Programmation de Modèles Linguistiques (I)

(L5SOPROG L3 Sciences du Langage)

2024-2025

TD8 Évaluer des outils de reconnaissance d'entités nommées (REN)

Caroline Koudoro-Parfait, Sorbonne Université.

Objectifs

- Programmer des fonctions
- Utiliser CounVectorizer et des mesures de distance
- stocker les résultats
- représenter les résultats par un graphique

Exercice 1 : Lire les fichiers json dans un dossier

En utilisant la fonction qui permet de lire des fichiers json vue dans les TDs précédents, vous lirez dans le dossier 'ressourcesTD8" les fichiers .json qui comprennent les dictionnaires stockant les entités nommées récupérées par spaCy avec les modèles *fr_core_news_sm* et *fr_core_news_lg* sur les trois versions des textes – sans effacer les fichiers .txt et sans bouger les fichiers .json :

– La référence : REF
et OCR (reconnaissance optique de caractères)
– La version transcrite avec Kraken
– La version transcrite avec Tesseract-Fra

Exercice 2 : Écrire une fonction pour calculer les distances

installer scikit-learn si ce n'est pas déjà le cas:

- `https://pypi.org/project/scikit-learn/`
- `https://scikit-learn.org/stable/install.html`

documentation de *scikit-learn* : `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.DistanceMetric.html`

A partir du TD7, vous écrirez une fonction qui utilise CountVectorizer et scikit learn pour calculer différente distances entre les 3 versions d'un même texte.

- Vous utiliserez différentes métriques
- vous utiliserez différents paramètres

Exercice 3 : Stocker les résultats

Vous élaborerez ce programme pour stocker les résultats dans un dictionnaire dans un format de fichier qui vous semble pertinent :

- le nom des textes comparés,
- les noms des métriques utilisées,
- les résultats correspondants, et non pas les tableaux de résultats.

```
{"Nom_fichier-ref -- Nom_fichier-hyp" :{
      "jaccard": [
        0.6041933418693982
      ],
      "braycurtis": [
        0.2020727818420559
      ],
      "dice": [
        0.4328632037610366
      ],
      "cosinus": [
        0.026162890836085473
      ]
    }
}
```

Exercice 4 : Écrire une fonction pour représenter les résultats

- Vous représenterez les résultats obtenus sous forme de graphiques. La visualisation des résultats doit permettre de pouvoir identifier quel est le meilleur OCR selon les résultats des métriques utilisées. Pour ce faire vous pouvez utiliser les bibliothèques :
  - Seaborn (`https://seaborn.pydata.org/`)
  - Matplotlib (`https://matplotlib.org/`)

Attention les deux bibliothèques ne prennent pas la même structure de données en entrée !

Vous écrirez quelques analyses des résultats obtenus par comparaison automatiques.

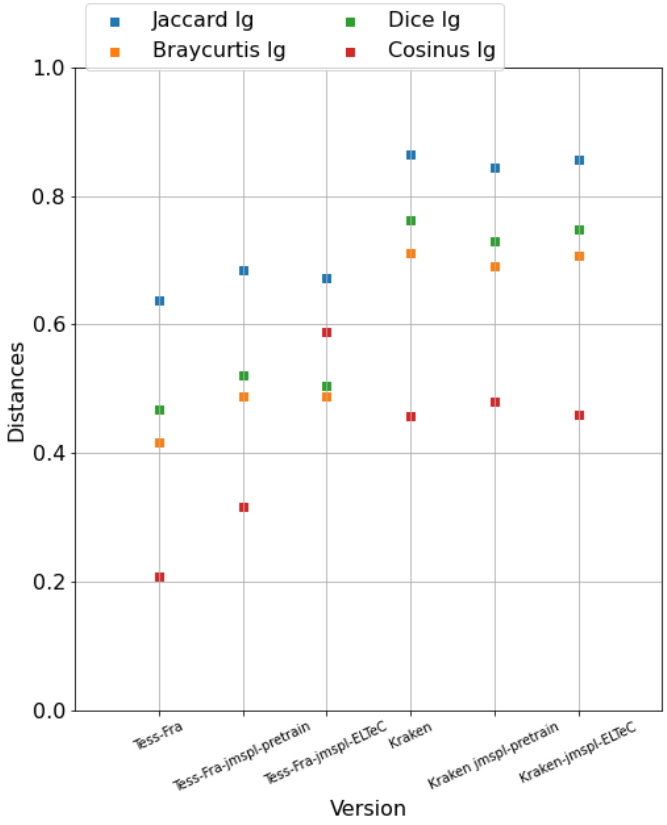

Figure 1: MAUPASSANT, spaCy3.5.1-lg

Devoir

- Utiliser des fonctions déjà vues en TD
- Écrire une fonction à partir d'un programme vu dans le TD7
- Produire des graphiques
- Produire une analyse des résultats

  Vous déposerez sur Moodle une archive zip nommée NUMETU.zip (où NUMETU est votre numéro d'étudiant) et contenant :

- Votre code exporté au format Python
- le PDF du document que vous avez produit

  Date limite : indiquée sur le Moodle !