****

**WQD 7005 DATA MINING**

**Lecturer: Dr.Teh Ying Wah**

**GROUP PROJECT**

**ZHU TING (WQD 180077)**

**Introduction**

In this assignment which the mainly objective I which it is do the stock prediction. it is about we do the predication about unknown future events. The techniques which it from data mining, statistic, modeling, machine learning to analyze current data to make predictions about future. Stock market decision making is a very difficult and important task due to the complex behavior and the unstable nature of the stock market. There is an important need to explore the enormous amount of valuable data generated by stock market. All investors usually have the imminent need of finding a better way to predict the future behavior of stock prices, this will help the in determine the best time to buy or sell stocks in order to achieve the best profit on their investment. Trading in stock market can be done physically or electronically. When an investor becomes an owner of the company, according to the ownership percentage of this company’s shares. This give the stockholders rights on the company’s dividends. Financial data of stock market is a complex nature, which makes it difficult to predict or forecast the stock market behavior. Data mining can be used to analyze the huge and complex amount of financial data, which leads to better results in predicting.

The prediction of stock market helps investors in their investment decisions, by providing them strong insights about stock market behavior them strong insights about stock behavior to avoid investment risks. It was found that news has an influence on the stock price behavior. Stock market prediction based on news. New mining can be defined as the process of extracting hidden, useful and potentially unknown patterns from news data to obtain knowledge.

Sentiment analysis is the prices of determine people’s attitudes, options, evaluations, appraisals and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes. Sentiments analysis considered a particular branch of data mining that classifies textual data into positive, negative and neutral sentiments.

**Related work**

Several approaches for predicting stock market behavior and price trend have been studied in the literature. Some of these studies focus on improving the accuracy of prediction based on sentiment analysis of news or tweets along with stock prices such as. Others focus on price prediction with different time frame. Moreover, different research approaches proved that there is a strong correlation between financial news and stock prices change. Finally, research studies were conducted to improve the prediction accuracy such as . all previous studies have a. challenge because of the complexity of dealing with unstructured data.

**Data acquisition**

In this assignment, we are required to get gather all needed data from multiple views from multiple resourced that has the potential for useful analytics and insights. This process will be repeated until the data is useful enough for analysis purposes.

Below is the assignment flow:

Web Scrapping

Record video (by individual)

View hdfs file using Hive

Insert .csv into hdfs

Convert Output into .csv

In overall, there are three web-domains to be used, which is I3 Investor, The Star and Market StockBiz. From all these web-domains, we have five views in total as stated below, at the same time we do attached the weblinks summary as well. All the views will be crawled repetitively for data analysis purpose.

|  |  |  |
| --- | --- | --- |
| Types of view | Source | Repeat Crawl |
| Quarter result | The Star | Repeat by quarterly |
| Annual result | The Star | Repeat by yearly |
| KLSE Index | The Star | Repeat by daily |
| Investors’ Comments | I3 Investor | Repeat by daily |
| Stocks’ News | I3 Investor, Malaysia Stock Biz | Repeat by daily |

|  |  |  |
| --- | --- | --- |
| Crawled Data | Crawl Method | Link |
| Stock Price  Stock Code  Quarter Results  Annual Results | Scrapy  \*Sapura Kencana as example | <https://www.thestar.com.my/business/marketwatch/stocks/?qcounter=SAPNRG> |
| KLSE Index | Scrapy | <https://www.thestar.com.my/business/marketwatch/> |
| Investors’ Comments | Scrapy  \*Sapura Kencana as example | <https://klse.i3investor.com/servlets/stk/5218.jsp> |
| Stocks’ News | Scrapy  \*Sapura Kencana as example, one of the newslink | <https://klse.i3investor.com/blogs/AmInvestResearch/197983.jsp> |
| Malaysiastockbiz News | Panda | <https://www.malaysiastock.biz/Blog/Blog-Headlines.aspx> |

Web crawl using scrapy

|  |  |  |
| --- | --- | --- |
| Step 1 | : | Unzip the folder, Klse\_spider. |
| Step 2 | : | Change the directory in Anaconda Prompt usind 'cd' function until the Klse\_spider is found , for instance, cd ' [your directory]/Klse\_spider'  The folder is the root of scrapy. It is created when the project is initialized. |
| Step 3 | : | Change the path of the Chromedriver to '[your directory]/Klse\_spider/Chromedriver.exe' in Klse.py, stock\_index.py, Stock\_News.py located in [your directory]/Klse\_spider/spiders |
| Step 4 | : | Type 'scrapy list' in Anaconda Prompt, three spiders will be shown:   * KLSE * Stock\_News * Stock\_index |
| Step 5 | : | Type “scrapy crawl {spider} xxx.csv” in Anaconda Prompt   \*Note\* Scrapy and Selenium need to be installed prior to running. |

**Webcrawl using Pandas**

|  |  |  |
| --- | --- | --- |
| Step 1 | : | Go to anaconda where to find environment n try to download spider this software. |
| Step 2 | : | Write the code |
| Step 3 | : | Press run then the related file will be saved in Desktop |

After all the web data has been crawled, the csv files output will be stored in a place accordingly, in here we will show two methods of storing in. The first method is run Apache Hive and load the table from local in path, while the second method is store the csv files output into hdfs then using Hive to open the tables.

**Method I:**

**Run Apache Hive Instructions in Windows 10**

Install Hadoop on Windows Guide: https://github.com/MuhammadBilalYar/Hadoop-On-Window/wiki/Step-by-step-Hadoop-2.8.0-installation-on-Window-10

Install Hive on Windows Guide: https://www.youtube.com/watch?v=npyRXkMhrgk

|  |  |  |
| --- | --- | --- |
| Step 1 | : | Open the first cmd, then type the following codes  cd C:\hadoop-2.8.0\sbin  start-dfs  start-yarn |
| Step 2 | : | Open the second cmd, then type the following codes  cd C:\derby\bin    startNetworkServer -h 0.0.0.0 |
| Step 3 | : | Open the third cmd, then type the following codes  cd C:\hive\bin  hive  CREATE TABLE i3news (stock\_code int, stock\_name string, title string, subhead string, para string) row format delimited fields terminated by ',';  load data local inpath "C:\Users\Leonard Teng\Desktop\csv files\i3news.csv"  OVERWRITE INTO TABLE i3news;  show tables; |

**Data preprocessing**

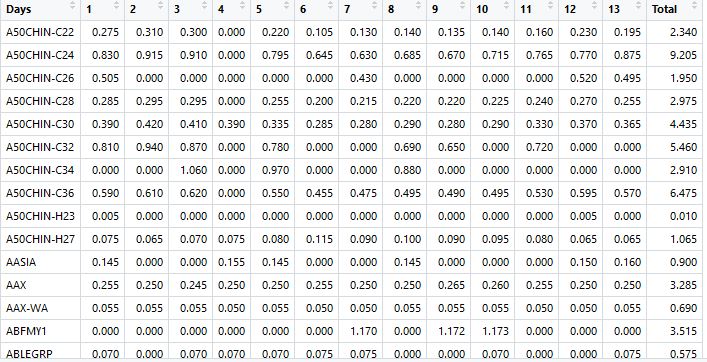
In this assignment, we are using two datasets. The first dataset is ss.csv from previous milestone and the second dataset are crawled data for 13 days.

First dataset are be used to calculate PCA and covariance between attributes. Second datasets are used to calculate SAX and PAA, because second dataset is time-series data.

For the second dataset, 13 days of data are compiled into one file using R. In order for us to perform SAX and PAA, data preprocessing is required such as cleaning data. For data preprocessing, we did:

1. Created a sum column at the right known as “Total” to sum all the stock values
2. Remove rows with sum value “0.000”
3. Replace NA with “0.000” in dataset
4. Change all columns attribute of those containing values to numeric

Figure 2.0.1 below shows the dataset that has already been cleaned or preprocessed. In the first column shows the list of companies, where from columns named “1” to “13” contain the stock values for each company from day 1 to day 13. The column named “Total” is the sum of all stock values from day 1 to day 13.



**Proposed model**

The proposed model helps investors avoid risks and financial crises when making investment decisions. The goal of the proposed model is to predict the stock market behavior, whether it is falling or raising. The proposed architecture combines the analysis of the stock market news and the historical prices together, in order to boost the classification accuracy of the stock market behavior . The study process performing text analysis on stock market news to determine the polarity of the news articles. Moreover, stock market historical prices opening, high, low and closing prices are analyzed to predict the future trends.

Open price is the open value of the stock in the current day, high and low prices are the highest and lowest values of the stock during a day respectively, and close price is the closing value of the stock for the current day.

**Sentiment analysis**

A close up of a white wall

Description automatically generated

**SAS modeling prediction**

A screenshot of a social media post

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Data Management**

To store the data acquired, we need a temporary storage and a warehouse. we evaluated two solutions, we evaluated two solutions, Google Big Query and Hadoop, which may be used for different use case of our data.

|  |  |  |
| --- | --- | --- |
|  | Google Big Query | Hadoop |
| Infrastructure | Serverless | Build on top of server and Network |
| On Premise | No (cloud as a service ) | Both Cloud and On Prem |
| Capacity | Petabytes | Petabytes |
| Database Support | Yes | SQL like interface |
| File Storage System | No | Yes |
| Distributed Processing | Database only | General Purpose, Structured and Unstructured data |
| Processing | Distributed, Realtime, Batch | Distributed, Batch |
| Data Types | Structured ( Row/Column ) | Structed and unstructured ( Data Lake) |
| Easy to setup and Use | Yes | Require physical setup of infrastructure on premise |
| SQL Query | Standard | Through eco system components, Hive |
| Search Speed | Very Fast for huge data | Very fast for huge data |

**Data warehouse**

In computing, a data warehouse also known as an enterprise data warehouse, is a system used for reporting and data analysis, and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate source. They store current and historical data in one single place that are used for creating analytical reports for workers throughout the enterprise.

The data stored in the warehouse is uploaded from the operational system. The data may pass through an operational data store and may require data cleansing for additional operations to ensure data quality before it is used in the DW for reporting.

On Hadoop platform, Hive is chosen mainly for below useful features to our analysis:

* Support structured and untrusted data
* Support SQL-like query language, a familiar to most researcher
* Support Thrift server , and interface allowing ODBC connection to Visualization tools such as PowerBI and Tableau.

**Data scheme**

The database schema of a database system is its structure described in a formal language supported by the database management system. The term “schema” refers to the organization of data as a blueprint of how the database is constructed. The formal definition of a database tables in the case of relational databases.

A screenshot of a cell phone

Description automatically generated

**Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted UP= TRUE** | **Precited UP=FALSE** | **Proportion** |
| **Actual UP=TRUE** | **1** | **3** | **4/15=26.67%** |
| **Actual UP=FALSE** | **9** | **2** | **73.3** |
| **Proportion** | **10/15=66.67%** | **5/15=33.33%** | **15** |

**Conclusion**

Based on the model, a potential strategy can be constructed as below. With below assumptions, a person can make:

The proposed model investigated the simultaneous effect of analyzing different types of news along with historical numerical attributes for understanding stock market behavior. Our proposal model improved the prediction accuracy for the future trend of stock market, by considering different types of daily news with different values of numeric attributes during a day.

Three categories of news data were considerate: news relevant to market, company news and financial reports that were published by financial experts about stocks. The proposed model consists of two stages, the first stage is to determine the news polarities to be either positive or negative.