ADRIEN VERGNE 21210014

Rapport

Ce rapport présente une analyse linguistique menée sur des textes de DAUDET et MAUPASSANT, en utilisant différentes méthodes de tokenisation et d'analyse morphosyntaxique. L'objectif est d'observer les différences entre la tokenisation native Python (split()) et la tokenisation avancée de spaCy, en examinant la loi de Zipf, la distribution des longueurs de tokens, les lemmes les plus fréquents et les catégories grammaticales les plus courantes.

Méthodologie

Deux corpus ont été traités :

DAUDET : "Le Petit Chose"

MAUPASSANT : "Une Vie"

Les analyses ont été effectuées sur deux versions de chaque texte :

Une version de référence (_ref)

Une version issue de l'OCR (Kraken-base)

Les analyses suivantes ont été réalisées :

1. **Tokenisation** avec split() et spaCy
2. **Lemmatisation** des formes lexicales
3. **Analyse morphosyntaxique** (POS tagging)
4. **Distribution des longueurs de tokens**
5. **Vérification de la loi de Zipf**
6. **Comparaison du nombre de caractères et de tokens entre les versions**

Résultats et Analyses

Loi de Zipf

Les courbes de la loi de Zipf montrent une distribution caractéristique où la fréquence des mots suit une décroissance exponentielle en fonction du rang.

DAUDET : La tokenisation spaCy produit une courbe plus lissée et proche de la loi de Zipf idéale, tandis que split() génère plus de fluctuations, notamment dans les mots les plus rares.

MAUPASSANT : Même observation, avec spaCy qui offre une meilleure capture des formes lexicales complexes.

La comparaison entre les versions _ref et Kraken-base indique que l'OCR introduit peu de bruit sur les mots les plus fréquents, mais plus de variations sur les mots peu fréquents.

Distribution des longueurs de tokens

Les histogrammes montrent une distribution classique :

La majorité des tokens ont une longueur de **2 à 6 caractères**, avec un pic autour de **3-4 caractères**.

Les versions _ref et Kraken-base suivent des distributions similaires, bien que Kraken-base contienne un léger bruit dû aux erreurs OCR.

Dans les deux corpus, spaCy segmente parfois mieux les contractions et les mots composés, produisant une répartition plus homogène.

Analyse des lemmes les plus fréquents

Les 10 lemmes les plus fréquents sont :

DAUDET (Kraken-base) : le, de, avoir, un, je, et, ce, que, il.

DAUDET (ref) : le, de, un, je, être, à, et, ce, que.

MAUPASSANT (Kraken-base) : le, de, un, et, lui, avoir, son, se, il.

MAUPASSANT (ref) : le, de, un, et, lui, son, à, se, il.

On remarque des différences minimes entre _ref et Kraken-base, montrant que l'OCR conserve globalement la structure lexicale des textes.

Analyse morphosyntaxique (POS Tagging)

Les classes grammaticales les plus fréquentes sont :

DAUDET (Kraken-base) : NOUN (18 000), DET (12 000), PRON (11 500), VERB (11 000), ADP (10 500).

DAUDET (ref) : NOUN (17 500), ADP (12 500), PRON (12 000), DET (11 500), VERB (11 000).

MAUPASSANT (Kraken-base) : NOUN (16 000), SPACE (12 000), VERB (10 500), DET (9 500), PRON (9 500).

MAUPASSANT (ref) : NOUN (15 800), DET (11 000), ADP (10 500), VERB (10 200), PRON (9 800).

L'analyse POS montre une cohérence générale entre _ref et Kraken-base, bien que la version OCR contienne des légères variations sur les adjectifs et les espaces mal détectés.

Nombre de caractères et de tokens

DAUDET : ref contient environ 490 000 caractères, Kraken-base 485 000 caractères.

MAUPASSANT : ref 450 000 caractères, Kraken-base 445 000 caractères.

Nombre de tokens :

DAUDET : ref (107 000 tokens), Kraken-base (115 000 tokens).

MAUPASSANT : ref (93 000 tokens), Kraken-base (100 000 tokens).

Les fichiers OCR tendent à produire légèrement plus de tokens, probablement en raison d’une segmentation plus fine par rapport aux versions de référence.

Conclusion

L’analyse a mis en évidence plusieurs points :

La tokenisation spaCy est plus précise que split(), notamment pour la segmentation des contractions et des mots complexes.

L’OCR conserve une structure lexicale et syntaxique stable, bien que de légères différences apparaissent dans les lemmes rares et la segmentation des tokens.

La loi de Zipf est bien respectée, validant la cohérence des données linguistiques.

Les POS les plus fréquents sont stables, confirmant une bonne analyse morphosyntaxique malgré des variations mineures dues à l’OCR.