Projet de spécialité Ensimag 2A
Classification de séminaires
Encadrement : Jean-Baptiste Durand (Jean-Baptiste.Durand@imag.fr),
Louis-Philippe Kronek (lpkronek@dataiku.com)

Des séminaires scientifiques, grenoblois ou autres, sont recensés par plusieurs pages web d'annonces, par exemple
https://portail.math.cnrs.fr/agenda
pour les maths,
https://www.liglab.fr/evenements/seminaires
pour l'informatique, ...
Certaines personnes souhaitent être avertis de séminaires traitant d'un sous-thème particulier, par exemple statistique mathématique, statistique appliquée, statistique computationnelle, statistique appliquée, probabilités appliquées, machine learning ou autre. Le but du projet est de développer une méthode de machine learning pour déterminer d'après chaque résumé, à quel sous-thème se rattache l'exposé. Pour cela, on a une base d'apprentissage obtenue par les flux RSS des journaux scientifiques :
http://tictocs.ac.uk/
avec par exemple The Annals of Statistics qui est emblématique de "statistique mathématique", etc. (on a accès aux titres et résumés).
Éventuellement on pourra commencer par un problème plus simple : classifier les séminaires de maths, biologie, physique et informatique par exemple.
Il s'agira de récupérer automatiquement les flux RSS et les séminaires pour alimenter la base d'apprentissage et de test, de traduire automatiquement en anglais les résumés des séminaires en français par des méthodes déjà disponibles en ligne, de calculer des descripteurs numériques caractérisant les résumés (Amini *et al.*, 2013) puis d'appliquer une méthode d'apprentissage statistique pour réaliser la classification.
La plateforme dataiku sera mise à disposition des étudiants pour l'analyse des données.

Ce projet est adapté à un ou deux trinômes, certains membres ayant des compétences en classification supervisée, et d'autres en récupération automatique de flux RSS ou autres contenus web.

Références

Amini Massih-Reza, Gaussier Eric : Recherche d'information, Applications, modèles et algorithmes, Eyrolles, 2013.

http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf

http://www.dataiku.com