Examen_R

22/01/2021

Introduction

L'objectif de ce rapport est de résumer puis d'évaluer Ce rapport vise à présenter et évaluer 5 travaux effectués par des étudiants de la promotion MSc Data Management 2020/2022 de PSB Paris.

L'ensemble des travaux est évalué selon les 5 critères ci-après:

- Aspect global du document ;
- Niveau de maitrise du sujet par l'auteur ;
- La qualité des codes ;
- Plue value ;
- L'intérêt intellectuel démontré ;

Je donnerais a la fin du document une autoévaluation des travaux auquels j'ai pris part.

1. Package "RandomForest"

a- Auteur : Thomas Massé Github dédié à ce package

b- Synthèse

Mon camarade a résussi a mettre en évidence l'usage du package Random Forest qui est un package très utilisé dans la prédiction et ce dans plusieurs secteurs. Il a étudié un cas d'utilisaion particulier. Il a dabord commencé par exliquer le jeu de données qui est un historique des parties d'un jeu vidéo.Il a également exliqué chacune des variables. L'objectif était de prédire le taux de réussite des joueurs en fonctions des données (la distance parcourue, le nombre d'armes ramassées..) Il a montré le mondèle d'entrainement avec l'algorithme "RandomForest". Celà a permi d'identifier les variables explicatives les plus importantes. Pour finir les résultats on été exposés pour chaque type de joueur.

c- Extrait de code

```
#Entraînement du modèle
system.time({
set.seed(123)
solo <- randomForest(winPlacePerc ~ ., data = head(trainsetSolo, 10000),
                     na.action = na.omit, importance=T, mtry=7, ntree=200)
             })
solo
```

Cette partie du code a permi de s'assurer de lexactitude des résultats de la prédiction.

d- Evaluation

- Aspect global du document : document bien rédigé
- Niveau de maitrise du sujet par l'auteur :sujet maitrisé par l'auteur
- La qualité des codes : lus ou moins bien fait
- Plue value : ce sujet est d'un très grand interet pour notre domaine
- L'intérêt intellectuel démontré : l'auteur a bien montré son interet pour le travail demandé.

Conclusion

L'auteur a su démontrer l'interet de l'usage du random forest. Un code un peu plus simple aurait permit de mieux apréhender le sujet qui represente quelque chose de très interessant.

2. Package "plotly"

a- Auteur : Imen Derrouiche & Olfa Lamti Github dédié à ce package

b- Synthèse

La data visualisation permet d'illustrer des données ou des informations de manière ludique, c'est une manière de capter l'attention du public.

Elle permet d'agréger des bribes d'informations minuscules, dispersées sur internet, par une représentation graphique interactive la plus ergonomique possible.

Le package Plotly permet donc de créer une variété de graphiques interactifs de qualité, de les organiser et de réaliser des dashborad dynamique.

c- Extrait de code

```
#Data visualisation
plot_ly (data,
         x = ~ jour,
         y = ~ nombre_ordre,
         type = 'scatter',
         mode = 'lines') %>%

  layout (title = 'Nombre de commandes quotidiennes dans un mois',
          yaxis = list (title = 'Nombre dordre'),
          xaxis = list (title = '1 mois'))
```

d- Evaluation

- Aspect global du document : document très bien rédigé
- Niveau de maitrise du sujet par l'auteur :sujet maitrisé par l'auteur
- La qualité des codes : codes clairs et bien faits
- Plue value : ce sujet est d'un très grand interet pour surtout pour l'aspect business
- L'intérêt intellectuel démontré : les auteurs on très bien montré son interet pour le travail demandé.

Conclusion

L'auteur a su démontrer l'interet de l'usage du package plotly.Le code était simple à appréhender et a tester.

3. Package "stringr"

a- Auteur : Léonard BOISSON Github dédié à ce package

b- Synthèse

stringR est un package très interessant qui permet de: manipuler des caractères individuels à l'intérieur des chaines de caractères un peu comme les regex, de réaliser des opérations locales, de manipuler et supprimer les espaces blancs

Dans le documentmon camarade expliques ces utilisations et donne des exemples.

c- Extrait de code Application en R:

```
library('stringr')
```

```
library(readxl)
```

```
d <- read_excel("C:/Users/leona/netflix_titles.xlsx", col_types = c("text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text"))
d
```

```
str_sub(d$type, 1, 4)
```

d- Evaluation

- Aspect global du document : document bien rédigé
- Niveau de maitrise du sujet par l'auteur :sujet maitrisé par l'auteur
- La qualité des codes : codes clairs et bien faits
- Plue value : ce sujet est d'un très grand interet pour l'eploration de données
- L'intérêt intellectuel démontré : l'auteur a bien montré son interet pour le travail demandé.

Conclusion

L'auteur a su démontrer l'interet de l'usage du package stringr. Malheureusement, les données sources n'ont pas été fournies, ce qui ne permet pas de tester le code comme il faut.

4. "LE PACKAGE GGPLOT2"

a- Auteur : ALLAKER Maxime et CHANEMOUGAM Siva Github dédié à ce package

b- Synthèse

ggplot2 est une librairie R de visualisation de données développée par Hadley Wickham. Les auteurs arès avoir présenté le package on montré comment installer celui ci et on exposés quelques cas d'utilisations.

c- Extrait de code Application en R:

```
#install.packages ( "ggplot2" )
#library(ggplot2) # installation de la librairie "esquise"
```

```
#chargement du jeu de données
data(iris)
head(iris)#affichage des 6 premieres lignes du jeu de données
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

d- Evaluation

- Aspect global du document : globalement le document est bien rédigé, on observe une partie du code pas très lisible.
- Niveau de maitrise du sujet par l'auteur :sujet maitrisé par l'auteur et contenu très fourni
- La qualité des codes : codes moyennement clair.

- Plue value : ce sujet est d'un très grand interet pour la visualisation des données.
- L'intérêt intellectuel démontré : l'auteur a bien montré son interet pour le travail demandé.

Conclusion

Le travail des auteurs a une réelle plus value. Il permet de cerner les cas d'utilisation rinciaux du package.Néamoins je ne le trouve pas très adapté pour des débutants.

5. "Hadoop et spark"

a- Auteur : Florine comlan et Ramya HOUNTONDJI [Github dédié à ce package]https://github.com/RamyaHTDJ/Psb_Ramya/blob/main/HadoopvsSpark.pdf)

b- Synthèse

Hadoop et sparks sont deux framework du big data qui ont des utilisations plus ou moins complémentaires. Ils sont des éléments indispensable a maitriser dans le domaine du big data et pour le traitement de données massives. Hadoop est très utile pour le stockage de données grace à HDFS. Map réduce lui permet de faire du traitement distribué. Spark quand à lui permet le traitement parallèle et integre un certain nombre de bibliothèque(grahe x, MLlib,. . .) L'utilisation de l'un ou de l'autre dépendra donc du besoin. Pour ce travail nous avons commencé par expliquer en quoi consistait nos deux framework un peu plus en détail, nous les avons ensuite comparé et nous avons montré comment installer hadoop via docker.

c- Extrait de code Comment installer Hadoop avec Docker?

Nous allons utiliser trois contenaires représentant respectivement un noeud maître (Namenode) et deux noeuds esclaves (Datanodes).

Etape 1: Installer Docker

Cliquez sur ce lien https://docs.docker.com/get-docker/ et suivez la procédure pour l'installer

Vérifiez la version installée avec :

`docker --version` **Etape 2: Télécharger l'image docker uploadée sur dockerhub** Cette image contient l'executable qui permet d'installer Hadoop (2.7.2), Spark (2.2.1), Kafka (2.11-1.0.2) et HBase (1.4.8). * `docker pull liliasfaxi/spark-hadoop:hv-2.7.2` **Etape 3: Créer les trois contenaires à partir de l'image téléchargée. Pour cela:** * Créer un réseau qui permettra de relier les trois contenaires:

`docker network create --driver=bridge hadoop` * Créer et lancer les trois contenaires (les instructions -p permettent de faire un mapping entre les ports de la machine hôte et ceux du contenaire):

```
docker run -itd --net=hadoop -p 50070:50070 -p 8088:8088 -p 7077:7077 -p 16010:16010 \
            --name hadoop-master --hostname hadoop-master \
            liliasfaxi/spark-hadoop:hv-2.7.2

docker run -itd -p 8040:8042 --net=hadoop \
        --name hadoop-slave1 --hostname hadoop-slave1 \
              liliasfaxi/spark-hadoop:hv-2.7.2

docker run -itd -p 8041:8042 --net=hadoop \
        --name hadoop-slave2 --hostname hadoop-slave2 \
              liliasfaxi/spark-hadoop:hv-2.7.2
```

Etape 4: Entrer dans le contenaire master pour commencer à l'utiliser. `docker exec -it hadoop-master bash` Le résultat de cette exécution sera le suivant: `root@hadoop-master:~#` Vous vous retrouverez dans le shell du namenode, et vous pourrez ainsi manipuler le cluster à votre guise. La première chose à faire, une fois dans le contenaire, est de lancer hadoop. Un script est fourni pour cela. Lancer ce script: `./start-hadoop.sh` Toutes les commandes interagissant avec le système Hadoop commencent par

hadoop fs. Ensuite, les options rajoutées sont très largement inspirées des commandes Unix standard. **Etape 5: Ajouter des données dans HDFS** Pour ce faire, nous allons utiliser le fichier purchases.txt.Ce fichier se trouve sous le répertoire principal de votre machine master. * Créer un répertoire dans HDFS, appelé input. Pour cela, taper: `hadoop fs -mkdir -p input` * Charger le fichier purchase.txt dans le répertoire input que vous avez créé:

`hadoop fs -put purchases.txt input` * Pour afficher le contenu du répertoire input, la commande est: `hadoop fs -ls input` * Pour afficher les dernières lignes du fichier purchases: `hadoop fs -tail input/purchases.txt` Vous touverez dans le tableau ci-dessous les commandes les plus utilisées pour manipuler les fichiers dans HDFS

Etape 6: Visualiser Hadoop offre plusieurs interfaces web pour pouvoir observer le comportement de ses différentes composantes. Le port 50070: permet d'afficher les informations de votre namenode: * `http://localhost:50070` Le port 8088: permet d'afficher l'avancement et les résultats de vos Jobs (Map Reduce ou autre): * `http://localhost:8088`

d- auto Evaluation

- Aspect global du document : le document a été rédigé simplement.
- Niveau de maitrise du sujet par l'auteur :Nous avons bien compris le sujet.
- La qualité des codes : peu de code.
- Plue value : ce sujet est d'un très grand interet pour le traitement de grand volules de données.
- L'intérêt intellectuel démontré : Nous avons bien montré l'interet du sujet.

Conclusion

Notre travail sur Hadoop et spark, constitue un très bon début pour des débutants souhaitant s'informer sur le sujet. Il permet également de voir les bases d'instalation de hadoop et didentifier quand l'utiliser. Je dirais que ce qui lui manque, est une étude plus appprofondie des deux packages.