Linguistique de corpus et traitement automatique

ECTS
3 crédits

Code Apogée
1MNTM511

Composante(s)
UFR Langues et Civilisations

Période de l'année
Semestre 1

En bref

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

Présentation

Description

Ce séminaire vise à initier les étudiants aux méthodes de traitement automatique modernes, en particulier en utilisant les techniques d'apprentissage automatique à partir de données textuelles volumineuses.

La première partie du cours sera consacrée à la classification de textes basée sur les données annotées. Nous aborderons les sujets suivants: comment préparer les données linguistiques pour la classification automatique, des différentes méthodes de classification, comment évaluer, interpréter et améliorer les résultats.

La deuxième partie présentera les méthodes de plongements lexicaux (ang. word embeddings). C'est une technique d'analyse sémantique basée sur l'hypothèse distributionnelle de Harris (1954). Nous allons construire des différents modèles de représentation sémantique de mots

à partir de grands corpus et discuter leurs propriétés.

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 sans connaissances en programmation ou du langage Python, sont invités à se signaler et suivre les tutoriels ci-dessous avant le début du cours

Heures d'enseignement

Linguistique de corpus et traitement automatique - CM	Cours Magistral	6h
Linguistique de corpus et traitement automatique - TD	Travaux Dirigés	18h

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)

Bibliographie

* Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.
* Mikolov, T., Chen, K., Corrado, G. Dean J. (2013). Ef#cient estimation of word representations in vector space. arXiv preprint arXiv:13013781
* Mikolov, T., Yih, W. T., Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 746-751.