Programmation de Modèles Linguistiques (I)

(L5SOPROG L3 Sciences du Langage)

2024-2025

TD6 Rappels

Caroline Koudoro-Parfait, Sorbonne Université.

Objectifs

Vérifier l'acquisition des connaissances suivantes :

- Opérations sur les fichiers : lire/écrire un fichier .txt
- Comprendre la différence entre txt, json et csv
- TXT: tokeniser, compter les effectifs, et faire une loi de zipf
- JSON: lire les modèles de langue
- JSON: stocker résultats de la NER
- CSV: lire fichier annotation manuelle

Exercice 1 : Opérations sur les fichiers : lire/écrire un fichier

Nous avons vu dans les TDs précédents comment lire un fichier et comment lire une série de fichier.

Voici les tâches que vous devez réaliser dans un nouveau notebook :

- Copier-coller la fonction permettant d'ouvrir un fichier comme une chaîne de caractères
- L'utiliser pour montrer les 200 premiers caractères du fichier `ressources_TD6/Texte/DAUDET/DAUDET_ref/DAUDET_petit-chose_ref.txt` figurant dans le dépôt *"ressources_TD6"* fourni sur Moodle
- Utiliser la librairie `glob` pour faire de même pour tous les fichiers txt du corpus

Exercice 2 : Comprendre la différence entre txt, json et csv

Identifiez les 3 fichiers de modèles `modeles.txt`, `modeles.csv`, `modeles.json`.

Utilisez l'un de ces fichiers pour écrire un programme permettant d'extraire les 10 mots les plus fréquents des langues suivantes : Polonais (pl) et Grec (el).

Expliquez en commentaire pourquoi vous avez choisi ce format.

Exercice 3 : Tokeniser, compter les effectifs de tokens, et loi de Zipf

Les noms des fichiers textes sont formés sur le patron suivant :
AUTEUR_version-OCR.txt
Pour les fichiers txt du corpus vous allez :

- Les ouvrir et les tokéniser
- Afficher le nombre de mots de chaque texte
- Faire un graphique de type loi de Zipf avec une courbe pour chacun de ces textes. Le nom du fichier doit être du type "AUTEUR_version-OCR.png" et ranger dans le dossier "AUTEUR"

Exercice 4 : Lire et manipuler les modèles de langue

En vous appuyant sur le fichier choisi dans l'exercice 2 vous calculerez l'intersection entre le modèle de langue anglais et chacune des autres langues. Puis l'intersection entre chaque langue entre elles si le résultats de l'intersection est non nul. Recette :

- lire le fichier
- calculer l'intersection entre le modèle de langue anglais et les autres modèles de langue,
- proposer à l'utilisateur d'entrer un terme en entrée et d'avoir en sortie les langues dans lesquelles le mot apparaît. Exemple de mots "and", "the", "to", "on", etc.

Exercice 5 : Stocker les résultats de la REN

Dans le TD5 il était attendu que le texte soit annoté automatiquement avec deux modèles de langue de spaCy :

- *fr_core_news_sm*
- *fr_core_news_lg*

Nous attendons ici que vous proposiez une fonction qui permet de faire la reconnaissance d'entités nommées à partir des lignes suivantes de l'exercice 2 du TD5

```
for ent in doc.ents:
        print(ent.text , ent. start char , ent. end char , ent. label)
```

et d'obtenir en sortie un dictionnaire stocké dans un fichier au format json nommé en adaptant l'appellation suivante : "AUTEUR_model-de-langue-spacy.json"

```
{
"entite_0": {
  "Entite": "Paris",
  "Label": "LOC"
}
"entite_1": {
  "Entite": "Norine",
  "Label": "PER"
}
}, ...
```

Puis vous écrirez un programme qui prend ce fichier json en entrée et qui permet de compter le nombre d'entités nommées reconnues par catégories (PER, LOC, MISC, ORG)

Exercice 6 : Lire des fichiers annotés manuellement

Dans le TD5 vous avez rencontré deux manières d'ouvrir et de lire des fichiers au format csv.

Dans cette exercice vous devez lire les fichiers csv compris dans le dossier "ANNOTATION_MANUELLES" en utilisant :

- la bibliothèque csv. (exercie 1 du TD5, mais attention dans le cas présent vous devez ouvrir et lire un fichier csv et non pas ouvrir puis écrire dans un fichier csv)
- la Bibliothèque Pandas. (exercice 3 du TD5)

Vous devez ensuite compter le nombre d'annotations (le nombre de signes X)par colonnes, pour déterminer :

- combien de tokens ont été annotés au total
- combien de tokens sont des PER, des LOC, des MISC ou des ORG.

Exercice 7 : Bonus : Accord inter-annotateurs

- Récupérer les fichiers `CSV` comportant les annotations de vos collègues dans les groupes du TD5 ;
- calculer l'accord inter-annotateurs en vous appuyant sur le programme suivant : `https://gist.github.com/cbuntain/9dd7e42d5d8ab34609162410e06f3270`

Devoir

- les programmes attendus
- des images des graphiques pour loi de Zipf
- quelques phrases de conclusion sur les résultats (qu'est-ce qui était attendu, qu'est-ce qui est inattendu ?)

  Vous déposerez sur Moodle une archive zip nommée NUMETU.zip (où NUMETU est votre numéro d'étudiant) et contenant :

- Votre code exporté au format Python .py (et pas ipynb)
- le PDF du document que vous avez produit

  Date limite : indiquée sur le Moodle !