Linguistique de corpus et traitement syntaxique

Code Apogée
2MNTM414

Composante(s)
UFR Langues et Civilisations

Période de l'année
Semestre 2

En bref

- **Mobilité d'études:** Oui
- **Accessible à distance:** Non

Présentation

Description

Les manipulations des données et la réalisation des travaux pratiques seront effectuées à l'aide du langage de programmation Python. Pour faciliter le déroulement du cours les étudiants sont demandés de suivre le cours 'Linguistique informatique : lexique' (5LNSE32) de Licence3 SDL au premier semestre (contactez Mme Anna Kupsc pour les détails) ou les tutoriels ci-dessous avant le début du cours.

La première partie du cours portera sur la collecte de données massives à partir du web (ang. Web scraping). Ce dispositif peut être appliqué par l'étudiant pour constituer son propre corpus ou bien pour récolter des données ciblées (ex. trouver des définitions dans un dictionnaire en ligne ou chercher des exemples dans les documents structurés).
La suite du cours se focalisera sur des traitements automatiques de données textuelles. Nous allons montrer des outils de catégorisation automatique de mots (ang. part of speech tagging), lemmatisation, troncation (ang. stemming) et d'analyse syntaxique. Nous allons discuter comment ces outils permettent de réaliser une étude linguistique de données langagières.

Informations complémentaires

Remise à niveau en programmation :

* https://www.learnpython.org/ (en anglais simple, partie 'Learn the Basics')
* https://www.datacamp.com/courses/intro-to-python-for-data-science (cours gratuit 'Python Basics', sur inscription)
* https://www.afterhoursprogramming.com/tutorial/python/python-overview/ (notions plus avancées)