How to run the Project

Executing the project locally

1. Install Java 8 and set all the path variables
2. Install python 3.8 and set all the path variables
3. Install Spark and set all the path variables
4. Run the given python file

In case the above method did not work out it might be due to the library version incompatibilities. Please use databricks notebook to execute the code.

Follow these steps to run the project on databricks.

1. Login to databricks.com

2. Create a cluster with the following python libraries(PyPi) installed on the cluster:

- s3fs

- boto3

- sklearn

- nltk

- networkx

- pandas

- bs4

- requests

2. create a new notebook by selecting import notebook => from URL =>

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4987911583861567/4345931203554285/3511878718732474/latest.html>

3. once the cluster is up and running attach the notebook to the cluster

4. select run all commands

5. You will find the plots created in the notebook and output files (analysis.csv and output.csv) in the S3 bucket after the commands are executed

6. In order to access the output files (analysis.csv and output.csv) paste the 2 urls of cmd 25 in the browser seperatly and press Enter.

Files will be automatically downloaded once the URL is pasted in the browser and enter is pressed.

\*\* Note : please allow atleast 30-35 minutes for all the commands to execute as the data file is large.