README

A Spark program for parallel processing of the predictive engine for stock portfolio losses using Monte Carlo simulation in Spark.

Input:

The input to the simulator is a randomly selected list of stocks, total fund amount in USD, and a time period for which prices are recorded for these stocks at Yahoo Finance.

To collect data from Yahoo Finance we use the pandas DataReader which returns the stock data given a stock ticker and time frame. The *resources/inputscript.py* handles the fetching of data from Yahoo Finance given the tickers and start & end date; and outputs .csv files containing the data for each ticker.

To Run:

1) To get the stock data follow either option A or option B. Make sure the ticker list you give here matches with the string, *Ticker_list"* in the *resources/Input.conf* file in the project folder.

Option A:

- ❖ Load the *src/main/resources/inputscript.py to* HDP Sandbox VM.
- ❖ To fetch the data, execute the following command in HDP Sandbox VM;

```
[root@sandbox-hdp ~]# python inputscript.py GE,GS,OIL,AAPL 2015-01-01 2019-01-15
```

Note:

i. The 2nd argument, *"GE,GS,OIL,AAPL"* represents the list of stocks for which we fetch the data.
ii. The 3rd and 4th argument are the start and end date respectively.

- ❖ Now execute the following command to store the generated csv files to an input directory

```
[root@sandbox-hdp ~]# hdfs dfs -put GE.csv GS.csv OIL.csv AAPL.csv input/stockData/
```

Where *input/stockData* is the location of the input directory. Make sure to create the directory before copying into it.

Option B:

- ❖ A sample pre-fetched input directory for the above data can be found in the project location src/main/input/*.csv. You can also copy these files to your HDP Sandbox VM input directory location.

2) To build the jar execute the command;

```
>> sbt clean compile assembly
```

3) To run the Spark job enter the following command,

```
spark-submit --class Main --master local EuniceDaphneHW3.jar  input/stockData/*.csv
outputfile
```

where *input/stockData/*.csv* is the location of the input stock data files and *outputfile* is the location of the output directory.

The Youtube link on steps to deploy the Spark program on Amazon EMR can be found here.

Implementation:

1) Initially, all the stocks will have equal distribution of the investment amount. After one Monte Carlo Simulation we get the worst, average and best-case performance of each stock.
2) Record the tickers whose value plateaued/depreciated and those whose value increased. We use the average case performance for each of the stocks to make the decision.
3) We now sell half the amount on tickers whose value plateaued/depreciated and use that amount to buy more stocks from the tickers whose value increased.
4) Now, we run the Monte Carlo Simulation again with the new stock values and record the gain/loss results in a single day.

Sample Output:

A sample output is of the following format;

```
Your initial investment is $100000

In the Worst Case your stock holding could be $97910.0
In the Most Likely Case your stock holding could be $100040.0
In the Best Case your stock holding could be $101870.0
```