Fiche Projet
Réalisation, validation d'un moteur de recherche bibliographique, utilisation dans un cas pratique

PROBLEMATIQUE:

Rationaliser la recherche bibliographique. Réaliser un programme qui à partir d'articles de références, établit via un modèle de Machine-Learning une liste qualifiée d'articles traitant du même sujet.
L'apprentissage et la validation du module se fera à partir d'un ou plusieurs articles sur le thème : "*Machine Learning dans le domaine de la santé*"
Elaborer des datavisualisation sur le poids des mots, les tailles d'échantillon, les spécialités, les pays en pointe, et autres marqueurs

CONTEXTE :

Les articles scientifiques sont très structurés : abstract, rationnel, discussion, références bibliographiques, mots-clefs. On peut profiter de cette structure pour, à partir d'article de référence, établir une liste le plus exhaustive d'articles connexes.
Les articles scientifiques sont souvent mis à disposition sur le Web. On retrouve généralement un à cinq sites de référence par thématique de recherche. De même, le format des références bibliographique est assez standardisé :
http://www.icmje.org/recommendations/translations/french2015.pdf
https://www.nlm.nih.gov/bsd/uniform_requirements.html
http://www.niso.org/publications/ansiniso-z3929-2005-r2010

OBJECTIFS:

Fonction Scraping

Élaborer un programme de scraping qui à partir d'un ou plusieurs articles scientifiques, parcours récursivement le chaînage des références bibliographiques et les mots clefs pour établir une banque des articles connexes. Le programme pourra prendre en entrée une liste d'URL d'articles et des articles au format pdf regroupé au sein d'un répertoire.
Le choix des bibliothèques de scraping devra être argumenté
Les articles retrouvés devront être stockés dans un format facilement exploitable pour la suite.
Choisir le format adapté au stockage (csv, Json, ElasticSearch, MongoDB).

Fonction Machine-Learning

A partir de bibliothèques existantes, établir un programme de Machine-Learning qui à partir d'une labélisation permet de classifier les articles retrouvés en fonction de leur thématique. Une base d'entrainement, une base de validation et une base de test seront constituées à

partir des résultats de recherche sur des articles sélectionnés. Ces articles traitants du sujet : « Machine Learning dans le domaine de la santé ».

Datavisualisation

Mettre en place des métriques utilisables lors de la mise au point du moteurs.
Élaborer des tableaux de bord quantitatif (mots déterminant, pathologies, domaines, pays,…) sur le ML dans le domaine de la santé.

ENVIRONNEMENT TECHNIQUE :

Langage : Python

Données : csv, json, Base noSQL, ETL à déterminer

Data-visualisation : Tableau ou PowerBI

Gestion de projet : Jira, GitHub

EQUIPE:

ALBAN ROUX

BERND STADELMAYER

PATRICK CHEVARIER