Big Data Architecture and Data Processing
Projet

Analyse et prédiction des infections COVID-19

Date limite: 15 janvier 2021, 15:00

Note 1 : Ce Projet doit être fait en **Spark**. Vous pouvez utiliser **Python, Scala, Java ou R.**
Note 2 : Vous pouvez utiliser qu'une partie des données

Des cas de nouveaux coronavirus ont été signalés pour la première fois à Wuhan, dans la province du Hubei, en Chine, en décembre 2019 et se sont depuis propagés dans le monde entier. Des études épidémiologiques ont indiqué une transmission interhumaine en Chine et ailleurs.

Des données épidémiologiques sont nécessaires pendant les épidémies émergentes pour mieux surveiller et anticiper la propagation de l'infection.
L'ensemble de données a été rendu public le 20 janvier 2020 contenant différentes informations sur les patients : clinique, démographique et géographique.

Vous avez le chois de choisir entre deux sources de données :

1. Les données épidémiologiques globales
Il peut être téléchargé sur (https://github.com/beoutbreakprepared/nCoV2019)

Vous pouvez aussi le télécharger ici :

https://docs.google.com/spreadsheets/d/e/2PACX-1vQU0SIALScXx8VXDX7yKNKWWPKE1YjFlWc6VTEVSN45CklWWf-uWmprQIyLtoPDA18tX9cFDr-aQ9S6/pubhtml#

Le dataset est également disponible sur la plateforme kaggle (https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge), choisissez le fichier COVID19_line_list_data.csv

Une description des champs de la base de données est présentée dans cet article: : *Epidemiological data from the COVID-19 outbreak, real-time case information*

2. Les données épidémiologiques en France
Vous pouvez aussi choisir le dataset concernant les infections avec le virus Covid19 en France, disponible ici :
https://github.com/lperez31/coronavirus-france-dataset
Choisissez le fichier deceased.csv.

Les deux fichiers seront aussi fournis par vos professeurs.

Note: *pour charger un fichier.csv avec Python, vous devez utiliser la function csv_read() de la librairie Pandas.*

L'objectif de ce Projet est d'utiliser Spark pour explorer ces données pour en extraire des connaissances afin d'aider la communauté à mieux comprendre la propagation du COVID-19.
Les questions demandées sont indicatives, vous êtes libre ***d'ajouter d'autres méthodes/modèles****.*

Le Projet est composé de l'ensemble des questions suivant :

Afin d'analyser l'ensemble des données, vous devez identifier et extraire certaines informations statistiques sur les données, par exemple : le type de données, les valeurs manquantes, les valeurs aberrantes, la

corrélation entre les variables, etc.

Dans le cas des valeurs manquantes, vous pouvez les remplacer par la moyenne, la médiane ou le mode de la variable concernée.

I. Analyse et prétraitement

1. Calculez les corrélations entre les variables. Quelles sont variables les plus corrélées avec la cible (‘result’)? Expliquez les résultats. Si variable ‘result’ n’existe pas, utilisez une autre variable cible.
2. Visualisez les données en deux dimensions en passant par l’ACP (analyse en composantes principales). Pouvez-vous utiliser une autre méthode ?

II. Apprentissage artificiel

1. Dans la suite, nous utilisons une méthode d'apprentissage automatique afin de prédire la classe : les patients sont soit «décédés» (‘died’) soit «sortis» (‘discharged’) de l'hôpital. Vous pouvez utiliser la classification par K-Nearest Neighbours (K-NN), l’arbre de décision ou le classificateur Bayes.

2. Les résultats obtenus doivent être validés en utilisant certains indices externes comme l’erreur de prédiction (matrice de confusion et précision) ou d'autres comme Rappel, F-Measure, …

3. Utilisez la régression pour prédire l'âge (age) des personnes en fonction d'autres variables. Vous avez le choix sur ces variables explicatives ? Comment choisissez-vous ces variables ? Calculez la qualité de la prédiction à l'aide de l'erreur MSE (Mean Squared Error).

4. Appliquer trois méthodes de clustering (K-means, NMF et CAH) sur l'ensemble de données pour segmenter les personnes en différents groupes. Utilisez l'index de Silhouette pour connaître le meilleur nombre de clusters.

5. Visualisez les résultats pour analyser visuellement la structure de clustering des trois méthodes.

6. Les données sont déséquilibrées. Vous pouvez les équilibrer en réduisant aléatoirement la classe majoritaire. Supposons que vous extrayez aléatoirement des échantillons équilibrés. Comment les résultats de la prédiction changeront-ils?

7. Comment pouvez-vous mieux gérer ce déséquilibre entre les classes ?

8. Pour trouver les meilleurs paramètres pour les modèles, quel algorithme pouvez-vous utiliser. Expliquez l'algorithme.

III. Architecture Map Reduce

1. Expliquez les propriété du challenge de « Big Data » concernant les données et la problématique du ce Projet

2. Présentez et expliquez l’avantage d’utiliser Spark par rapport à aux techniques classiques dans ce contexte.

3. Ecrivez le pseudo code MapReduce d’une méthode de classification que vous avez utilisé.