Proposition de TER pour la formation
Master Informatique Paris Descartes

Information concernant l'encadrant

Encadrant(s) : Séverine Affeldt et Lazhar Labiod, MCF, Université de Paris
Email : severine.affeldt@u-paris.fr, lazhar.labiod@u-paris.fr

Description générale du projet

Intitulé du projet :

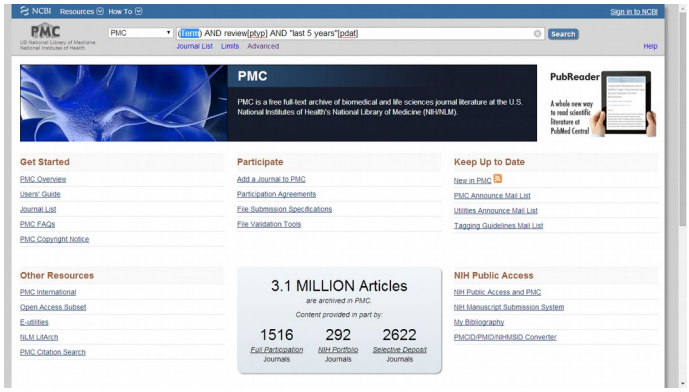
– Dashboard pour la constitution et l'analyse d'un corpus biomédical –

Contexte:

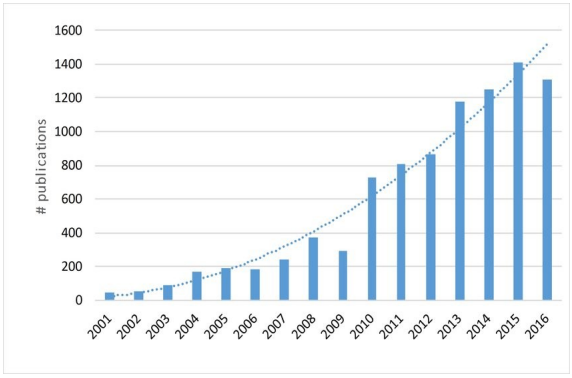
Le NLP (Natural Language Processing) appliqué au texte permet l'automatisation d'opérations telles que la constitution de *corpus* (ensemble de documents) ou leur annotation. Les méthodes existantes, largement accessible via les librairies R ou Python, permettent d'analyser et de valoriser de larges corpus comportant par exemple des news, des compte-rendus d'entretiens ou des commentaires de consommateurs.

Dans le domaine biomédical, de très nombreux articles sont aujourd'hui disponibles en ligne (Fig.1(a)) et leur exploitation peut permettre d'identifier des relations d'intérêt pour une meilleure prise en charge des patients. Toutefois, on produit de nos jours bien plus d'articles biomédicaux qu'on ne peut en lire (Fig.1(b)), et recouper l'ensemble des documents mise à disposition demande l'exploitation d'approches de NLP et de fouilles de texte avancées.

Figure 1. PMC/PubMed – Base de documents biomédicaux en ligne



(a) Page d'accueil PubMed



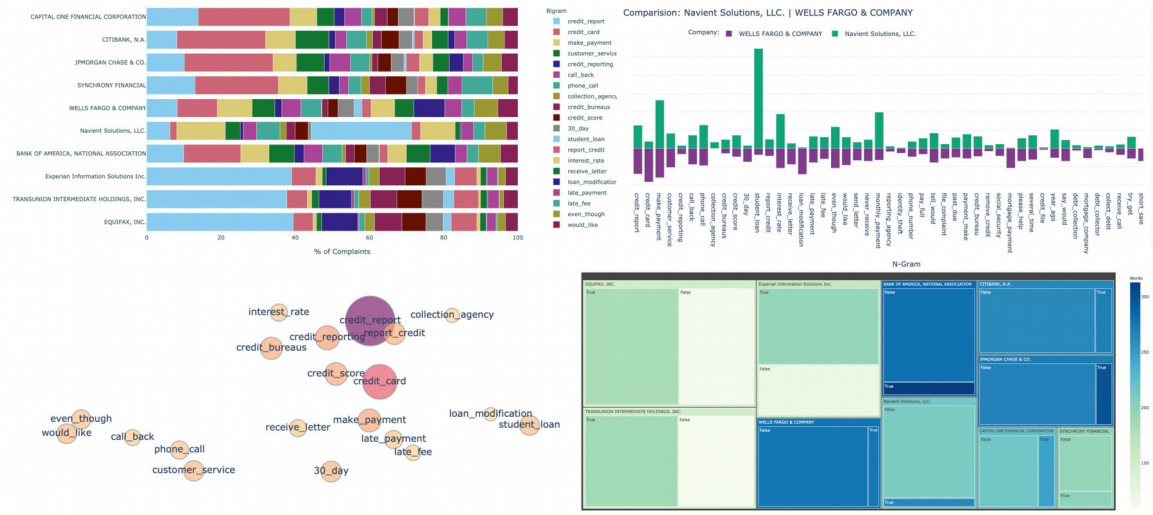
(b) Evolution du nombre de nouveaux articles PubMed

Objectifs:

Ce TER a pour premier objectif d'implémenter un outil ergonomique de création et d'analyse d'un corpus biomédical. Il doit pouvoir constituer un corpus de documents à partir de la base PubMed *via* un ou plusieurs mot-clefs et permettre le *nettoyage* du texte (eg. suppression de la ponctuation, des symboles) pour une exploitation NLP.

Un dashboard interactif (Fig.2) est requis pour la visualisation des données collectées et l'appel des méthodes de nettoyage. Il intégrera également l'appel à des méthodes de co-clustering classiques et la visualisation de leurs résultats.

Figure 2. Exemple de dashboard interactif avec Plotly Express et Dash



Réalisations attendues:

1. Constitution d'un corpus à partir de la base PubMed
2. Création d'un dashboard interactif avec Python Dash pour l'analyse des documents
3. Implémentation des méthodes de nettoyage de texte et intégration au dashboard
4. Utilisation des approches classiques de co-clustering et intégration au dashboard

Attention: Toutes les réalisations sont attendues en Python.