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

14 décembre 2025

Résumé

Ce rapport synthétise les travaux réalisés pour la Partie A du projet. Il détaille le pipeline de preprocessing appliqué au corpus NER (2221 phrases), présente une analyse statistique des données, et justifie les choix techniques (nettoyage, lemmatization) en analysant leurs avantages, inconvénients et impacts sur les tâches d'extraction d'entités (Partie B) et de relations (Partie C).

Table des matières

1 Introduction

Le projet vise à extraire des 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 pour les algorithmes d'apprentissage automatique.

1.1 Le Corpus

Le dataset utilisé (`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/biographiques.

— **Annotations** : Tags BIO (Begin, Inside, Outside) pour les entités Personnes (PER), Lieux (LOC), Organisations (ORG), etc.

2 Méthodologie de Preprocessing

Contrairement à une approche classique de "Bag of Words" (TF-IDF), nous avons opté pour une approche de **preprocessing enrichi** qui préserve la structure séquentielle des données, essentielle pour le NER.

2.1 Pipeline Mis en Place

Le traitement a été réalisé en Python (via un notebook Jupyter) et comprend les étapes suivantes :

1. **Nettoyage (`cleaned_text`)** :
   — Conversion en minuscules (lowercase) pour réduire la dimensionnalité.
   — Suppression des caractères spéciaux non alphanumériques (sauf espaces).
   — Normalisation des espaces multiples.
2. **Lemmatization (`lemmatized_text`)** :
   — Utilisation de la librairie **spaCy** (modèle `en_core_web_sm`).
   — Transformation des mots en leur forme canonique (ex : "running" → "run", "better" → "good").
3. **Enrichissement du Dataset** :
   — Au lieu de créer des matrices séparées, nous avons ajouté les colonnes `cleaned_text` et `lemmatized_text` directement au fichier original.
   — **Sortie** : `data_preprocessed.csv`.

3 Analyse Statistique du Corpus

L'analyse réalisée 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%**, ce qui densifie l'information sans perte majeure de sens.

3.2 Distribution des Entités (NER)

L'analyse des tags `ner_tags` montre un déséquilibre de classe classique :
— Le tag **O (Other)** est ultra-majoritaire (tokens non-entités).
— Les entités **B-PER**, **B-LOC**, **B-ORG** sont présentes mais minoritaires.
— **Conséquence** : Les modèles devront gérer ce déséquilibre (ex : via des fonctions de perte pondérées).

4 Analyse Statistique Détaillée

Cette section présente les visualisations générées à partir de l'analyse du corpus.

4.1 Distribution des Longueurs

La figure 1 montre la distribution de la longueur des phrases (en nombre de mots). On observe une moyenne d'environ 24 mots, avec une variabilité importante typique des textes encyclopédiques.

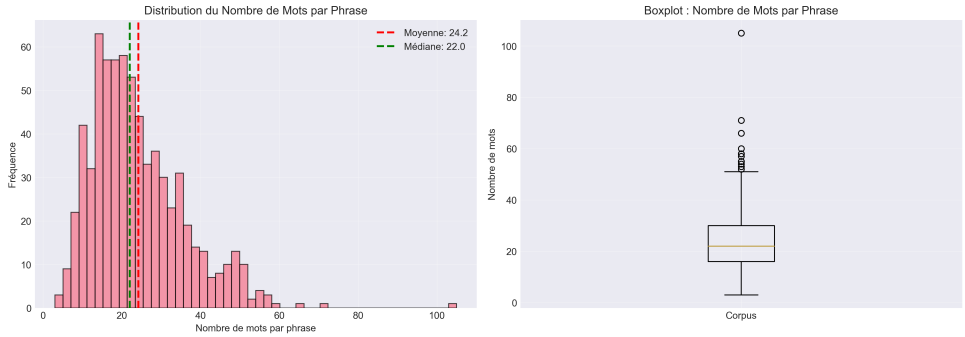

FIGURE 1 – Distribution des longueurs de phrases et boxplot

4.2 Impact du Preprocessing

Le preprocessing (nettoyage et lemmatization) permet une réduction significative de la taille du vocabulaire et de la longueur des séquences (en caractères), comme illustré en figure 2.

4.3 Analyse du Vocabulaire

Le nuage de mots (Figure 3) et le top 20 des mots fréquents (Figure 4) mettent en évidence les thématiques du corpus (histoire, géographie, biographie) et l'effet de la lemmatization qui regroupe les termes (ex : "was/is" → "be").

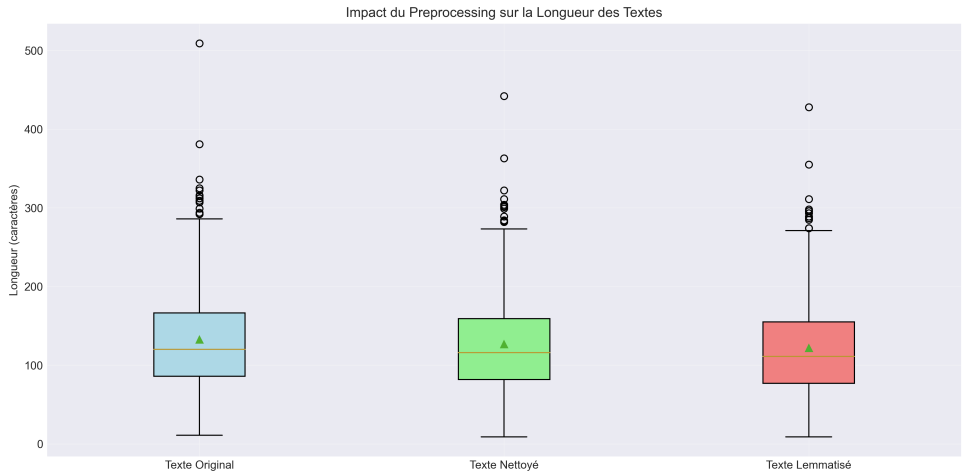

FIGURE 2 – Impact du preprocessing sur la longueur des textes

Métrique	Valeur
Taille vocabulaire original	4990
Taille vocabulaire lemmatisé	4280
Réduction de vocabulaire	**14.2%**
Réduction caractères (Cleaning)	4.6%
Réduction caractères (Lemma)	3.7%

TABLE 1 – Statistiques de réduction (sur échantillon)

4.4 Distribution des Entités Nommées (NER)

La figure 5 confirme le fort déséquilibre des classes. Le tag "O" est prédominant. Parmi les entités, les personnes (PER) et les lieux (LOC) sont les plus fréquents.

5 Analyse Critique des Choix

5.1 Choix 1 : Lemmatization vs Stemming

— **Choix** : Lemmatization (spaCy).

— **Avantages** : Produit des mots réels (formes dictionnaire), préserve mieux le sens sémantique pour l'extraction de relations (Partie C). Plus précis que le stemming (qui tronque brutalement).

— **Inconvénients** : Plus coûteux en temps de calcul. Peut perdre certaines nuances flexionnelles (temps des verbes) utiles pour la temporalité.

5.2 Choix 2 : Lowercasing (Mise en minuscule)

— **Choix** : Tout mettre en minuscule dans `cleaned_text`.

— **Avantages** : Réduit drastiquement la taille du vocabulaire. Associe "Apple" (fruit) et "apple" (fruit).

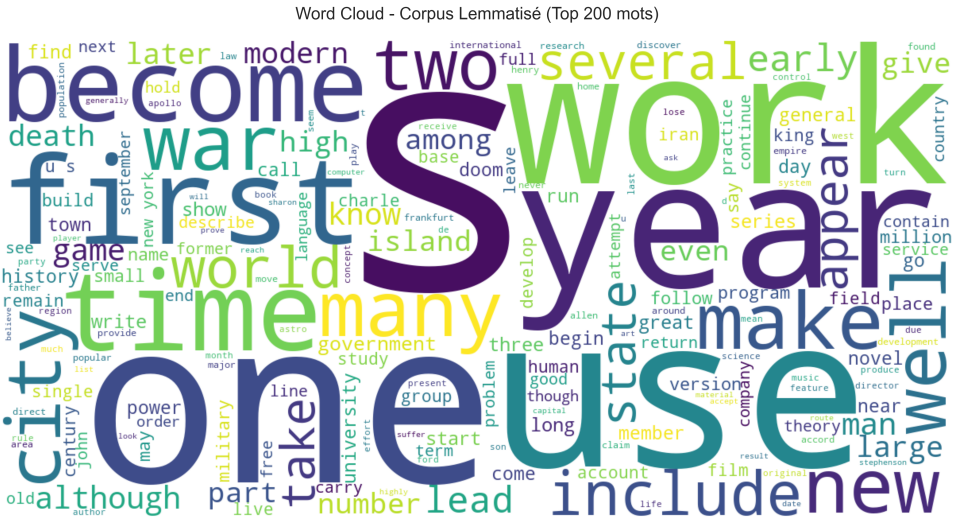

FIGURE 3 – Nuage de mots du corpus lemmatisé

FIGURE 4 – Top 20 des mots les plus fréquents

— **Inconvénients (Critique pour le NER)** : **Perte de l'information de capitalisation**, qui est un indice crucial pour détecter les entités nommées (ex : "Apple" l'entreprise vs "apple" le fruit).

— **Mitigation** : Nous avons conservé la colonne `text` originale. Les modèles NER modernes (BERT, etc.) ou les features manuelles peuvent utiliser le texte brut pour récupérer la casse.

5.3 Choix 3 : Conservation de la Structure Séquentielle

— **Choix** : Ne pas vectoriser en TF-IDF (Bag of Words) pour l'export final.

— **Avantages** : Indispensable pour le NER et l'extraction de relations qui dépendent de l'ordre des mots et du contexte local. Permet l'utilisation de modèles de Deep Learning (RNN, Transformers).

— **Inconvénients** : Fichiers texte plus volumineux que des matrices creuses.

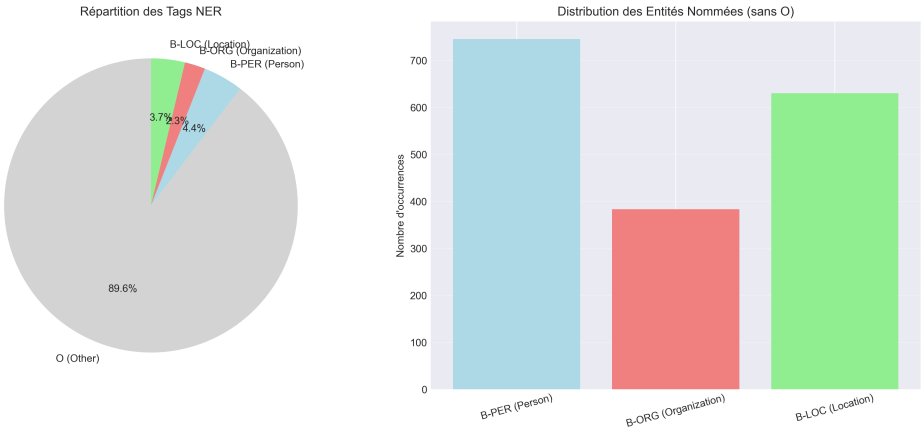

FIGURE 5 – Distribution des tags NER

6 Impact sur la Suite du Projet

6.1 Impact sur la Partie II (NER)

Le preprocessing fournit une base propre (`lemmatized_text`) qui généralise mieux. Cependant, pour maximiser les scores F1, il sera probablement nécessaire d'utiliser :

1. Le **texte original** pour les features de capitalisation.
2. Le **texte lemmatisé** pour les embeddings sémantiques (Word2Vec/GloVe) afin de regrouper les variantes d'un même mot.

6.2 Impact sur la Partie C (Extraction de Relations)

L'extraction de relations (triplets Sujet-Verbe-Objet) bénéficie grandement de la lemmatization :

— Elle normalise les verbes (ex : "est né", "naquit" → "naître"), simplifiant la détection de patterns de relations.

— Elle réduit la complexité des arbres de dépendance syntaxique.

7 Conclusion

Le preprocessing réalisé est un compromis entre **réduction de bruit** (nettoyage, lemmatization) et **préservation de l'information** (conservation du texte brut et de la séquence). Ce socle de données enrichi (`data_preprocessed.csv`) est robuste et adapté aux défis des parties suivantes : la reconnaissance fine d'entités et l'extraction complexe de relations sémantiques.