README

Environment Details:

- Ubuntu 16.10
- Python 3
- Jupyter Notebook
- Spark 2.3
- MySQL
- MySQL Java Connector jar

Step:1 Setup the development environment:

- We have setup the environment for Jupyter Notebook and Python using Anaconda with following commands:

```
$ wget http://repo.continuum.io/archive/Anaconda3-4.1.1-Linux-x86_64.sh
$ bash Anaconda3-4.1.1-Linux-x86_64.sh
```

- To setup the Spark, following commands were used.

```
$ wget
http://archive.apache.org/dist/spark/spark-2.3.0/spark-2.3.0-bin-hadoop2.7.tgz
$ sudo tar -zxvf spark-2.3.0-bin-hadoop2.7.tgz
```

- Download the MySQL java connector jar file so that we can access the MySQL from the Jupyter Notebook.

```
Mysql-connector-java-5.1.46.jar
```

- Setup the MySQL database. Create a blank database and import the SQL dump provided by the Yelp here:

https://www.yelp.com/dataset/download

- To reduce the data storage size and avoid the recomputation every time, we have performed some operations at the SQL level. Please see the feature engineering section of the Project report for queries to needed to run

Step:2 Export the Path:

Export following path in the User profile file .bashrc.

export PATH="/home/purvesh/anaconda3/bin:$PATH"

```
export SPARK_HOME='/home/purvesh/spark-2.3.0-bin-hadoop2.7'
export PATH=$SPARK_HOME/bin:$PATH
export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH
export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
```

Step:3 Configure Spark:

As the application is computation intensive, we need to increase the driver memory for the Spark. Change the following property in conf/spark-defaults.conf file at Spark location:

```
spark.driver.memory              8g
```

Step:4 Start Jupyter Notebook with following command:

To start the Jupyter Notebook with pyspark, run the following command. This command ensures that SparkContext is ready and available to use in the Notebook.

```
pyspark --jars /home/purvesh/mysql-connector-java-5.1.46.jar
```

Done.

Spark is running in the Jupyter Notebook with mysql connection on **http://localhost:8888.**

Step for Data Visualization using Gephi:

- Installation: Prerequisites is Java JRE versions 7/8.

- Installation Link: Download the file for windows from https://gephi.org/users/download/

- Windows: gephi-0.9.2-windows.exe file. Run the file.

- Ubuntu: Download the file for and execute following command to run the gephi tool:

```
tar xvf

cd /gephi/bin

./gephi
```