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

Caroline Koudoro-Parfait

12 Décembre 2022

Nom de l'étudiant :
n° d'étudiant :

Projet

1 Modalités :

Le projet :

- est individuel,
- sera constitué :
  - d'un programme python commenté dans ses grandes lignes,
  - d'un répertoire DATA et d'un répertoire CODE
  - d'un fichier texte qui comportera votre réponse à la question posée à la fin de l'énoncé.
- Vous devez vous appuyer sur les exercices effectués les semaines précédentes.
- il doit être envoyé par email le lundi 12 décembre 2022 à 19h30 au plus tard, à caroline.parfait@sorbonne-universite.fr.
- en cas de problèmes pour la remise du fichier en parler avant le 12 décembre 2021 19h15.

1.1 Corpus

Pour effectuer ce projet, vous devez disposer d'un corpus en langue française. Il s'agit du corpus préparé dans les semaines précédentes.

1.2 Lecture du Corpus

Écrire une fonction qui permet de lire les fichiers de votre corpus à partir du chemin relatif vers les données.

1.3 Spacy

Ci-dessous les modèles de Spacy pour le français. Choisissez en deux.

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

2 Tokenisation et Segmentation avec Spacy

2.1 Token

- Écrire une fonction qui permet de Tokeniser les textes de votre corpus en français. Stocker les sorties dans une liste.

2.2 Segmentation

- Écrire une fonction qui permet de segmenter les textes de votre corpus. Stocker les sorties dans une liste.

2.3 Tokenisation des phrases

- Écrire un programme qui permet de compter le nombre de tokens par phrase en vous aidant de ce que vous avez fait dans les parties 2.1 et 2.2.
- Votre programme doit permettre de stocker dans un dictionnaire le nombre de tokens par phrase de la manière suivante :

```
{
"Segment_0": {
  "Phrase_0": "Depuis un nombre innombrable d'hivers, c'est dans la maison de Norine.",
  "nombre de token": "11",
  "Liste de tokens" :["Depuis","un",....]
}
},

...
```

Il est attendu que le programme génère un fichier de sortie pour chaque fichier d'entrée de votre corpus, au format .json.

3 Reconnaissance d'entités nommées

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

- Écrire un programme qui permet de récupérer les entités nommées et leur label (PERS : personne ; LOC : localisation ; MISC : Miscelleanous ; ....) sur chacun des textes de votre corpus en utilisant spaCy et qui permet de stocker un dictionnaire au format json.

```
{
"entité_00": {
  "Entité": "Paris",
  "Label": "LOC"
}
},

...
```

- ce programme doit permettre de récupérer les entités nommées sur votre corpus en utilisant deux modèles de langue de spaCy. Par exemple :
  - *fr_core_news_sm*
  - *fr_core_news_lg*
- l'opération peut-être longue, appliquez votre programme sur au moins 1 de vos textes pour générer une sortie. (Vous devez tout de même programmer le fait que le programme soit applicable sur tous les textes de votre corpus)
- Rédiger de manière synthétique quelques observations sur les sorties de reconnaissance d'entités nommées. Pour ce faire vous choisirez l'un des textes de votre corpus et ses sorties pour chacune des deux versions générées avec les deux modèles de langue française de votre choix.