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

Caroline Koudoro-Parfait

5 décembre 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,
- points de participation : il doit être envoyé par email le lundi 12 décembre 2022 au plus tard, à caroline.parfait@outlook.fr.

2 Gestion des modèles de langue de Spacy

Spacy à plusieurs modèles de langue selon les langues. Vous pouvez travailler avec les modèles pour le français :

- *fr_core_news_sm*
- *fr_core_news_md*
- *fr_core_news_lg*

2.1 Sentence Segmentation avec SPACY

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

- Segmenter votre corpus avec deux modèles de spacy pour le français.
- donner le nombre de phrases, pour chacun des textes de votre corpus, trouvés avec chacun des deux modèles de spacy. Y-a-t-il une différence ? Vous Donnerez des exemples.

2.2 Tokenisation avec SPACY

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

- Tokeniser votre corpus avec deux modèles de spacy pour le français.
- donner le nombre de token, pour chacun des textes de votre corpus, trouvés avec chacun des deux modèles de spacy. Y-a-t-il une différence ? Vous Donnerez des exemples.

3 Stocker vos résultats au format json

3.1 Type construit de données : Dictionnaire

stocker vos résultats dans un fichier json de la façon suivante pour les phrases. Et en remplaçant ***phrase*** par ***token*** pour les résultats de la tokenisation.

```
 {
"phrase_00": {
  "texte": "Ha-\n"
},
"phrase_01": {
  "texte": "Elle passe, a bon droit, pour la\nplus regardante de Saint-Brunelle;"
},
"phrase_02": {
  "texte": "Il paratt difficile de rencontrer dans les pa-\nlais, savez-vous,"
},
#.... Jusqu'à la fin du texte
}
```

3.2 Stockage : json

Vous stockerez les résultats des programmes suivants en utilisant la fonction suivante. Lorsque vous faites tourner votre programme il doit stocker 1 fichier de sortie pour 1 fichier d'entrée et chacun des fichiers de sortie doit être généré durant le même run.

```python
import json

def stocker( chemin, contenu):

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