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

Caroline Koudoro-Parfait

6 Février 2023

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 1 à 3 pages répondant aux questions ci-dessous,
- il doit être envoyé par email le 20 février 2023 à 15h30 au plus tard, à caroline.parfait@sorbonne-universite.fr.
- en cas de problèmes pour la remise du fichier en parler avant le 20 février 2023.

2 But du projet

Le but du projet consiste à manipuler les library csv, pandas, seaborn pour l'exploitation et la visualisation des données.

3 Création du jeu de données

Concevoir un programme qui prend en entrée un texte et qui donne en sortie un dictionnaire. Stocker ce fichier au format json.

```
 {
"entité_00": {
  "entité texte": "Paris",
  "label" : "LOC",
  "contexte gauche" : "Je vais à",
  "contexte droite" : "pour voir ma soeur"

}
```

Préparer vos données en déterminant celles qui sont des Vrais Positifs, c'est-à-dire que le label qui leur est distribué est bien le bon (Paris = LOC, Marine = PERS ...), les Faux Positifs (les labels attribués ne sont pas les bons : Charlotte = ORG, Limoges = Pers, Aller = LOC)

4 CSV

- Concevoir un programme qui prend en entrée le jeu de données constitué au point 2 et qui donne en sortie un tableau CSV.
- Votre tableau doit contenir les colonnes : id (identifiant), texte de l'entité nommée, label de l'entité, contexte gauche, contexte droit, une colonne "annotation" qui comporte la mention "Vrai Positif" ou "Faux Positif" selon le cas.
- Votre tableau doit être complété automatiquement par les données figurants dans le dictionnaire stocké au format json.
- Votre tableau doit être stocké au format .csv

N'hésitez pas à vous servir d'internet !

5 Réalisation d'un tableau de données avec Pandas

`https://pandas.pydata.org/`

- Concevoir un tableau avec la librairie Pandas à partir du jeu de données au format .json
- Votre tableau doit contenir les colonnes : id (identifiant), texte de l'entité nommée, label de l'entité, contexte gauche, contexte droit, une colonne "annotation" qui comporte la mention "Vrai Positif" ou "Faux Positif" selon le cas.

6 Visualisation avec Seaborn

`https://seaborn.pydata.org/`

Concevoir un programme qui permet de représenter graphiquement les données de votre jeu de données à partir de votre fichier .json annoté.

Par exemple un graphique qui permet de montrer la quantité d'un type de label (LOC, PER, etc.) et s'il y a plus ou moins de "Vrai Positif" ou de "Faux Positif pour un type de label".