Programmation de Modèles Linguistiques (II)

(L5SOPROG L3 Sciences du Langage)

2024-2025

TD2 Analyses morphosyntaxique, lemmatisation et tokenisation

Caroline Koudoro-Parfait, Sorbonne Université

Objectifs

- Analyses linguistiques appliquées à un texte: Lemmatisation, Tokenization, POS tagging.
- Organiser son/ses scripts de manière lisible
- représenter graphiquement des résultats

Principaux outils nécessaires :

- spaCy
- Matplotlib/ Seaborn

1 Utiliser différents outils de spaCy pour l'analyse linguistique

Utiliser les outils proposés par spaCy pour effectuer les traitements suivants :

1. Lemmatisation
2. POS tagging

➜ Vous stockerez les résultats dans la structure de données et dans un fichier dont l'extension vous semble les plus appropriées.

➜ Vous proposerez une représentation graphique pour chacune des tâches qui vous semble pertinente.

➜ Vous proposerez une analyse des résultats et des graphiques pour les deux tâches.

2 Différentes tokenisations

étape 1 Tokeniser le texte en utilisant la fonction native python split()

étape 2 Tokeniser les textes en utilisant spaCy

étape 3 stocker les résultats dans une/des structures de données qui vous semblent appropriées

étape 4 Exprimer les résultats sous forme d'un graphique représentant une loi de Zipf. Ce graphique permet de comparer la courbe de la loi de Zipf pour les résultats avec la tokenisation split() et la tokenisation de spaCy. Quelles observations pouvez-vous formuler ?

Devoir

- 1 script python .py
- 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 !