LETTRES
SORBONNE
UNIVERSITÉ

Extraction d'Entités Nommées au format BIO avec spaCy dans différentes versions du texte

Weronika BIEGOWSKA

numéro d'étudiant: 21211975

Projet S2

Enseignement : Programmation de Modèles Linguistiques (II)

Enseignante : Caroline Koudoro-Parfait

Date : 30/03/2025

Table des matières

I. Introduction

A. Context & Objectif

La reconnaissance des entités nommées (REN) est un des tâches fondamentales dans le traitement automatique du langage (TAL). Ainsi, la Commission Nationale de l'Informatique et des Libertés la définit comme « sous-tâche d'extraction d'informations qui cherche à localiser et classifier les mentions d'entités nommées dans du texte non structuré en catégories prédéfinies »[1]. Il s'agit donc d'extraire les sous-chaînes d'un texte qui nomment des objets du monde réel et à déterminer leur type (par exemple, s'il s'agit de personnes (PER) ou d'organisations(ORG) etc[2].

L'objectif de ce projet est d'évaluer la performance du NER sur différentes versions d'un texte donné, y compris les changements introduits par la reconnaissance optique de caractères (OCR). L'OCR introduit souvent des erreurs qui peuvent affecter l'extraction des caractéristiques. Il est donc crucial d'évaluer la robustesse du modèle spaCy NER lorsqu'il est appliqué à de telles données bruitées.

B. Corpus & Data

J'ai reçu un corpus composé de 15 documents : 3 versions différentes du même texte pour chaque auteur:

Tab.1: Corpus

Auteur	Nom du fichier de référence	Nom du fichier de OCR: Kraken	Nom du fichier de OCR: Tesseract
DASH	DASH_chateau-de-Pinon-V1_PP	DASH_chateau-de-Pinon-V1_Kraken-base	DASH_chateau-de-Pinon-V1_TesseractFra-PNG
DAUDET	DAUDET_petit-chose_PP	DAUDET_petit-chose_Kraken-base	DAUDET_petit-chose_TesseractFra-PNG
FLAUBERT	FLAUBERT_education-sentimentale_PP	FLAUBERT_education-sentimentale_Kraken-base	FLAUBERT_education-sentimentale_TesseractFra-PNG
MAUPASSANT	MAUPASSANT_une-vie_PP	MAUPASSANT_une-vie_Kraken-base	MAUPASSANT_une-vie_TesseractFra-PNG
NOAILLES	NOAILLES_la-nouvelle-esperance_PP	NOAILLES_la-nouvelle-esperance_Kraken-base	NOAILLES_la-nouvelle-esperance_TesseractFra-PNG

1 Commission nationale de l'informatique et des libertés. (s.d.). Reconnaissance d'entités nommées. CNIL. Consulté le 30 mars 2025, à l'adresse https://cnil.fr/fr/definition/reconnaissance-dentites-nommees.

[2]Keraghel, I., Morbieu, S., & Nadif, M. (2024). Recent advances in named entity recognition: A comprehensive survey and comparative study. *arXiv*. https://arxiv.org/pdf/2401.10825

Il peut être observé que outre la version de référence, il existe pour chaque texte deux options réalisées à l'aide de l'OCR. Ainsi, la reconnaissance optique de caractères est une technologie qui convertit le texte imprimé ou manuscrit d'images numérisées en texte lisible par une machine. Tesseract est un moteur OCR open-source développé par Google, connu pour sa flexibilité et son large support linguistique, ce qui le rend adapté aux tâches OCR générales. Cependant, sa précision peut être affectée par la qualité de l'image et l'alignement du texte.[3] Kraken, un dérivé d'OCRopus, est particulièrement efficace pour les documents historiques et les scripts complexes, offrant une grande précision pour les cas d'utilisation spécialisés comme les textes anciens, grâce à son approche basée sur l'apprentissage en profondeur.[4]

II. Méthodologie

Cette section présente une méthodologie choisi pour ce projet. L'objectif principal du projet est d'identifier et classifier les entités (telles que les personnes, les organisations et les lieux). Pour ce faire, on utilisera spaCy pour la reconnaissance d'entités, en traitant des versions de texte au format .txt en utilisant:

1) Le **format BIO** (Beginning, Inside, Outside), un schéma d'annotation standard, couramment utilisé pour annoter les entités dans le traitement du langage naturel, qui structure les résultats de la reconnaissance des entités de manière claire et systématique.

2) Ensuite, on va effectuer la reconnaissance d'entités et le **POS tagging:** l'étiquetage de la partie du discours avec l'aide de l'outil de morphologie syntaxique ou **Part-of-speech tagging**, en se focalisant sur la récupération des tokens dont le label est "PROPN" (Proper Noun)

On utilise **spaCy**, une bibliothèque NLP populaire, choisie en raison de son efficacité et de la disponibilité de modèles pré-entraînés pour diverses langues. Le processus commence par la préparation du texte, comprenant la tokenisation, la lemmatisation et la gestion des erreurs potentielles liées à la reconnaissance optique de caractères (OCR).

Les modèles ont ensuite été appliqués sur les différentes versions de textes pour extraire les entités et comparer les performances entre les versions. On va comparer les sorties de la référence et des OCR et examiner les intersections entre les noms propres et les entités reconnues.

---

[3] Smith, R. (2007). An Overview of the Tesseract OCR Engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition, 629–633.
Retrieved from https://research.aimultiple.com.

[4] Rühling, A. (2018). Kraken: A Deep Learning-Based OCR Engine for Historical and Complex Scripts. Academia. consulté depuis : https://www.academia.edu.

Grâce à diverses visualisations, notamment des graphiques de distribution des entités et des diagrammes de Venn, on va mettre en évidence les divergences entre les versions du texte et à analyser les performances de la NER dans les données traitées par l'OCR.

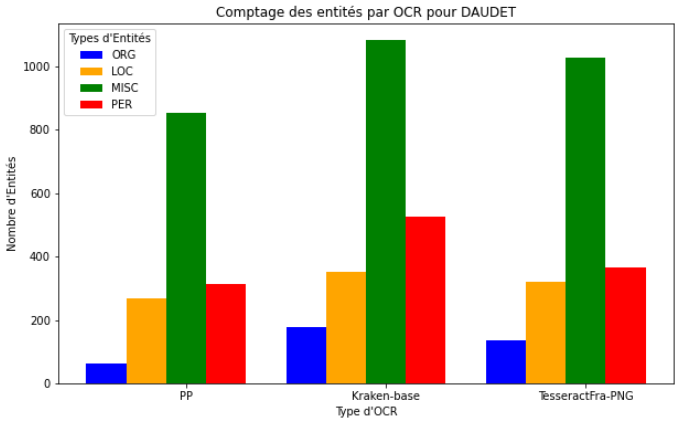
Les libraries principales pour effectuer ce projet incluent:

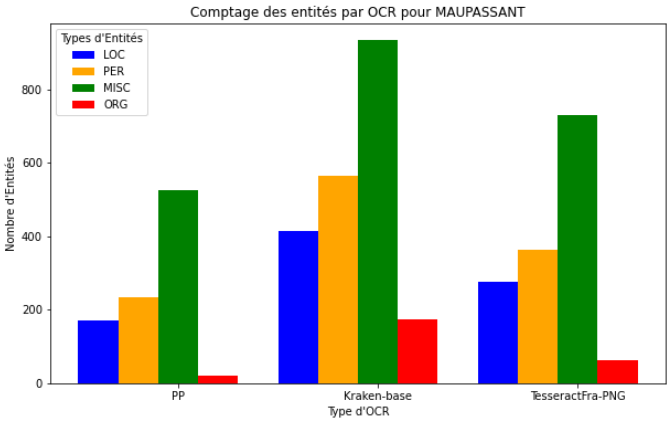
- **SpaCy**: Bibliothèque de traitement du langage naturel;
- **Glob**: pour récupérer plusieurs fichiers correspondant à un motif;
- **Json** : pour la gestion des fichiers JSON;
- **Re** : pour la manipulation des expressions régulières;
- **Os**: pour gérer les fichiers et dossiers;
- **Csv**: pour l'écriture et la lecture de fichiers CSV
- **Matplotlib**
- **Seaborn**
- **Pandas**

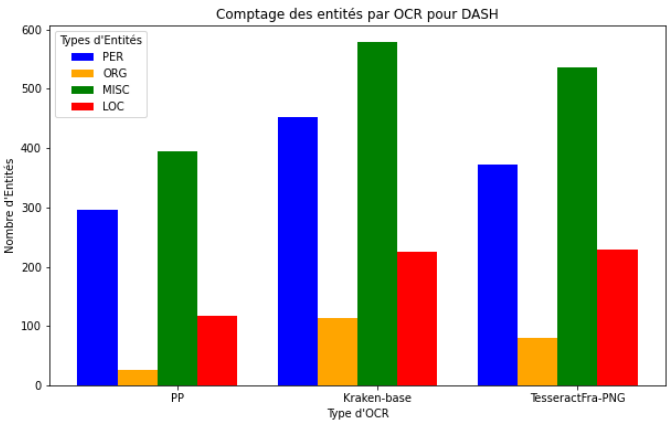
III. Résultats

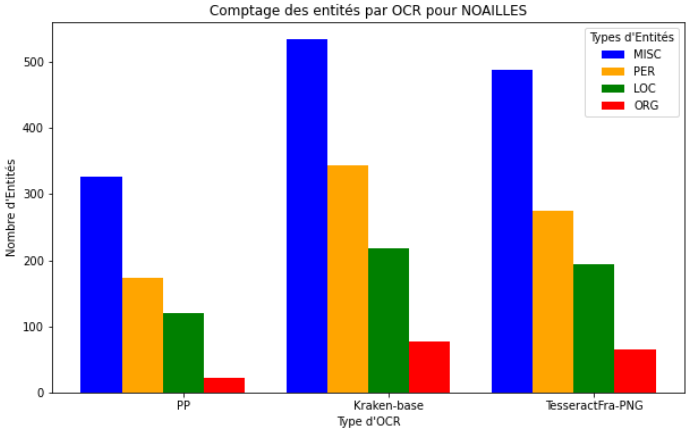
A. BIO: Comparaison des sorties de la référence et des OCR

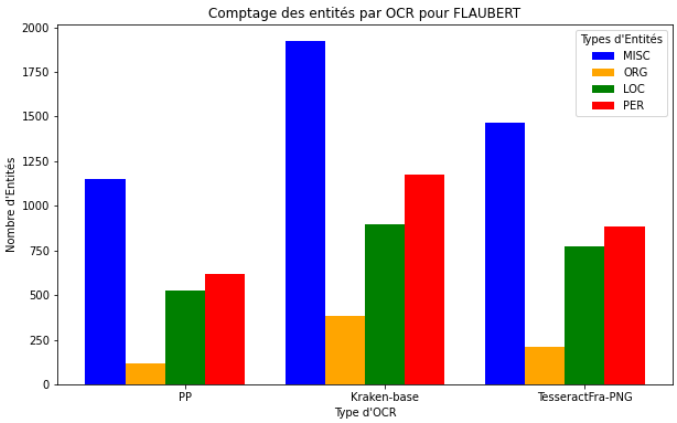
1) La proportion d'entités pour chaque label sémantique selon les différentes versions des textes.





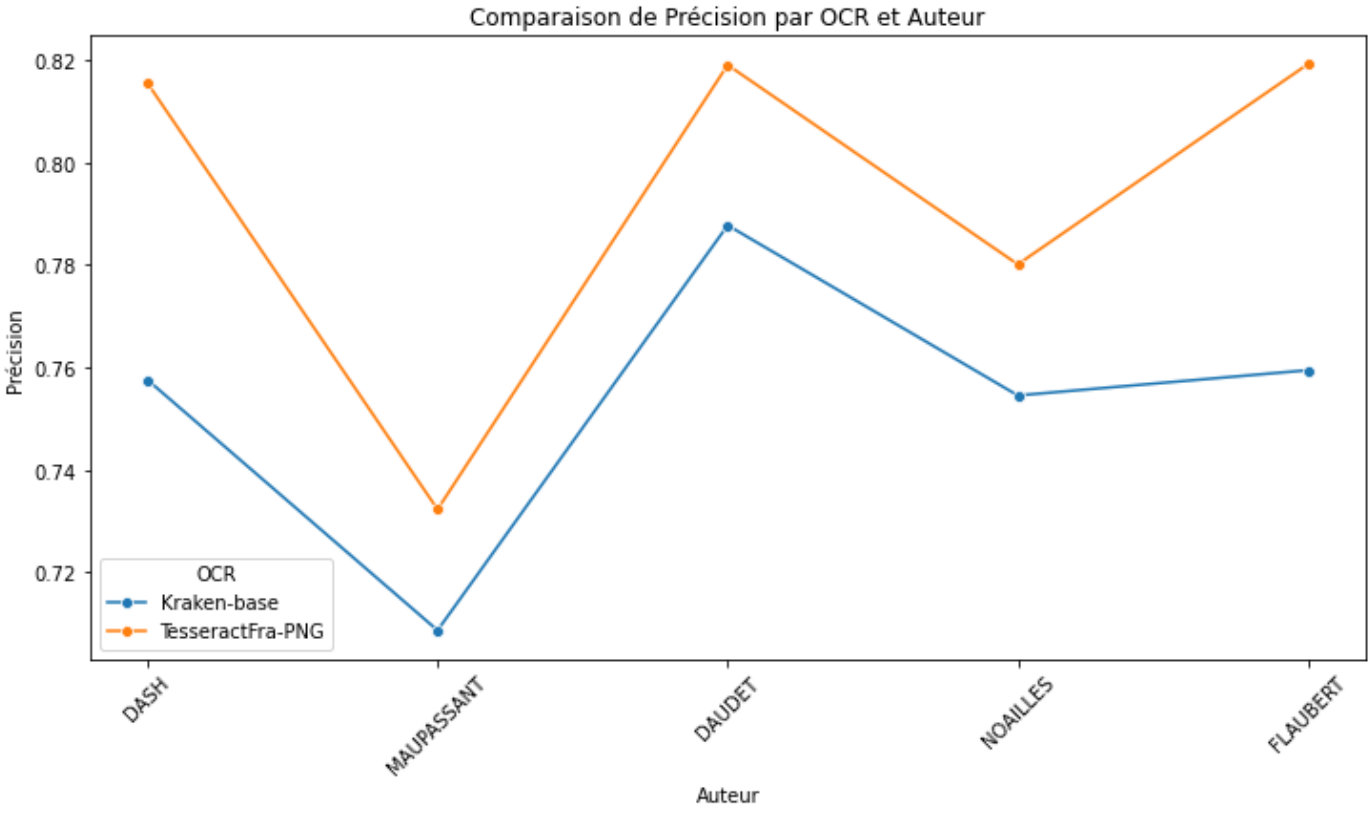


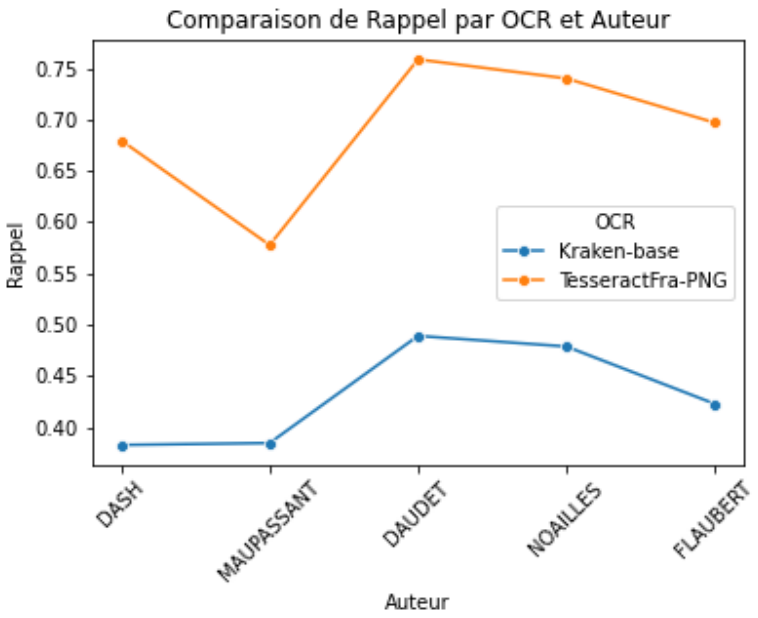


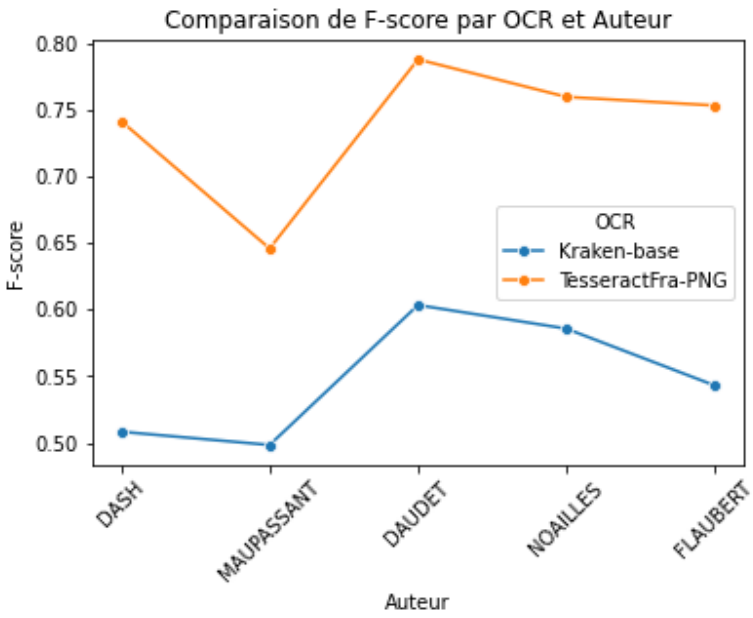


On observe que globalement plus des entités nommées ont été retrouvées dans les textes OCR que dans la version de référence. Le label sémantique qui est le plus est indubitablement « MISC » dans tous les textes et toutes les versions.

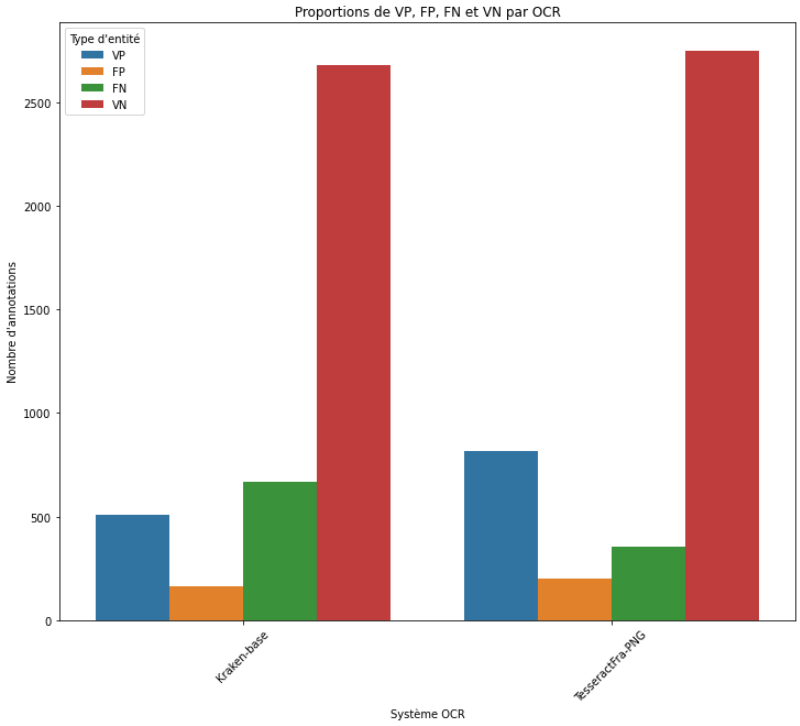
2) Précision, le rappel et le f-score.







3) Les proportions de VP, FP, VN, FN dans la REN



<u>Analyse de résultats:</u>

Comparaison des scores F

Globalement, Tesseract-Fra-PNG surpasse systématiquement Kraken-base chez tous les auteurs. Les deux systèmes d'OCR présentent des performances similaires pour tous les auteurs. Les scores F les plus élevés pour les deux systèmes sont obtenus avec Daudet et les plus bas sont obtenus avec Maupassant. L'écart de performance entre les systèmes est relativement constant (environ 0,2-0,25 points de différence).

Comparaison du rappel

Comme pour le score F, le rappel de Tesseract-Fra-PNG est supérieur à celui de Kraken-base. Les deux systèmes ont le meilleur rappel avec Daudet et le plus faible avec Maupassant. Le modèle de performance entre les auteurs est cohérent entre les deux systèmes.

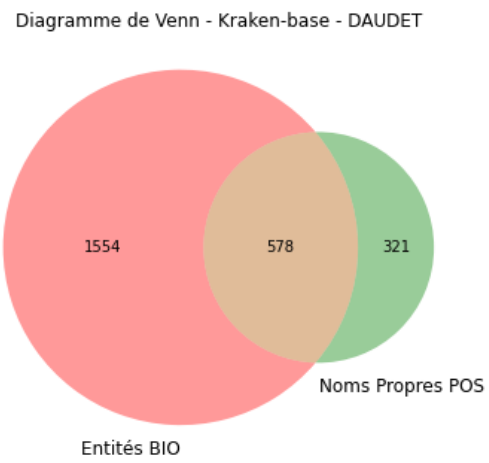
Comparaison de la précision

Tesseract-Fra-PNG maintient une précision plus élevée que Kraken-base pour tous les auteurs. Les deux systèmes montrent une plus grande variation de la précision entre les auteurs que le F-score et le rappel. Flaubert et Daudet présentent la précision la plus élevée pour Tesseract-Fra-PNG tandis que Maupassant présente la précision la plus faible pour les deux systèmes. Daudet a la meilleure précision pour Kraken-base

Distribution des types d'entités

Le diagramme à barres montre le nombre de différents types d'entités (VP, FP, FN, VN) reconnus par chaque système d'OCR : VN (Vrais Négatifs) est la catégorie dominante pour les deux systèmes. Tesseract-Fra-PNG identifie plus de VP (Vrais Positifs) que Kraken-base. Kraken-base produit plus de FN (Faux Négatifs). Le nombre de FP (Faux Positifs) est relativement similaire d'un système à l'autre.

B. Les intersections entre les sorties de REN et les sorties de Part-of-speech tagging ”PROPN” avec des diagrammes de Venn.



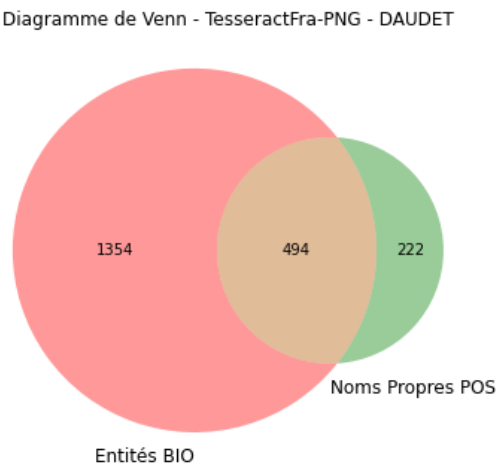


Diagramme de Venn - Kraken-base - MAUPASSANT

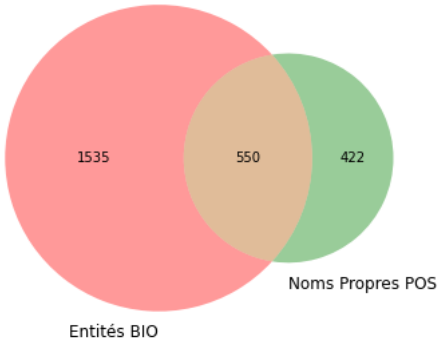


Diagramme de Venn - TesseractFra-PNG - MAUPASSANT

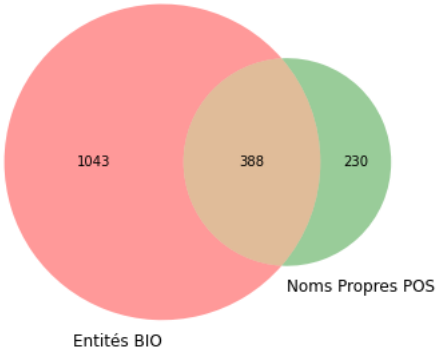


Diagramme de Venn - Kraken-base - DASH

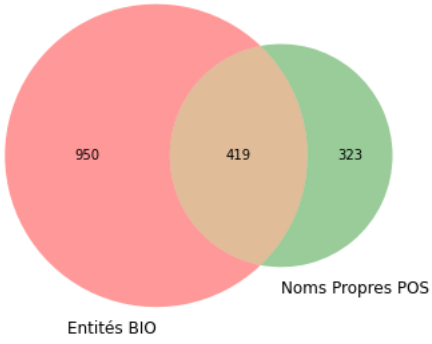


Diagramme de Venn - TesseractFra-PNG - DASH

Diagramme de Venn - Kraken-base - NOAILLES

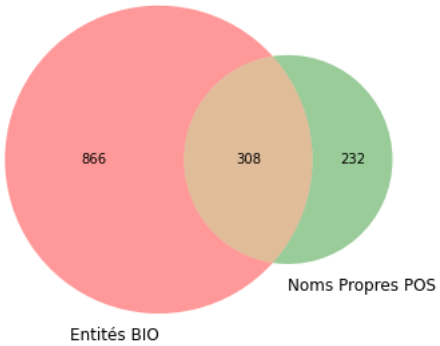
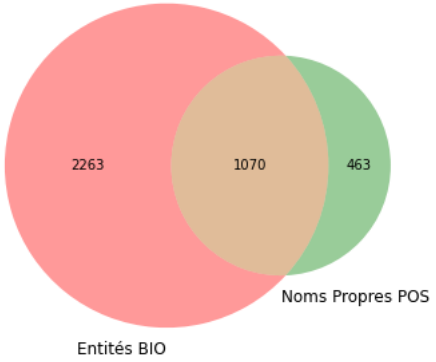


Diagramme de Venn - TesseractFra-PNG - NOAILLES

Diagramme de Venn - Kraken-base - FLAUBERT

Diagramme de Venn - TesseractFra-PNG - FLAUBERT



Ces diagrammes de Venn fournissent une comparaison détaillée entre deux systèmes OCR (TesseractFra-PNG et Kraken-base) pour cinq auteurs (Flaubert, Noailles, Dash, Maupassant et Daudet), montrant la relation entre les entités BIO (reconnaissance des entités nommées) et l'étiquetage POS (part-of-speech) pour les noms propres.

Analyse de résultats:

Comparaison par système OCR: TesseractFra-PNG vs. Kraken-base

Kraken-base identifie systématiquement plus d'entités pour l'ensemble des auteurs. Il montre des comptes plus élevés dans les trois domaines : Entités BIO uniquement, noms propres POS uniquement, et intersection entre les deux. Cependant, l'augmentation du nombre d'entités avec Kraken-base n'indique pas nécessairement une meilleure précision (pour rappel, les graphiques précédents montraient que TesseractFra-PNG avait une meilleure précision).

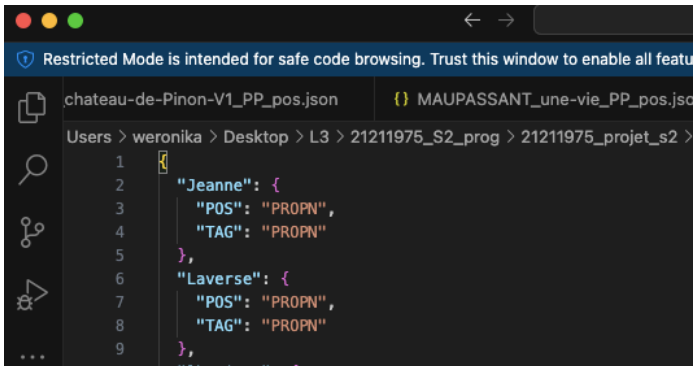
Différences méthodologiques : Le chevauchement limité entre les entités BIO et les noms propres POS suggère que ces approches sont complémentaires plutôt que redondantes.

Cohérence : Les proportions relatives entre les catégories restent relativement cohérentes d'un auteur à l'autre, ce qui suggère que les méthodologies de marquage sont stables dans les différents styles d'écriture.

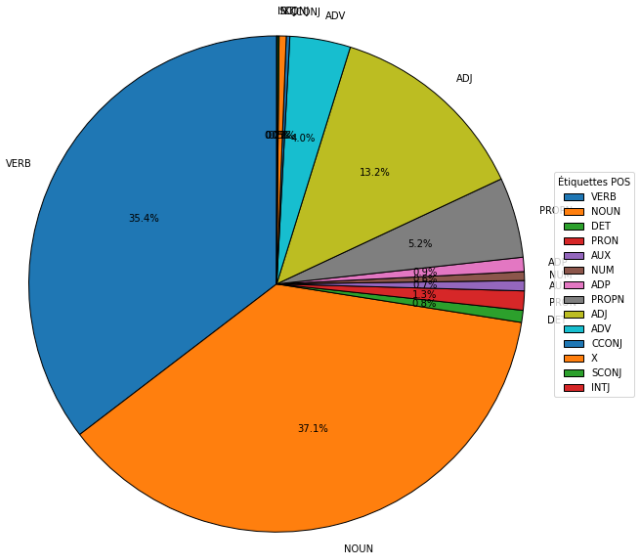
Signification statistique : Les différences entre les systèmes d'OCR semblent suffisamment importantes pour avoir un impact sur les applications NLP en aval, en particulier pour des auteurs comme Flaubert et Maupassant.

C. La proportion de verbe, d'adjectif, de nom commun etc. qui ont été annotés comme des Entités nommées.

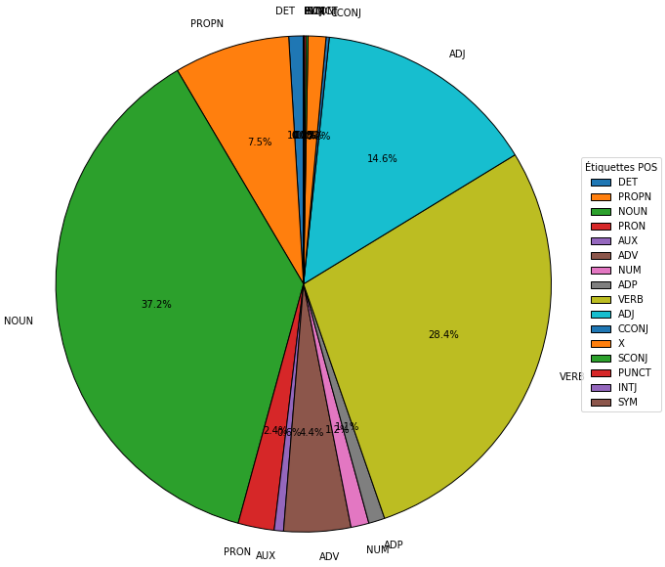
Malheureusement, après avoir essayé tous les modèles de spal pour le français dans tous les dictionnaires, l'étiquette POS ne différait pas de celle de TAG, et je n'ai donc pas pu faire ce graphique : il m'aurait montré un diagramme camembert avec un seul type : NPROP. C'est pourquoi j'ai fait des graphiques qui montrent la catégorie POS de tous les tokens et pas seulement des noms propres à la place. Comme ils sont nombreux et qu'ils prendraient plusieurs pages, je ne joins que des exemples, tout le reste se trouve dans le dossier :



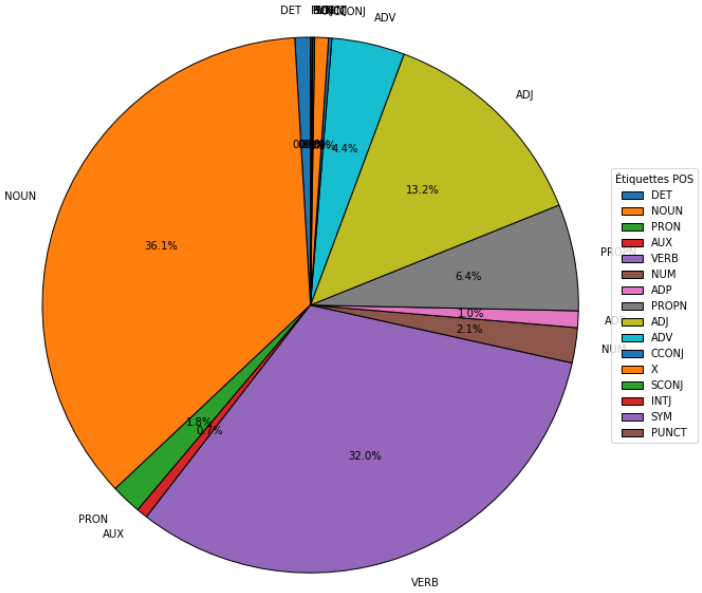
Proportion des étiquettes POS dans le DAUDET_petit-chose_PP_pos



Proportion des étiquettes POS dans le DAUDET_petit-chose_Kraken-base_pos



Proportion des étiquettes POS dans le DAUDET_petit-chose_TesseractFra-PNG_pos



Dans le cas de Daudet, on observe que les etiquettes les plus fréquentes sont dans tous les versions: les noms. Leur pourcentage varie en fonction de version: 37.1% pour la version de référence, 37.2% pour Kraken-base et 32% pour Tesseract. En outre, les noms propres (PROPN) constituent 5.2% de la version de référence, 7.5% pour Kraken-base et 6.4% pour Tesseract. Ici on observe que la difference entre la référence et OCR est assez considerable dans les deux cas.

IV. Conclusion et perspectives

Cette étude a exploré l'extraction d'entités nommées au format BIO avec **spaCy** sur différentes versions d'un texte, y compris des versions OCR. Les résultats ont montré que la qualité du texte influence fortement la précision du REN, les erreurs OCR perturbant la détection des entités. L'analyse via **POS tagging** a mis en évidence certaines confusions entre entités nommées et noms propres.

Pour améliorer les performances, l'utilisation de **modèles entraînés sur des textes OCR bruyés**, ainsi que des **post-traitements linguistiques ou automatiques**, pourrait réduire ces erreurs. Une comparaison avec des **modèles transformers** comme BERT pourrait également offrir des résultats plus précis. Ce travail souligne les défis du NER sur des textes de qualité variable et ouvre des perspectives pour des approches plus robustes.

V. Bibliographie

1 Commission nationale de l'informatique et des libertés. (s.d.). Reconnaissance d'entités nommées. CNIL. Consulté le 30 mars 2025, à l'adresse https://cnil.fr/fr/definition/reconnaissance-dentites-nommees.

2 Keraghel, I., Morbieu, S., & Nadif, M. (2024). Recent advances in named entity recognition: A comprehensive survey and comparative study. *arXiv*. https://arxiv.org/pdf/2401.10825

3 Smith, R. (2007). An Overview of the Tesseract OCR Engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition, 629–633. Consulté depuis https://research.aimultiple.com.

4 Rühling, A. (2018). Kraken: A Deep Learning-Based OCR Engine for Historical and Complex Scripts. Academia. consulté depuis : https://www.academia.edu.