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 :

– Fouille de texte pour l'exploration des maladies similaires au COVID-19 –

Contexte:

Un nouveau coronavirus responsable de la maladie à coronavirus 2019 (COVID-19; Fig.1), a provoqué une épidémie de pneumonies graves en Chine et dans d'autres pays à partir de décembre 2019. Actuellement, le nombre de patients atteints de la COVID-19 a augmenté rapidement, mais il est encore difficile de définir un traitement efficace contre la maladie ou qui permet d'éviter les formes graves.

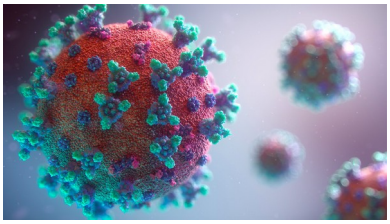


Figure 1: Représentation du coronavirus SARS-CoV-2 ©
Fusion Medical Animation on Unsplash

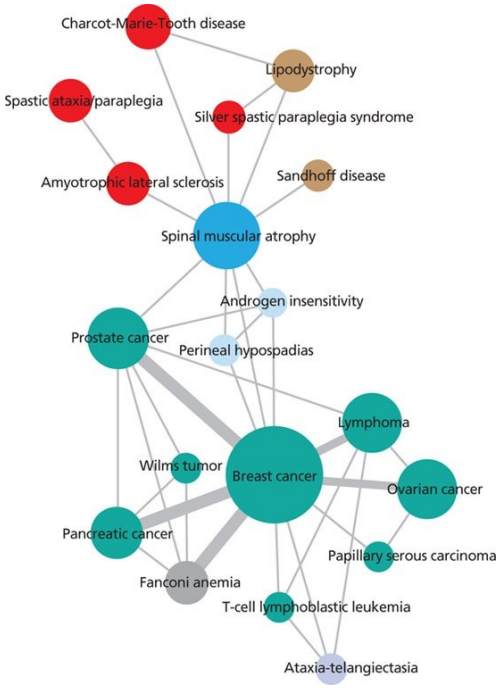


Figure 2: Exemple de réseau de similarités entre les maladies

La mise en évidence de similarités entre certaines maladies aujourd'hui bien connues et la COVID-19 offrirait des pistes intéressantes pour le développement de nouveaux traitements efficaces (voir un exemple Fig.2). Certaines prises en charges médicamenteuses exploitées dans le cadre d'autres infections pourraient également s'appliquer aux patients atteints de formes graves afin de réduire leurs symptômes.

Objectifs:

Dans ce projet, nous souhaitons explorer les similarités entre des maladies déjà connues et la COVID-19 afin de proposer des traitements innovants susceptibles de réduire la sévérité de la maladie. Nous proposons pour cela d'exploiter la litérature scientifique disponible en ligne.

A partir d'un corpus biomédical portant sur la COVID19, et à l'aide d'approches de fouilles de texte avancés -- telles que le clustering simultané de document et de mots (co-clustering), il serait possible d'identifier les maladies dont les caractéristiques se rapprochent le plus de ceux de la COVID-19. D'autres outils issus du Natural Language Processing (NLP) -- tels que les vecteurs de mots --, pourraient également nous permettrent d'explorer les relations entre les symptômes et les traitements.

Réalisations:

(1) Ce TER a pour premier objectif de créer une corpus biomédical autour de la COVID19 et des maladies similaires.

(2) Le second objectif de ce projet est l'exploitation d'approche de fouille de texte et de NLP pour l'identification de nouveaux traitements.

Proposition de PPD pour la formation
Master 2 MLDS - Université de Paris

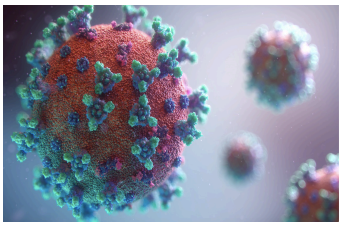
Information concernant les encadrants

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

Intitulé du projet :

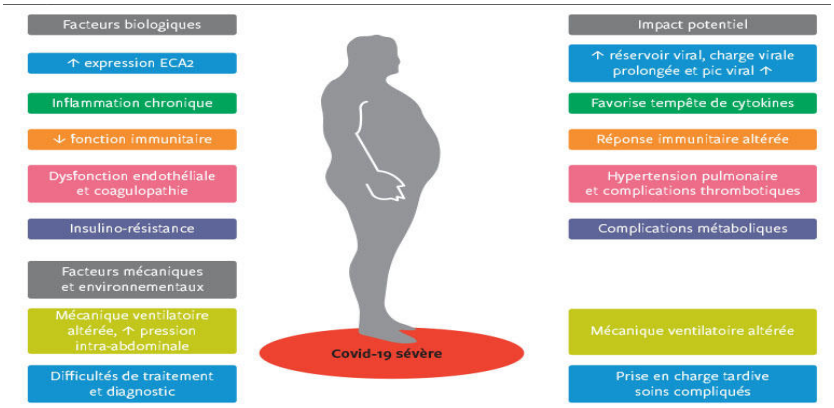
– Fouille de texte pour l'exploration des facteurs de sévérité du COVID-19 –

Contexte:



Un nouveau coronavirus responsable de la maladie à coronavirus 2019 (COVID-19), a provoqué une épidémie de pneumonie graves en Chine et dans d'autres pays à partir de décembre 2019. Le nombre de patients atteints de la COVID-19 a augmenté rapidement. Malgré les données recueillies, il est encore difficile d'établir et d'associer des profils de patients susceptibles de développer une forme grave et leurs comorbidités (ie. maladies et/ou troubles s'ajoutant à la maladie initiale).

Certaines comorbidités ont été identifiées pour certains profils de patients (Fig. 1). Toutefois, d'autres comorbidités peuvent également jouer un rôle important chez différents profils et accroître fortement la sévérité de la maladie. Une détection précoce des patients susceptibles de développer une forme sévères de COVID-19, et les comorbidités associées, est nécessaire pour réduire l'impact de cette maladie.



Objectifs:

Ce projet vise tout d'abord à identifier les profils de patients susceptibles de développer une forme grave de la maladie. Parallèlement à ce travail, il est également demander d'identifier les comorbidités principales en lien avec la COVID19. Plus précisémment, il s'agit d'identifier des marqueurs de prédictions d'intérêt pour la sévérité de la maladies.

Nous proposons pour cela d'exploiter la litérature scientifique disponible en ligne. Ce travail se fera à partir d'un corpus biomédical portant sur la COVID19, et à l'aide d'approches de fouilles de texte avancés, telles que le clustering simultané de document et de mots (ie., co-clustering).

D'autres outils issus du Natural Language Processing (NLP), tels que les vecteurs de mots, pourraient également permettrent d'explorer les relations avec les comorbités qui sont des facteurs aggravant des symptômes connus.

Réalisations:

(1) Ce PPD a pour premier objectif d'analyser une corpus biomédical autour de la COVID19, et plus particulièrement des comorbidités et des facteurs de sévérité de la maladie.

(2) Le second objectif de ce projet est l'exploitation d'approche de fouille de texte et de NLP pour l'identification de ces facteurs et de leurs interactions.

Projet de big data avec Apache Spark

Stanislas Morbieu

Mai 2021

Consignes générales

L'objectif de ce projet est d'utiliser Apache Spark pour implémenter un algorithme de partitionnement tel que k-means. Le projet est à réaliser en binome.

Le rendu attendu est double : un **rapport** et le **code associé**. Une attention toute particulière devra être portée à :

- l'**explication du fonctionnement** de Spark (ex : "Comment se fait la parallélisation des données pour telle partie de code ?", "Quelles sont les limites ?", ...) ;
- l'**interprétation** des résultats (ex : "On obtient une valeur de xx pour le score yy, ce qui signifie que les classes sont bien séparées.") ;
- l'**illustration** par quelques exemples (ex : "Voici la visualisation des résultats du clustering pour un jeu de données généré de la manière suivante... ").

Analyse descriptive

Clustering avec des implémentations disponibles

4. Partionner les points avec l'algorithme k-means, en spécifiant le nombre de clusters comme étant égal au nombre de classes du jeu de données (l'objectif ici n'est pas de trouver le nombre de clusters idéal) :
   (a) en utilisant scikit-learn
   (b) en utilisant une implémentation disponible dans Spark (voir la bibliothèque MLlib incluse dans Spark). Deux versions sont disponibles : une avec l'API DataFrame et une autre avec l'API RDD.
5. Analyser les résultats et la qualité de la partition obtenue. On pourra par exemple utiliser l'information mutuelle normalisée (NMI) dont l'implémentation est disponible dans scikit-learn.
6. Mesurer le temps d'exécution des différentes méthodes. Discuter des avantages potentiels des deux méthodes (scikit-learn et Spark).

Implémentation de K-means

7. Donner une implémentation de l'algorithme k-means avec Apache Spark. On pourra se limiter dans un premier temps au cas unidimensionnel (chaque point est représenté par une seule dimension). Pour celà :
   (a) Définir la fonction `compute_centroids` qui prend en argument deux RDD :
   `points` : chaque élément est la valeur du point pour sa seule dimension ;
   `cluster_ids` : chaque élément est l'id du cluster auquel est associé le point.
   Cette fonction retourne un RDD de couples (cluster_id, moyenne). On utilisera par exemple les fonctions `reduceByKey`, `mapValues`, `sortByKey`, `zip` et `map` et on suivra les étapes suivantes :
      i. Construire le RDD `sum_by_cluster_id` où chaque élément est un couple constitué de l'id du cluster et de la somme des éléments contenus dans ce cluster.
      ii. Construire le RDD `count_by_cluster_id` où chaque élément est un couple constitué de l'id du cluster et du nombre d'éléments contenus dans ce cluster.
      iii. Construire le RDD de couples (cluster_id, moyenne).
   (b) Définir la fonction `assign_clusters` qui prend en argument deux RDDs :
   `points` : comme défini précédemment ;
   `centroids` : retourné par la fonction `compute_centroids` définie précédemment.
   Cette fonction retourne un RDD dont chaque élément est l'id du cluster auquel est associé le point. On décomposera ce code en trois étapes :
      i. Définir la fonction `squared_distances` qui prend deux arguments : la valeur pour le point et la liste Python des moyennes des différents clusters. Cette fonction doit retourner une liste des carrés des distances entre le point et les différentes moyennes des clusters.
      ii. Récupérer les moyennes issues du RDD `centroids` sous forme de liste Python. Quelle supposition a-t-on faite pour effectuer cette opération ?
      iii. Utiliser la fonction `numpy.argmin` pour retourner le RDD des assignations.
   (c) Implémenter l'étape d'initialisation et l'itération.
8. Adapter l'implémentation précédente si nécessaire pour gérer le cas multidimensionnel
9. Analyser les résultats et les comparer aux méthodes précédentes.
10. Implémenter une variante de k-means. On pourra par exemple implémenter k-médoïdes ou Spherical k-means.
11. Utiliser l'implémentation précédente sur des exemples pertinents.

Projet Big Data

M2 MLDS/AMSD - 2021-2022

Contexte et objectifs

Lorsque les données sont très volumineuses, il n'est plus possible d'appliquer des méthodes qui supposent de les avoir toutes en mémoire en même temps sur une seule machine.
Le but de ce projet est de voir plusieures méthodes qui visent le même objectif : classifier des points de manière non supervisée. K means est choisi comme objet d'étude. Chaque partie correspond à une manière de l'implémenter, aucune n'est universellement meilleure que les autres puisque chacune fait des choix différents pour répondre à certaines contraintes.

A. Implémentation de k-means séquentiel (Python)

B. Implémentation d'une version *streaming* de k-means (Python)

C. Implémentation de k-means distribué (Apache Beam)

D. Implémentation de k-means séquentiel distribuée (Apache Beam)

E. Implémentation d'une version *streaming* et distribuée de k-means (Apache Beam)

Plus de détails sur github

Tous mes projets sont disponibles publiques sur mon Github : github.com/SeyfGoumeida

2020/2021 : (NLP) Fouille de texte pour l'exploration des maladies similaires au COVID-19
- https://github.com/SeyfGoumeida/Text_mining_Covid-19_Similaire_disease

2021/2022 : (NLP) Avancement sur le méme projet aprés son succés en université , Fouille de texte pour l'exploration des facteurs de sévérité du COVID-19
- https://github.com/SeyfGoumeida/PPD_NLP_sivirity_factors_COVID19

2020/2021 : (Apache Spark) Projet de BigData
- https://github.com/SeyfGoumeida/PPD_NLP_sivirity_factors_COVID19

2021/2022 : (Apache beam) Projet de BigData
- https://github.com/SeyfGoumeida/ApachBeam

2021/2022 : (Apprenssage supervisé) Projet de fin d'étude
- En alternance chez carrefour on travail sur un projet pour définir le nomnbre des caisses idéal à mettre en place dans chaque magasin, on se basant sur les KPIs quotidiens , hébdo et annuels
https://github.com/SeyfGoumeida/PFE

D'autre projets sont également présents sur le github , je vous invite à les consulter .