INALCO - Licence LMFA - TNM L3, Algorithmique et structures de données

Caroline Koudoro-Parfait

21 Novembre 2022

Nom de l'étudiant :
n° d'étudiant :

Projet

1 Modalités :

Le projet :

- est individuel,
- sera à terme, constitué :
  - d'un programme python commenté dans ses grandes lignes,
  - d'un fichier PDF de 2 pages minimum répondant aux questions ci-dessous,
- il doit être envoyé par email le lundi 28 novembre 2022, à caroline.parfait@outlook.fr.
- en cas de problèmes pour la remise du fichier en parler avant le 5 décembre 2022.

2 Installation de Spacy sur votre machine

Suivez bien les instructions du site : `https://spacy.io/usage`

- Lire attentivement le site web
- Écrire une note d'une vingtaine de ligne pour résumer ce que permettent les différentes fonctionnalités de l'outil Spacy

3 Stocker vos résultats au format json

Pour stocker les résultats des programmes suivants, vous utiliserez la fonction suivante. Expliquer ce que fait cette fonction de manière précise et quels sont ses paramètres.

```
import json

def stocker( chemin, contenu):

    w =open(chemin, "w")
    w.write(json.dumps(contenu , indent = 2))
    w.close()
    print(chemin)
```

4 Tokeniser

`https://spacy.io/usage/linguistic-features#tokenization`

- Écrire un programme qui permet de tokeniser le(s) textes de votre corpus
- Ajouter à ce programme la possibilité de compter le nombre de token de votre corpus
- Déterminer la fréquence de chaque token dans votre corpus
- Stocker les différents résultats dans un document json

5 Sentence Segmentation

`https://spacy.io/usage/linguistic-features#sbd`

- Écrire un programme qui permet de segmenter le(s) textes de votre corpus en phrase
- Ajouter à ce programme la possibilité de compter le nombre de phrase de votre corpus
- Stocker les différents résultats dans un document json
- Enfin déterminer en observant les sorties, dans votre document json, si la segmentation est faite de la manière suivante : Majuscule pour le premier mot de la phrase et point à la fin de la phrase. Qu'en dites-vous ? rédiger quelques lignes.