Pre-processing

Read the stock files inside the data directory. Stock files contain open & close price, high and low, volume and time related data about per stock.

date_time,Date,Time,Open,High,Low,Close,Volume
27.01.2016 11:00,27.01.2016,11:00,3.08,3.08,3.08,3.08,1950
27.01.2016 13:00,27.01.2016,13:00,3.08,3.08,3.08,3.08,3
27.01.2016 14:00,27.01.2016,14:00,3.08,3.08,3.07,3.07,1650
27.01.2016 15:00,27.01.2016,15:00,3.07,3.09,3.07,3.09,3065
27.01.2016 16:00,27.01.2016,16:00,3.09,3.09,3.07,3.07,6993

Read the files into pandas dataframe and make series of preprocessing operations before the training process.

Preprocessing occurs according to constraints given while running the run file

python3 run.py predict -i data -o info.txt -min_s 7 -min_d 800 -sdate '01.01.2015' -t 1

"-i", "input_path" - Path to csv files to be processed
"-o", "output_path" - Path to text file to store the short info
"-min_s", "min_time_interval_size" - help="Minimum time interval size
"-min_d", "min_number_of_distinct_days" - Minimum number of distinct days
"-sdate", "start_date_inp" - Start date to process the input files '01.01.2015' e.g.
"-t", "tiflag" - Flag for the additional technical indicators

In addition to these contsraints there are two other parameters as the open and volume thresholds which help us to discard the data according to these threshold values.

There are 100 files in the data directory. After the preprocessing the number of remaining files are around 20~30.

The remaining files construct a final dataframe which will be normalized between 0 and 1 and split into two parts as train and test dataset.

The dataframe is converted into a numpy array at the end of the preprocessing.

Train

On an example run:
Final Data Shape: (13520, 7, 5)
The shape info means that there are 13520 day, 7 hours data per day and 5 columns as date_time,date,time,Stock,Sma,Volume,Open

date_time,date,time,Stock,Sma,Volume,Open
07.03.2016 10:00,07.03.2016,10:00,41,0.0475074810364427 5,0.0038350719687727784,0.048884190744655866
07.03.2016 11:00,07.03.2016,11:00,41,0.048075807835950966,0.0006715672290514819,0.04871975569649989
07.03.2016 12:00,07.03.2016,12:00,41,0.048180194390962675,0.0010090273495828065,0.048696264975334747
07.03.2016 14:00,07.03.2016,14:00,41,0.049061680855506044,0.01048901444841534,0.04947145877378436
07.03.2016 15:00,07.03.2016,15:00,41,0.05022153146674732,0.014858839292895047,0.05048155978388537

Train data: %70
Test data: %30

Then, it starts creating models using the parts splits as 2 hours, 3 hours ... 6 hours of train data respectively. It creates 5 different models.

The model uses first two hours of all days in the data to train the models and uses open price, volume and sma columns as inputs. It calculates a heuristic price as an output.
Then it saves the model and repeats the same process for 3 hours split, 4 hours split etc.

Test

In the test phase, it makes predictions for all of the different splits of each day.

It loads the relevant model let's say the model trained with 2 hours split and then takes the first open price, volume and sma data for the first 2 hours data of that day.
Then it makes prediction using that data split as an input and it tries to predict the heuristic price for that split.

Then, it repeats the process for 3 hours, 4hours, 5 and 6 hours splits.

Based on that prediction, it makes buy and sell decision while comparing the prediction to certain threshold values.
It also calculates modelGain and DataGain parameters based on these trading operations.