Proposition de sujet Impact
Fouille de textes

Sujet encadré par Philippe Preux, Professeur d'informatique, Université Lille 3, équipe SequeL du Laboratoire d'Informatique Fondamentale de Lille et de l'INRIA-Nord Europe.

Ce sujet ne se poursuit pas en mémoire de M2 maths.

Ce sujet concerne l'étude la fouille de textes. Par exemple, étant donné un ensemble de textes en langue naturelle, on veut regrouper les textes (ou mieux, définir une métrique sur les textes) par thèmes ; autre problématique classique : on veut déterminer la catégorie d'un texte étant donné un ensemble de textes déjà catégorisés (cela permet notamment de déterminer les spams parmi les emails).

Ces questions sont très bien étudiées en fouille de données depuis une quinzaine d'années. Les techniques qui ont sont issues ont fait la fortune d'entreprises comme Google, Yahoo !, Amazon, et feront la fortune d'autres dans l'avenir. Elles revêtent une importance capitale à l'heure du Big Data.

Le but de ce travail d'impact est d'étudier ces problématiques et mettre en œuvre différentes techniques constituant l'état de l'art sur le sujet. Le travail sera constitué d'une partie fondamentale permettant de bien poser mathématiquement ces problèmes et d'une partie expérimentale importante.

La partie expérimentale tirera partie de l'environnement disponible en R.

L'organisation temporelle du travail peut être la suivante :

- 1re quinzaine de novembre : étude des fondements mathématiques et prise en main des outils disponibles sous R
- 2e quinzaine de novembre : mise en application sur la classification supervisée de textes. On étudiera la mise en œuvre pratique de différentes approches algorithmiques pour cela, ainsi que les problèmes de passage à l'échelle
- décembre : classification non supervisée (en particulier, cartes de Kohonen et LDA)
- janvier : LSA, fouille de textes multi-langues
- février : récupération automatique des textes par *crawl* du web et leur fouille
- mars : rédaction rapport et soutenance

Candidat recherché : dynamique, autonome, curieux ; bonne **compréhension** des notions de mathématiques niveau prépa. D'un point de vue technique, le/a candidat(e) est à l'aise devant un ordinateur. S'il/elle connaît R, c'est encore mieux. Un avantage à celui/celle qui travaille sous Linux sur son ordinateur personnel.