Annotations sémantiques et applications en recherche d'information

Infos pratiques

> **ECTS** : 3.0
> **Nombre d'heures** : 24.0
> **Période de l'année** : Enseignement neuvième semestre
> **Méthodes d'enseignement** : En présence
> **Forme d'enseignement** : Cours magistral
> **Ouvert aux étudiants en échange** : Oui
> **Composante** : Philo, Info-Comm, Langages, Littératures & Arts du spectacle
> **Code ELP** : 4L9TL02P

Présentation

Ce cours exploite les savoirs et données touchant à la classification automatique de documents à partir de critères sémantiques. Il présentera ainsi des méthodes, modèles et applications qui s'intéressent à la catégorisation sémantique de documents annotés de façon automatique, semi-automatique ou manuelle. Les diverses catégories sémantiques abordées seront principalement explorées à l'aide de vecteurs de mots (*word embeddings*), ceux-ci faisant actuellement l'objet de nombreuses recherches applicatives et théoriques en linguistique. Tout au long du semestre, les étudiants développeront leurs propres projets en groupe et utiliseront plusieurs méthodes sur les données annotées de leur choix dans le but de les évaluer.

Objectifs

L'objectif est de réussir à mettre en place une chaîne de traitement de bout en bout afin de proposer un modèle permettant la classification automatique de documents sur la base de critères sémantiques pertinents. Il s'agira ainsi d'évaluer l'importance d'une annotation sémantique (lexicale, distributionnelle, grammaticale) aux applications en classification de documents.

Évaluation

M3C en 2 sessions

- Régime standard session 1 – avec évaluation continue (au moins 2 notes, partiel compris) : .....

ou

- Régime standard session 1 – avec évaluation terminale (1 seule note) : Compte-rendu de groupe écrit à rendre à la fin du cours (100%)

Un projet par groupe

- Régime dérogatoire session 1 : Compte-rendu écrit à rendre à la fin du cours (100%)

Un projet par groupe

- Session 2 dite de rattrapage : Compte-rendu de groupe écrit à rendre à la fin du cours (100%)

Un projet par groupe

Pré-requis nécessaires

Notions de développement en Python.

Compétences visées

Comprendre et savoir manipuler les vecteurs de mots, les algorithmes de classification standards, construire un corpus de documents structurés et développer une réflexion sur les critères sémantiques pertinents pour la classification automatique de documents.

Bibliographie

Kisselew, Max & Rimell, Laura & Palmer, Alexis & Padó, Sebastian. (2016). Predicting the Direction of Derivation in English Conversion. 93-98. 10.18653/v1/W16-2015.

Lapesa, G., Kawaletz, L., Plag, I., Andreou, M., Kisselew, M., & Padó, S. (2018). Disambiguation of newly derived nominalizations in context: A Distributional Semantics approach. *Word Structure*, *11*(3), 277-312.

Ye, Z., Li, F., & Baldwin, T. (2018, August). Encoding sentiment information into word vectors for sentiment

analysis. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 997-1007).