Voici le programme pour l'enseignement d'ouverture "Approche statistique pour la fouille automatique des textes".

Si vous êtes intéressé(es), merci de vous inscrire auprès de :
Julien VELCIN (julien.velcin@univ-lyon2.fr).

(Réponse vendredi 27 avril 2018 dernier délai)

Si vous avez des questions, n'hésitez pas à contacter par mail :

Dr. Julien VELCIN

e-mail : julien.velcin@univ-lyon2.fr
Laboratoire ERIC, bureau K198 -
5 av. Pierre Mendès-France - 69676 Bron Cedex, France
(0) 478 772 414 -

Les cours auront lieu à :
L'Université Lyon2 (prendre le T2 en direction St-Priest - Bel Air)
5 av. Pierre Mendès-France - 69676 Bron Cedex, France
Bâtiment i, 2ème étage, salle i203, au pied du tram T2, station "Parilly - Université - Hippodrome"

A l'issue de cette formation, Julien VELCIN vous transmettra une attestation de suivie.
Cette formation sera validée en formation scientifique à saisir dans siged onglet "formations.

Approche statistique pour la fouille automatique des textes

L'objectif de ce cours est de donner les bases pour le traitement automatique des données textuelles, mais également de présenter des avancées récentes réalisées à la frontière entre apprentissage automatique (machine learning) et traitement automatique des langues (natural language processing). Ce cours alternera la présentation des principales notions avec des travaux pratiques en utilisant le logiciel R. L'idéal est de venir avec un ordinateur portable où sera installé RStudio avec les bibliothèques "TM" (text mining), "rJava" et "mallet". Cependant, des ordinateurs fixes seront mis à la disposition des doctorants qui en feront la demande.

Le cours s'articule en quatre séances :

- lundi 7 mai 2018, 14h-17h : introduction générale, principales applications, représentations vectorielles des textes et tout premiers traitements avec la librairie "TM". Les doctorants sont encouragés à venir avec un corpus de leur choix comportant un volume suffisant de texte (au moins ~5000 phrases) et d'une qualité convenable (éviter les textes trop courts et mal écrits). Sinon, la librairie "gutenbergr" de R permet d'avoir accès à des milliers d'ouvrages libres de droits issus du site http://gutenberg.org/.

- lundi 14 mai 2018, 14h-17h : prétraitements standards des données textuelles (tokenization, suppression des mots-outils, stemming, etc.) et réalisation d'un petit moteur de recherche.

- lundi 28 mai 2018, 14h-17h : premiers éléments plus avancés de traitement des langues (étiquetage automatique des catégories grammaticales, désambiguïsation du sens des mots) et intégration de connaissances issues du Web (exemple avec WordNet).

- lundi 4 juin 2018, 14h-17h : extraction de thématiques et ouverture vers les techniques de plongement de mots (word embedding) en lien avec les dernières avancées en apprentissage profond (deep learning). Les expérimentations concerneront uniquement l'utilisation du modèle LDA (Blei et al., 2003) et seront réalisées avec la librairie "mallet", qui elle-même repose sur la librairie "rJava".