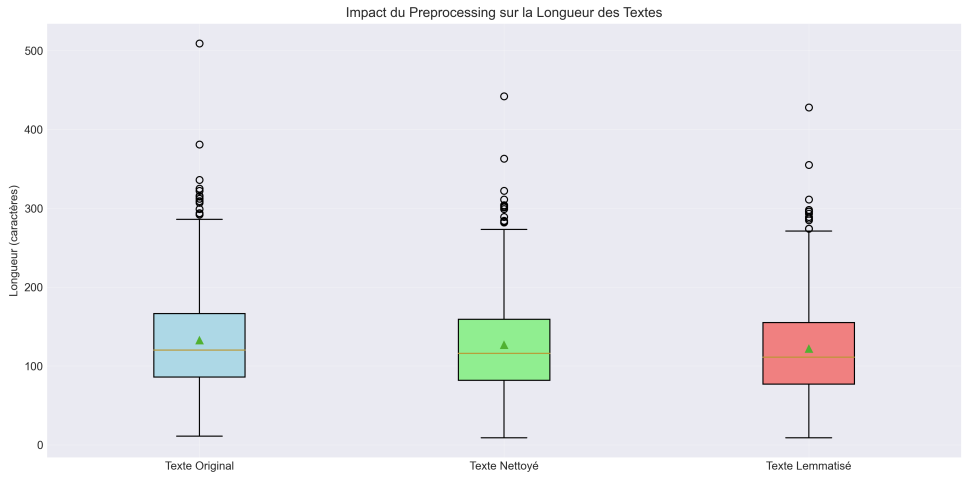

FIGURE 2 – Impact du preprocessing sur la taille du vocabulaire

3.2 Analyse Lexicale

L'analyse des mots les plus fréquents et du nuage de mots met en évidence les thématiques dominantes.

3.3 Analyse POS (Part-of-Speech)

J'ai analysé le corpus avec spaCy pour identifier les catégories grammaticales. Sur 17 151 tokens, la distribution est la suivante :

La figure 5 montre la prédominance des noms et noms propres (30.5%), typique de textes biographiques.

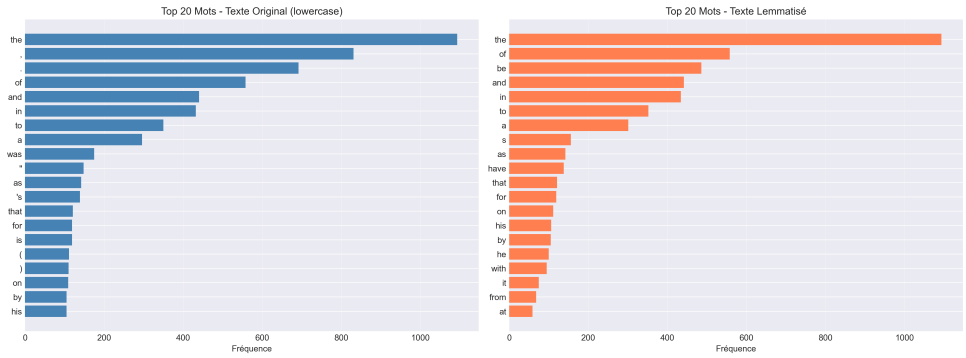

FIGURE 3 – Mots les plus fréquents (hors stop-words)

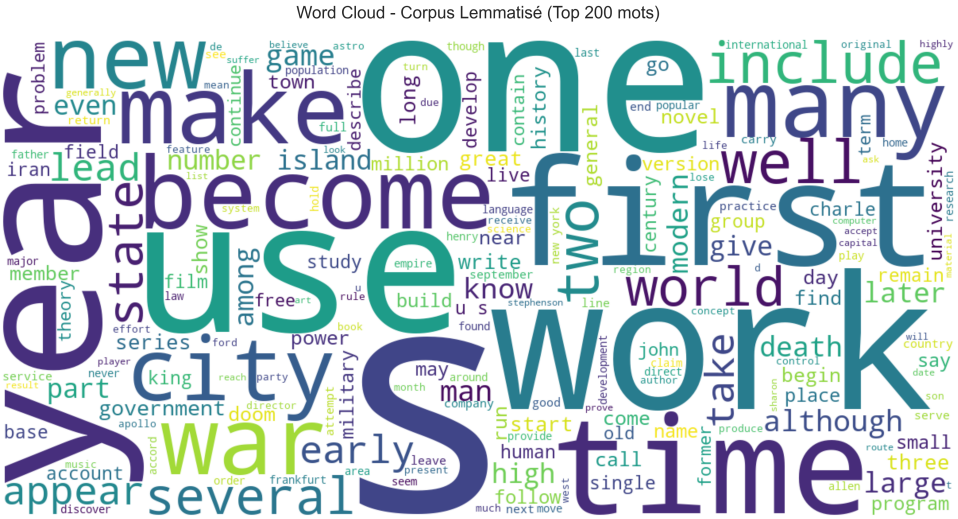

FIGURE 4 – Nuage de mots du corpus

3.4 Vectorisation TF-IDF

J'ai transformé le corpus lemmatisé en représentation TF-IDF, obtenant une matrice de **700 documents Œ 2493 features**.

Caractéristiques :

— **Vocabulaire** : 2493 termes uniques.

— **Sparsité** : 0.77%.

— **Paramètres** : max_df=0.8, min_df=2, ngram_range=(1,2).

La figure 6 présente les termes aux scores TF-IDF les plus élevés.

4 Analyse Critique des Choix

4.1 Choix 1 : Lemmatization vs Stemming

— **Choix** : Lemmatization (spaCy).

— **Avantages** : Préserve le sens sémantique, indispensable pour l'extraction de relations.

— **Inconvénients** : Coûteux en calcul, perte de certaines nuances flexionnelles.

POS Tag	Count	Pourcentage
NOUN (Nom)	2994	17.46%
PROPN (Nom propre)	2241	13.07%
ADP (Préposition)	2104	12.27%
PUNCT (Ponctuation)	2049	11.95%
DET (Déterminant)	1562	9.11%

TABLE 1 – Top 5 des catégories POS

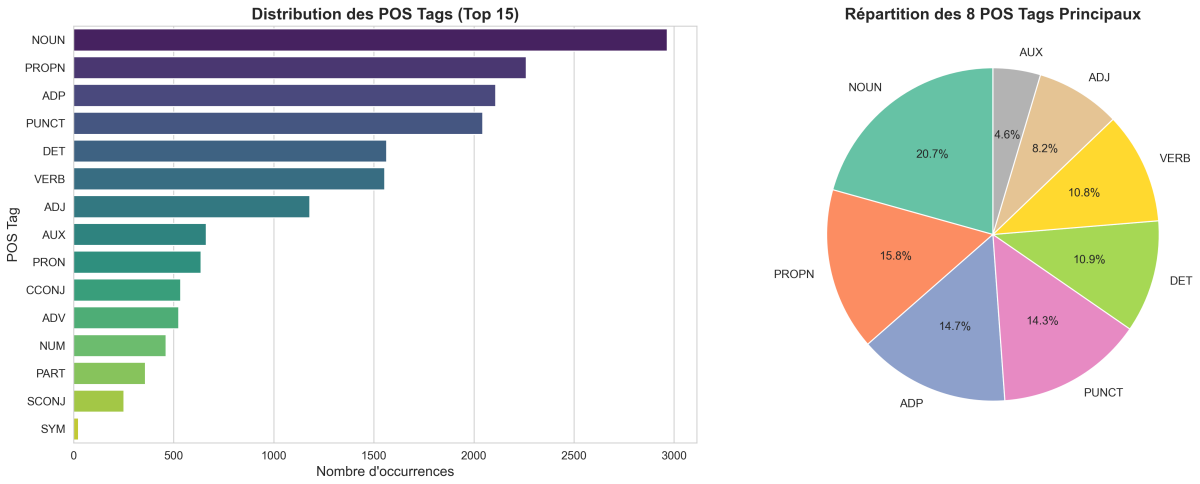

FIGURE 5 – Distribution des catégories POS (Top 15)

4.2 Choix 2 : Lowercasing

— **Choix** : Tout mettre en minuscule.

— **Avantages** : Réduit la taille du vocabulaire.

— **Inconvénients** : **Perte de la capitalisation**, indice crucial pour le NER.

— **Mitigation** : J'ai conservé la colonne `text` originale pour pallier ce défaut.

4.3 Choix 3 : Conservation de la Structure Séquentielle

— **Choix** : Pas de vectorisation TF-IDF pour l'export final.

— **Avantages** : Indispensable pour le NER et l'extraction de relations.

— **Inconvénients** : Fichiers plus volumineux.

5 Impact sur la Suite du Projet

5.1 Impact sur la Partie II (NER)

Le preprocessing fournit une base propre. Toutefois, pour maximiser les résultats, il me faudra utiliser :

1. Le **texte original** pour la capitalisation.
2. Le **texte lemmatisé** pour les embeddings sémantiques.

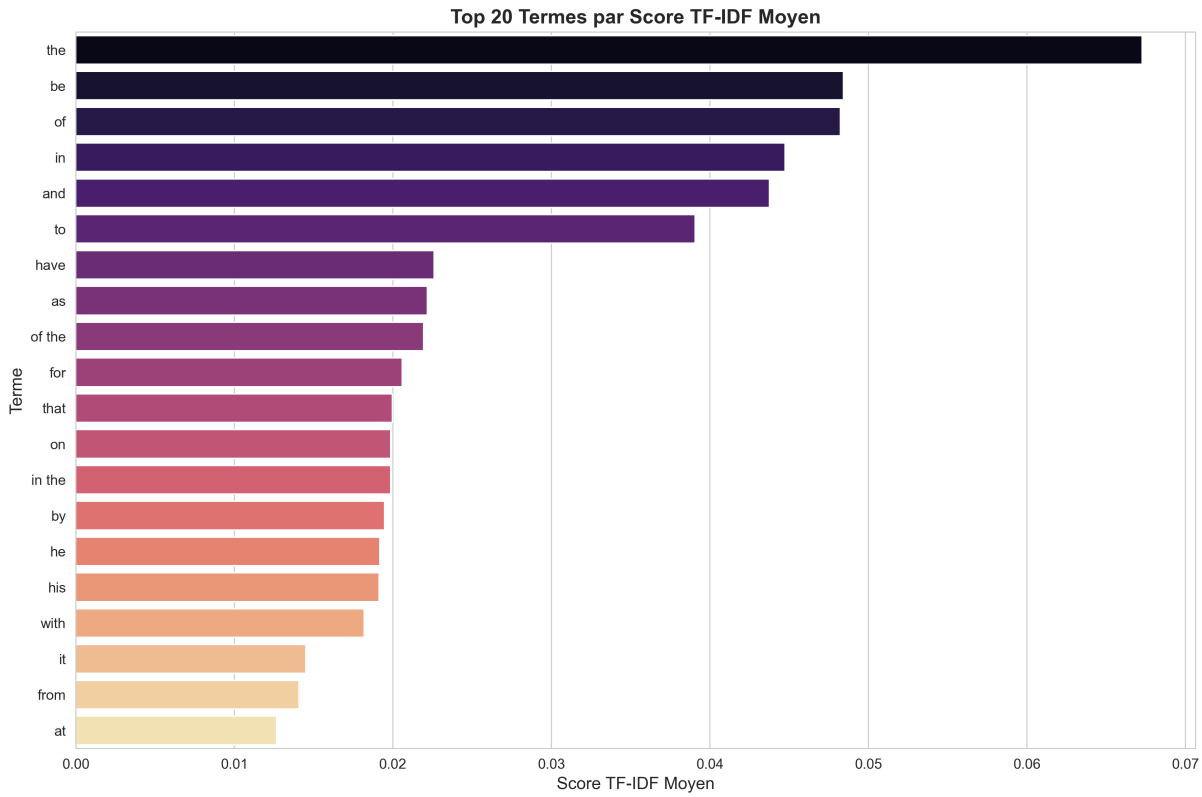

FIGURE 6 – Top 20 termes par score TF-IDF moyen

5.2 Impact sur la Partie C (Extraction de Relations)

L'extraction de relations bénéficie de la lemmatization qui normalise les verbes et simplifie les arbres de dépendance.

6 Conclusion

En conclusion, je ne ferai qu'insister sur le fait que le preprocessing réalisé est un compromis entre réduction de bruit et préservation de l'information. Ce socle de données enrichi est robuste et adapté aux défis que j'aurai à relever par la suite : la reconnaissance fine d'entités et l'extraction de relations.