**MINISTRY OF EDUCATION AND TRAINING**

**UNIVERSITY OF TECHNOLOGY AND EDUCATION**

**FACULTY OF INFORMATION TECHNOLOGY**



**SUBJECT PROJECT**

**MACHINE LEARNING**

**Stock Prediction using K-nearest neighbor Algorithm**

**STUDENT LIST:**

Vu Tuan Anh 17110004

Pham Vu Thien 17110075

Tran Ngoc Minh Quan 17110068

**LECTURER:** PhD. Le Vinh Thinh

**Ho Chi Minh City - 2021**

# **Student List**

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Full Name | Email | Student ID |
| 1 | Vu Tuan Anh | [17110004@student.hcmute.edu.vn](mailto:17110004@student.hcmute.edu.vn) | 17110004 |
| 2 | Pham Vu Thien | [17110075@student.hcmute.edu.vn](mailto:17110075@student.hcmute.edu.vn) | 17110075 |
| 3 | Tran Ngoc Minh Quan | [17110068@student.hcmute.edu.vn](mailto:17110068@student.hcmute.edu.vn) | 17110068 |

# **Task Assignment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Member in charge** | **Expected product** | **Rate** |
| Searching document | Teams | Knn documents | 100% |
| Coding demo using matlab | Tuan Anh | Demo for team members | 100% |
| Searching financial papers about stock | Thien | Data using for test and analyst documents research | 100% |
| Coding Knn to predict stock | Thien+Quan | Predict stock complete | 100% |
| Report | Team |  | 100% |

## **Introdution:**

“Security prices fully reflect all available information” (Fama, 1991). In her 1991 paper on Effective Capital Markets, Fama reiterates that a security prices at any time will “fully reflect” all available information. Based on this assumption she makes, the world today is dedicated to researching on ways to predict future stock prices based on their current information. And for this, many researches and data mining is done based on the data from stock market. In this paper, we attempt to do such analysis but with an emphasis of using a machine learning algorithm. We applied k-nearest neighbor algorithm in order to predict stock prices for a sample of seven major companies listed on the NASDAQ stock market to assist investors, management, decision makers, and users in making correct and informed investments decisions. According to the results, the kNN algorithm is mildly robust with a good accuracy; consequently the results were rational and also reasonable. In addition, depending on the actual stock prices data; the prediction results were close and fairly parallel to actual stock prices.

Even though, a lot of businesses are vested in researching on predicting the Stock market, financial data is considered as complex data to forecast and or predict. Predicting market prices are seen as problematical, and as explained in the efficient market hypothesis (EMH henceforth) (Fama 1991). The EMH is considered as bridging the gap between financial information and the financial market; it also affirms that the fluctuations in prices are only a result of newly available information; and that all available information reflected in market prices. The EMH assert that stocks are at all times in equilibrium and are difficult for inventors to speculate. Furthermore, it has been affirmed that stock prices do not pursue a random walk and stock prediction needs more evidence (Gallagher and Taylor, 2002).

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Price | Current Price |
| Open | Opening price of a stock for the day |
| Close | Closing price of a stock for the day |
| High | Highest price of a stock for the day |
| Low | Lowest price of a stock for the day |

**Figure 1: Main variables that affect stock movement**

In stock predictions, a set of pure technical data, fundamental data, and derived data are used in prediction of future values of stocks. The pure technical data is based on previous stock data while the fundamental data represents the company's’ activity and the situation of market. When we combine these information about a company and its stock, we assume that we should be able to yield a prediction of the future price of the stock. In classification approaches using machine learning algorithms, a data set is divided into training data set and testing set. kNN uses similarity metrics to compare a given test entity with the training data set. Each data entity represents a record with n features. In order to predict a class label for unknown record, kNN selects k records of training data set that are closest to the unknown records.

## **Literature Review**

There are a lot of researches going on in the field of data mining and future prediction. Since financial securities can yield a lot of profit from small investment in time and capital, a lot of researches are being done especially in this field. Since there are a large amount of financial information sources in the world that can be valuable research areas, getting hands on to these data and using these data is not a difficult ask anymore. Stock prediction becomes increasingly important especially if number of rules could be created to help making better investment decisions in different stock markets. Hellstrom and Holmstrom used a statistical analysis based on a modified kNN to determine where correlated areas fall in the input space to improve the performance of prediction for the period 1987-1996 (Hellstorm, Holmstrom 1998). The study developed potential guidelines to mine pairs of stocks, stock-trading rules, and markets; it also showed that such approach is useful for real trading. Moreover, Qian and Rasheed adopted KNN as prediction techniques much like in this paper (Qian and Rasheed, 2007).

## **Research Methodology and Analysis**

K Nearest Neighbor is an instance-based, competitive learning, and lazy learning algorithm. Instance based algorithms, sometimes called memory-based learning, are those algorithms that, instead of performing explicit generalization, use the instances seen in the training as a comparison standard. For kNN, the entire training dataset is the model. When a prediction is required for a unseen data instance, the kNN algorithm will search through the training dataset for the k-most similar instances. kNN is a competitive learning model because a majority vote is performed among the selected k records to determine the class label and then assigned it to the query record. kNN is considered a lazy learning that does not build a model or function previously, but yields the closest k records of the training data set that have the highest similarity to the test (i.e. query record).

### **When do we use KNN algorithm?**

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique, we generally look at 3 important aspects:

* Ease to interpret output
* Calculation time
* Predictive power

Let us take a few examples to place KNN in the scale:

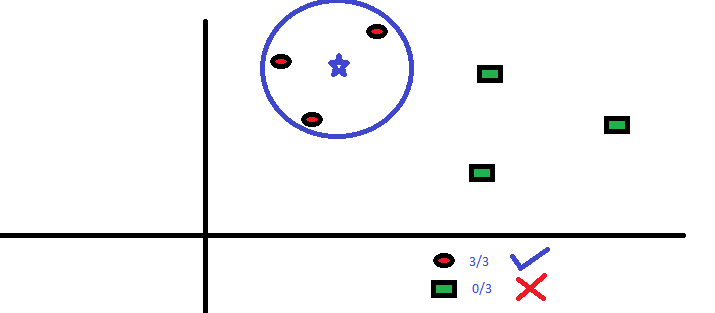
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Logistic Regression | CART | Random Forest | KNN |
| Ease to interpret output | 2 | 3 | 1 | 3 |
| Calculation time | 3 | 2 | 1 | 3 |
| Predictive power | 2 | 2 | 3 | 2 |

KNN algorithm fairs across all parameters of considerations. It is commonly used for its easy of interpretation and low calculation time.

### **How does the KNN algorithm work?**

Let’s take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS):

You intend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else. The “K” is KNN algorithm is the nearest neighbor we wish to take the vote from. Let’s say K = 3. Hence, we will now make a circle with BS as the center just as big as to enclose only three datapoints on the plane. Refer to the following diagram for more details:



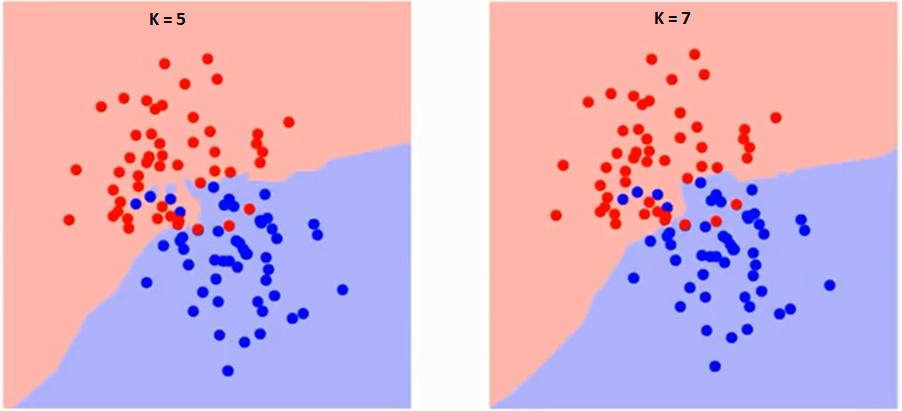
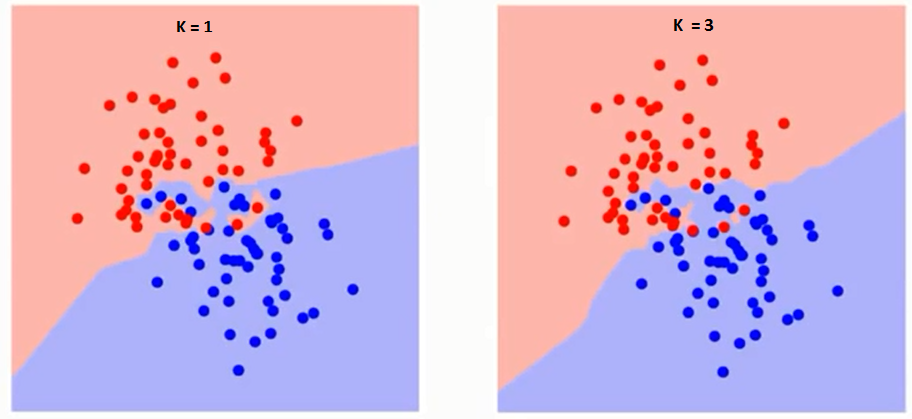
**Figure 2: KNN algorithm applied**

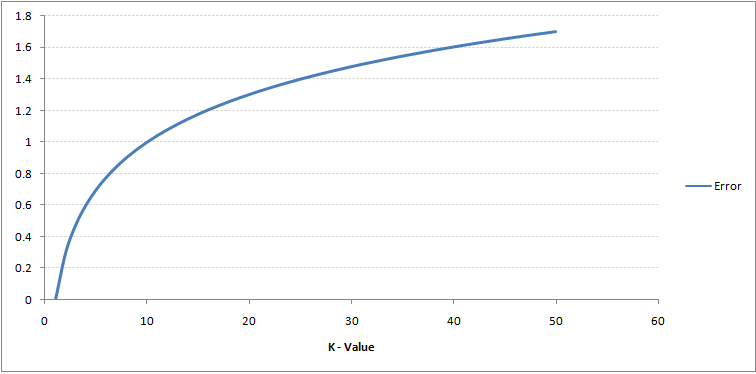
The three closest points to BS are all RCs. Hence, with a good confidence level, we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm. Next, we will understand what are the factors to be considered to conclude the best K.

### **How do we choose the factor K?**

First let us try to understand what exactly does K influence in the algorithm. If we see the last example, given that all the 6 training observations remain constant, with a given K value we can make boundaries of each class. These boundaries will segregate RC from GS. In the same way, let’s try to see the effect of value “K” on the class boundaries. The following are the different boundaries separating the two classes with different values of K.

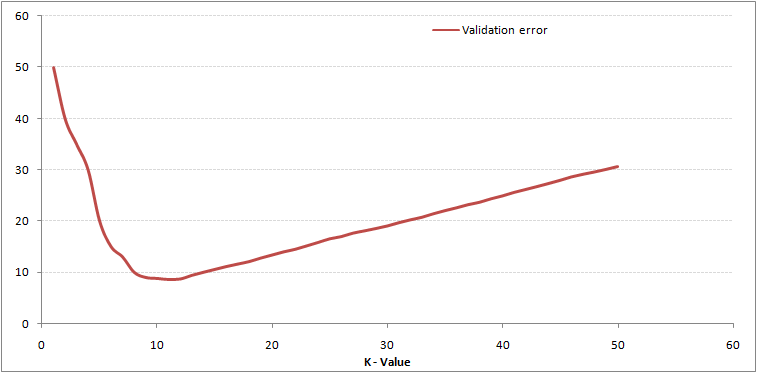
**Figure 3: The differences between variables k**



If you watch carefully, you can see that the boundary becomes smoother with increasing value of K. With K increasing to infinity it finally becomes all blue or all red depending on the total majority. The training error rate and the validation error rate are two parameters we need to access different K-value. Following is the curve for the training error rate with a varying value of K:

**Figure 4: Relationship between the rate of error and the value k**

As you can see, the error rate at K=1 is always zero for the training sample. This is because the closest point to any training data point is itself. Hence, the prediction is always accurate with K=1. If validation error curve would have been similar, our choice of K would have been 1. Following is the validation error curve with varying value of K:



**Figure 5: The relationship between the rate of validation error and the value k**

This makes the story more clearly. At K=1, we were overfitting the boundaries. Hence, error rate initially decreases and reaches a lowest point. After the minima point, it then increases with increasing K. To get the optimal value of K, you can segregate the training and validation from the initial dataset. Now plot the validation error curve to get the optimal value of K. This value of K should be used for all predictions.

### **Breaking it Down – Pseudo Code of KNN**

1. Handle data: Get data using yahoo finance and save and the load the dataset from CSV and split into test/train datasets
2. Similarity: Calculate the distance between two data instances
3. Neighbors: Locate k most similar data instances
4. Response: Use a majority vote for the class labels of k nearest neighbors and generate a response from a set of data instances
5. Accuracy: Summarize the accuracy of predictions
6. Main: Tie it all together

## **Mathematical Calculations and Visualizations Models**

**Accuracy:** Our data is structured such that we see “up” or “down” values for each day if its stock price has gone up or down respectively. So, since there are only two possible outcomes of the process, so we directly compare the predicted outcome with the actual outcome to calculate the accuracy of the model. We can see this implemented in the getAccuracy() function in our code. **Visualization:** For visualization, we have constructed a graph, for each stock, that shows the actual stock prices on a given day with the prediction for that day. We also have a graph for each stock that shows the predicted against actual change in stock price movement. And finally, we also have created, for each stock, we show their stock performance using their return.

## **Data Description, Results, And Analysis**

In this article, data from the NASDAQ stock exchange was analyzed and a brief data analysis was presented to provide the reader with the fundamental concepts of data attributes. Also, the obtained results of prediction of the NASDAQ stock exchange are provided.

* 1. **Data Description**

The sampling data were obtained from Yahoo Finance's NASDAQ stock exchange. The analysis contained inventory data from seven selected NASDAQ firms. Each one has six attributes, including date, opening price, high, low, adjusted closing price, and state change of stocks as set out in Figure 7 as an example training data set from Jan 1, 2011 to April 16, 2017. (this only has home randomly selected data to show how the data looks like since showing all data will take lots of space). The key factor influencing the prediction of a given stock using KNN algorithm is closing values. We used the Adjusted Closing price in our case, which displays business actions for the closing price of an equity on the day.

All of these estimates are obtained from Yahoo funding, so for any stock on a regular basis we compute the "price change" value. The adjusted stock selling price for that day is deducted from the day's adjusted stock closing price. And if this remaining number is positive, we'll give it an up mark that means that the market price has risen, otherwise we'll label the state adjustment as "down" to mean stock value decline. On 20294 documents the kNN algorithm is used.

|  |  |
| --- | --- |
| **Stock Name** | **Company** |
| AMZN | Amazon INC |
| DIS | The Walt Disney Company |
| SBUX | Starbucks |
| TWLO | Twilio Inc |
| TWTR | Twitter INC |

**Figure 6: The Stocks used that are listed in NASDAQ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Yesterday Close | Price Change |
| 1/2/2002 | 10.93 | 11 | 10.48 | 10.96 | up |
| 1/3/2002 | 11.13 | 11.94 | 11.05 | 11.9 | up |
| 1/4/2002 | 12.02 | 12.4 | 11.95 | 12.25 | up |
| 1/7/2002 | 12.08 | 12.51 | 12.08 | 12.34 | up |
| 1/8/2002 | 12.27 | 12.32 | 11.75 | 11.85 | down |
| 1/9/2002 | 11.96 | 12.17 | 11.3 | 11.53 | down |
| 1/10/2002 | 11.66 | 11.67 | 10.87 | 11.04 | down |
| 1/11/2002 | 11.03 | 11.34 | 10.93 | 11.03 | down |
| 1/14/2002 | 10.8 | 10.83 | 10.09 | 10.11 | down |

**Figure 7: Sample Data**

* 1. **Analysis And Results**

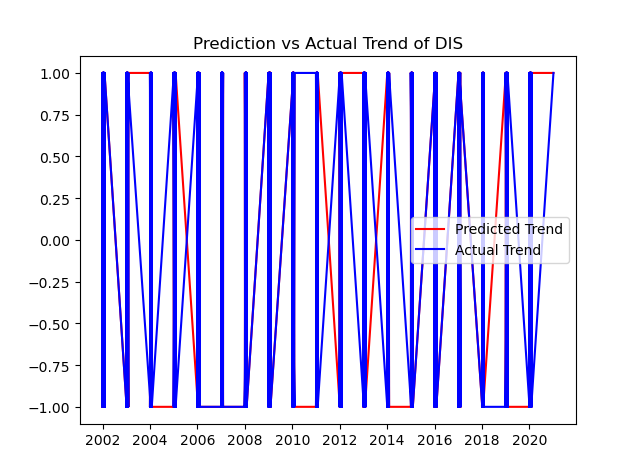
The results of the predicted stock price trend for each individual company used in the sample are presented as graphs along with the actual trend. The actual trend can be seen in blue and the predicted trend can be seen in red from figure 9 to 13. From figure 14 to 18, the stock chart for each stock can be seen. However, only this chart does not help much in determining the trend for analysts.

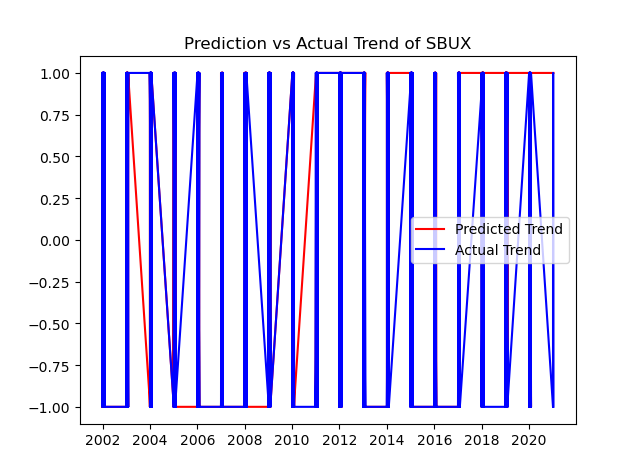
|  |  |  |
| --- | --- | --- |
| Stock | Data used | Accuracy(in %) |
| AMZN | Train: 3262 Test: 1532 | 74.186 |
| DIS | Train: 3267 Test: 1527 | 65.880 |
| SBUX | Train: 3181 Test: 1613 | 68.530 |
| TWLO | Train: 788 Test: 361 | 74.792 |
| TWTR | Train: 1192 Test: 617 | 78.119 |
|  |  | Average: 72.301 |

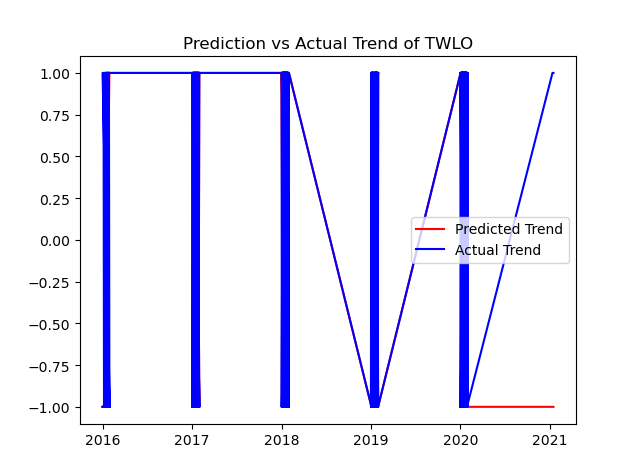
**Figure 8: Final result and accuracy**

# 

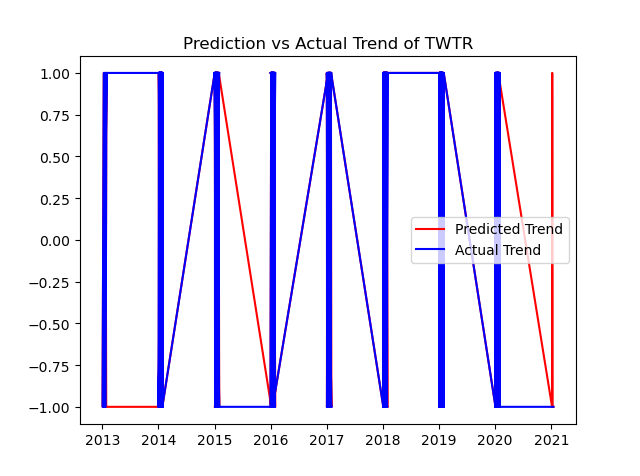
**Figure 9: Trend of Amazon**

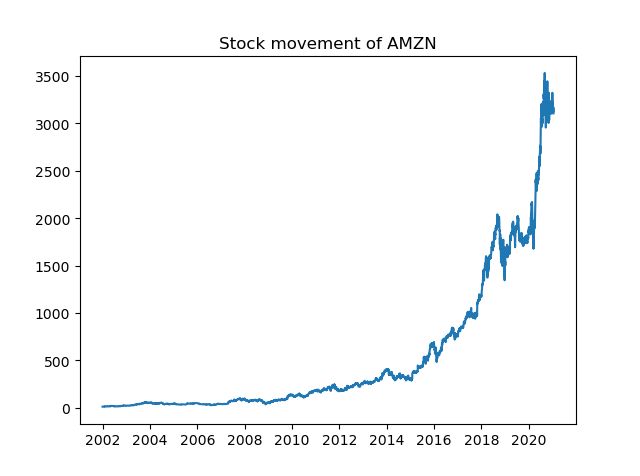
**Figure 10: Trend of Disney**

**Figure 11: Trend of Starbucks**

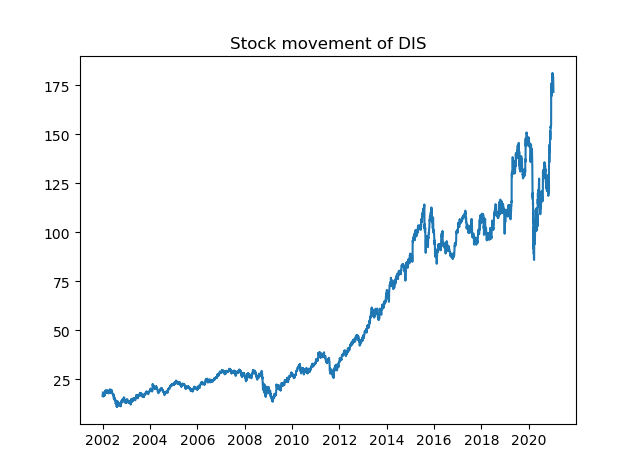
****

**Figure 12: Trend of Twilio**

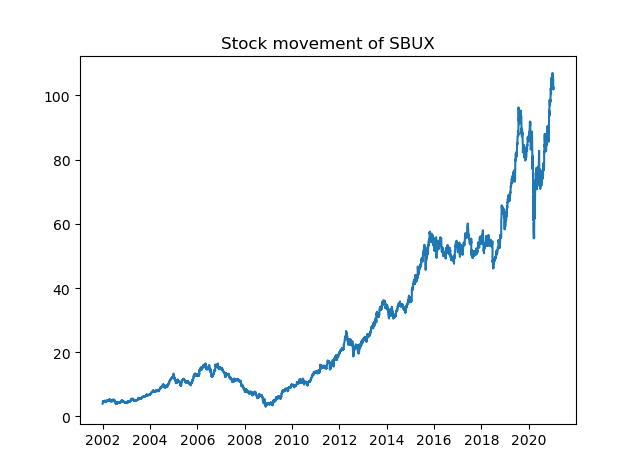
**Figure 13: Trend of Twitter**

****

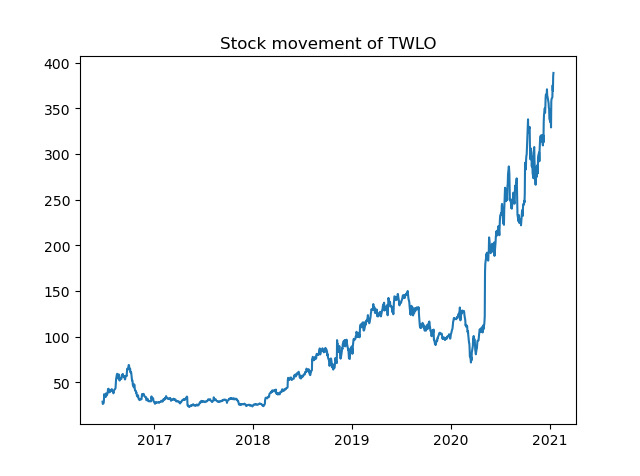
**Figure 14: Movement of Amazon**

****

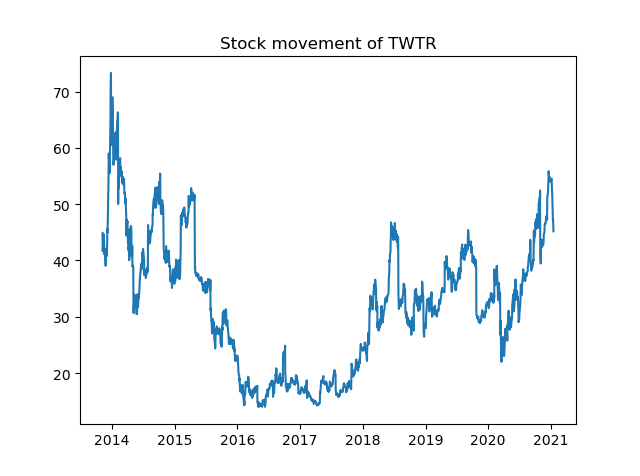
**Figure 15: Movement of Disney**

****

**Figure 16: Movement of Starbucks**

****

**Figure 17: Movement of Twilio**

****

**Figure 18: Movement of Twitter**