Projet Master Humanités Numériques

Contexte :

Le projet LIFRANUM vise à constituer et analyser le corpus des productions littéraires francophones nativement numériques (produite exclusivement sur et pour le web). Il s'agit d'un projet financé par l'Agence Nationale de la Recherche (ANR) qui regroupe un laboratoire de sciences humaines (MARGE), un laboratoire d'informatique (ERIC) et la Bibliothèque Nationale de France (BnF).

L'une des difficultés de ce projet est la grande quantité de données à indexer et analyser pour répondre à un ensemble de question : Comment les auteurs utilisent-ils les dispositifs numériques pour écrire, diffuser des textes et rejoindre leurs lecteurs ? Comment se servent-ils des dispositifs numériques pour parler de littérature et imaginer de nouvelles formes littéraires ?

Pour répondre à ces questions, les membres du projet ont choisi d'employer des techniques de traitement automatique de la langue. Il est aussi envisagé de prendre en compte le contexte d'énonciation de ces productions (réseau).

Projet :

Un des niveaux d'étude pertinents dans ce cadre est celui des thématiques traitées par les auteurs. En effet, il peut permettre de détecter des formes d'écriture très particulières (les chroniques, les digressions, etc.), de regrouper les auteurs, et pourquoi pas d'identifier la littérarité ou non d'un texte.

En utilisant des méthodes d'apprentissage non-supervisées type LDA (Latent Dirichlet Allocation), il faudra produire des représentations thématiques au niveau des textes et des auteurs dans un premier temps. Il serait intéressant de tester des méthodes plus récentes, comme BERTopics (https://github.com/MaartenGr/BERTopic), si les étudiants le souhaitent.

Ces méthodes permettront dans un second temps des analyses plus approfondies, que ce soit par du clustering ou de la classification automatique, mais aussi en travaillant de manière plus étroite avec les chercheurs en littérature au sein du projet. Pour cela il faudra développer également une interface de visualisation des résultats permettant d'intégrer pleinement ces derniers dans la démarche.

Pour approfondir, l'information de structure entre les blogs et réseaux d'auteurs (citations, commentaires, likes, …) issus des différents crawls réalisés par la BnF notamment pourra également être ajoutée afin d'affiner la détection de communauté.

Données :

Ce travail s'articulera autour d'un ensemble de blogs extraits d'API de Blogger et Wordpress (par exemple http://carnets-haijin.blogspot.com/, https://terreurterreur.wordpress.com/, …). En effet, elles permettent d'obtenir des données propres et structurées. L'auteur est à

chaque fois clairement identifié et le texte facilement exploitable. Pour chaque blog, l'ensemble des posts sera disponible ainsi que les commentaires qui lui sont associés.

Contacts :

- Pour le Master HN :
  - Julien Velcin : julien.velcin@univ-lyon2.fr
- Pour le Projet LIFRANUM :
  - Alice Pantel-Cassagnaud : alice.pantel@univ-lyon3.fr
  - Enzo Terreau : enzo.terreau@univ-lyon2.fr