SPARK K-MEANS CLUSTERING

(MASTER II SEP CS)

L'objectif principal de ce projet est de proposer un k-means clustering de Bristol City Bike en fonction de l'emplacement des stations vélos en utilisant spark. Le fichier ***BRISBANE***-city-bike.json contient des informations concernant l'emplacement de chaque vélo.

lien vers les données :BRISBANE-city-bike.json

```
{"number":122,
"name":"122 - LOWER RIVER TCE / ELLIS ST","address":"Lower River Tce / Ellis St",
"latitude":-27.482279,
"longitude":153.028723}
```

Pour se faire:

1- Instancier le client Spark Session.

2- Créer un fichier properties.conf contenant les informations relatives à vos paramètres du programme en dur.

```
[Bristol-City-bike]
Input-data=data/Bristol-city-bike.json
Output-data =exported/
Kmeans-level= 3
```

Utiliser le fichier de configuration pour récupérer les path.

```
import configparser
config = configparser.ConfigParser()
path-to-input-data= config['Bristol-City-bike']['Input-data']
path-to-output-data= config['Bristol-City-bike']['Output-data']
num-partition-kmeans = config['Bristol-City-bike'][Kmeans-level]
```

3-Importer le json avec spark : bristol = spark.read.json. en utilisant la variable `path-to-input-data`

4-créer un nouveau data frame Kmeans-df contenant seulement les variables latitude et longitude.

5-k means.

```
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.clustering import KMeans
features = ('longitude','latitude')
kmeans = KMeans().setK(num-partition-kmeans).setSeed(1)
assembler = VectorAssembler(inputCols=features,outputCol="features")
dataset=assembler.transform(Kmeans-df)
model = kmeans.fit(dataset)
fitted = model.transform(dataset)
```

6- quels sont les noms des colonnes de fitted ? vérifier qu'il s'agit de longitude, latitude, features, predictions.

7- Déterminer les longitudes et latitudes moyennes pour chaque groupe en utilisant spark DSL et SQL. comparer les résultats

8-Bonus: Faire une visualisation dans une map avec le package leaflet ***(correction) (La ville est en Australie et s'appelle BRISBANE et non pas BRISTOLE)***

9- Exporter la data frame fitted après élimination de la colonne features, dans le répertoire `path-to-output-data`

10- Commit & push dans un répo git en respectant l’architecture du projet, et en ajoutant une bonne documentation qui expose tous les résultats.