Projet Data Science

Master 2 ISiDIS 2015 / 2016
Projet à rendre le 15/01/2016 à 23h59

Le but de ce projet est de mettre en œuvre la démarche d'un data scientist qui doit analyser un volume conséquent de données hétérogènes à l'aide de technique de machine learning pour en extraire une information pertinente.

1 Enoncé

Le projet doit se décompose en trois étapes :

1. Formuler une question qui peut être étudiée à l'aide de données à votre disposition
2. Constituer une base donnée nosql à partir des données
3. Analyser les données pour répondre à la question formulée en utilisant des techniques de machines learning

2 Guide

2.1 Délivrables

Trois documents doivent être remis :
— le code qui vous permet de constituer la base de données,
— le code qui vous permet de mener l'analyse
— un document écrit (au format pdf) décrivant :
  — la question étudiée,
  — la source des données et la constitution de la base de données,
  — les outils utilisés en argumentant leur utilisation,
  — l'analyse des données et la réponse à la question.

Il peut être difficile de délivrer directement les données toutefois vous trouvez peut-être un moyen de les communiquer.

2.2 Source des données

De nombreuses données sont maintenant disponibles. On peut penser à plusieurs sources :
— Activités humaines : emails, photos, vidéo, logs, likes, etc.
— Activités des machines : données de capteurs,

— Open data : données publiques françaises (`https://www.data.gouv.fr/fr/`), association de l'open data (`http://www.opendatafrance.net/`), les concours (`https://www.kaggle.com/` ou `https://www.datascience.net/`), etc.
— Open API des réseaux sociaux : voir `http://www.programmableweb.com/`
— Le web : pages web, etc.

Naturellement, vous pouvez croiser plusieurs sources de données. Quelque soit les données que vous utilisez, veuillez indiquer les sources et les licences d'utilisation.

2.3 Environnement de développement

Vos développements doivent être réaliser sous l'environnement hadoop et/ou les outils associés (scoop, hive, spark, HBase, etc.).

2.4 Analyse à l'aide de techniques de machine learning

Plusieurs outils contenant les techniques de machine learning peuvent être utiliser pour l'analyse :
— Spark est un écosystème de traitement big data conçu pour effectuer des traitements rapides en RAM. Il contient beaucoup outils très bien intégrés en particulier Spark MLlib qui est une librairie de machine learning,
— Apache Mahout est un framework dédié au machine learning,
— Environnement R propose aussi une interface avec hadoop : RHadoop
— Un environnement python est aussi possible en combinant pydoop, pig, sciki-learn dans ipython. Vous trouverez un exemple ici : `http://nbviewer.ipython.org/github/ofermend/IPython-notebooks/blob/master/blog-part-1.ipynb`

3 Evaluation

Les critères d'évaluation de ce projet :
— La pertinence de la question étudiée
— Le volume de données analysées
— La nature et la diversité des données
— Le choix de la base de données nosql en fonction de la nature des donnée et de leur utilisation
— La qualité des données retenues dans la base après "nettoyage"
— Les techniques de machine learning mise en œuvre
— La qualité de l'analyse (arguments justifiés par les données)
— La justification et la maîtrise des outils utilisés (hadoop, spark, etc.)