TD3 : analyse morpho-syntaxique

Entités Nommées

LO17

Fatma Chamekh

2021-2022

Objectifs :

L'objectif de ce TD est d'explorer les principales fonctions de la librairie SPACY pour réaliser une analyse morpho-syntaxique en mettant en œuvre la reconnaissance des entités nommées. Il comporte deux parties :

- Partie 1 : analyser le corpus en mettant en place la chaine de traitement morpho-syntaxique en ajoutant la reconnaissance des entités nommées.
- Partie 2 : Comparer les résultats de reconnaissances d'entités nommées SPACY vs HeidelTime https://heideltime.ifi.uni-heidelberg.de/heideltime/.

Exercice 1 :

L'objectif de cet exercice est d'analyser les termes les plus fréquents dans les noms de produits de l'openfood database (https://fr.openfoodfacts.org/data). l'OpenFood Database, une base de données alimentaire qui est enrichie de manière collaborative.

Au passage, cela permet de réviser les étapes d'analyse textuelle et d'explorer les enjeux de reconnaissance d'entités nommées :

- Tokenisation
- Retrait des stop words
- Dépendance entre les mots
- Lemmatisation
- Stemming
- Nuage de mots (il faut utiliser wordcloud)
- Reconnaissance d'entités nommées

Ecrivez un programme Python qui permet de :

- Importer le modèle de reconnaissance de langage qui sera utilisé par la suite ainsi que le corpus Anglais utilisé par spacy.
- Importer/afficher les données (le fichier des données est disponible sur Moodle).
- Créer une fonction de nettoyage des noms de produits effectuant les étapes suivantes : tokeniser le texte en question et retirer la ponctuation et les stopwords.
- Appliquer cette fonction à l'ensemble des noms de produits (variable product_name)
- Visualiser avec spacy.displacy le résultat d'une reconnaissance d'entités nommées sur 50 données aléatoires. Cela vous semble-t-il satisfaisant ?

Exercice2 :

Vous avez à disposition une archive contenant les données textuelles à manipuler :

- Répertoire CorpusBEA/TXT-TP : 158 compte-rendu des évènements (incidents ou accidents aériens). Les fichiers ont été générés automatiquement à partir des pages Web. Certains éléments typographiques peuvent persister.
- Répertoire CorpusBEA/MEDATA-TP : 158 fichiers tabulés (format CSV/TSV) contenant les métadonnées relatives à l'évènement. Le nom du fichier de métadonnées est le même que celui du fichier texte contenant le compte-rendu de l'évènement. Chaque fichier contient une ligne. Les métadonnées sont séparées par des tabulations. Le fichier CorpusBEA/Metadata.txt contient une description des métadonnées.

En utilisant SPACY, faites une analyse morpho-syntaxique y compris la reconnaissance des entités nommées. Ecrire une fonction qui permet de calculer le nombre d'entité détecté pour chaque catégorie d'entités nommées. Est-ce que votre programme a détecté des entités nommées temporelles et les mentions d'altitude et de vitesse.

Dans un deuxième temps, vous utilisez HeidelTime (https://dbs.ifi.uni-heidelberg.de/research/heideltime/) un système de reconnaissance d'entités nommées multi-langues. Il existe deux façons pour l'utiliser :

- La version en ligne : https://heideltime.ifi.uni-heidelberg.de/heideltime/.
- Pour les plus courageux, l'installer : https://github.com/HeidelTime/heideltime. Dans ce cas, l'extraction des entités nommées peut être réaliser en adaptant le script shell suivant :

```
mkdir CorpusBEA/HDT-TP
for f in CorpusBEA/TXT-TP/*.txt ; do
  echo $f
  fout=CorpusBEA/HDT-TP/`basename ${f/.txt/.hdt}`
  heideltime-standalone -l french $f > $fout
done
```

- Ecrivez une fonction qui permet de calculer le nombre d'entités détectées par HidelTime pour chaque catégorie d'entités nommées. Vous pourrez Comparez les résultats avec ceux de Spacy. Que pouvez-vous conclure ?
- Ecrivez un programme vous permettant d'identifier les altitudes et les vitesses mentionnés dans les documents.