K-means Clustering

The script kmeans_job_opt.py is an implementation of clustering with k-means, the script is fully commented in order to explain better the approach.

What does the script do?

- Initializing the Spark Session
- Reading Json data
- Splits the dataframe into 2 : df_train, df_target (this is in order to speed up things instead of using the whole dataframe to find the best k-means model, we use just the df_train to find the model with a less error and then we predict the whole dataframe using the chosen model)
- Creates the vector features (I have used location : latitude + longitude in order to clusterize)
- Make the first model using k = 2 and we compute the global error
- Use a loop to iterate over k values from 3 to 99 999 (this value is arbitrary since we will never reach this value…), in each step we create a model and compute its error and then we compare the error with the previous one; we keep iterating until the error is no longer decreasing.
- Now that we have our model with the lowest error, we predict the cluster values using the original dataframe; new column prediction which contains the cluster values is added
- We store the whole dataframe as a parquet file for reusing it later, parquet is the best way if we want to keep track of the schema.

Output?

The output is a parquet file which can be found in the directory result, however for the sake of this exercise I have stored the result dataframe as a csv file using the code below:

```
parquetFileDF = spark.read.parquet("lat_long_result.parquet")
parquetFileDF = parquetFileDF.drop("features")
parquetFileDF.repartition(1).write.mode("overwrite").csv("data.csv", se
p=';',header=True)
```

How to launch?

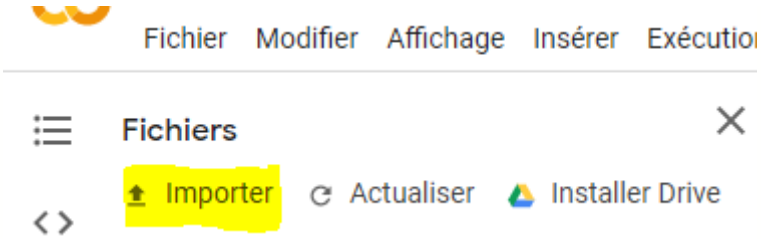
- Using Spark-submit command (“--master yarn” if yarn is used):

  ```
  spark-submit --master sparkMaster/local kmeans_job_opt.py
  ```

- Using Google Collab (Where I coded this script) , you can create your own Spark standalone and test the script by following these steps :
  1. Connect to https://colab.research.google.com/notebooks/intro.ipynb
  2. Create your python3 notebook
  3. Use this script to set up the whole environment

```
# ******* Spark Env Setup For Google Collab ****** #

# download spark 2.3
!wget -q http://mirror.its.dal.ca/apache/spark/spark-2.3.4/spark-
2.3.4-bin-hadoop2.7.tgz

# extract spark archive
!tar xf /content/spark-2.3.4-bin-hadoop2.7.tgz

# install findspark, numpy, pandas, matplotlib
!pip install -q findspark
!pip install -q numpy
!pip install -q pandas
!pip install -q matplotlib

# install ssh server
!sudo apt-get install -y openssh-server

# generate key
!ssh-keygen

# move authorize generated keys
!sudo cat /root/.ssh/id_rsa.pub >> /root/.ssh/authorized_keys

# restart ssh service
!sudo service ssh restart

# start all spark daemons including master slave (standalone)
!sudo spark-2.3.4-bin-hadoop2.7/sbin/start-all.sh

# check processes are started OK
!jps
```

4. Now we upload the zip file containing the data and the script manually using “import” as show in the picture below



5. Unzip the archive

```
#and then unzip it with
!unzip Brisbane_CityBikeNew.zip
```

6. Now the whole env is set just use the spark-submit command to try the script:

```
!spark-2.3.4-bin-hadoop2.7/bin/spark-submit --
master sparkMaster/local kmeans_job_opt.py
```