Report 1

Tomi Theodor Kuusik, Kaarel Kristjan Katmuk, Kristen Surva

29.11.2021

Business understanding

Background

News can be catalysts for stock market fluctuations or the other way around. The hypothesis is that some keywords in news headlines can predict market movement trends e.g "Oil prices rise to record levels" would correlate with the rise energy market stocks (in this case the keywords "oil" and "rise" would be the words). However correctly identifying such keywords is very difficult because the keywords do not always lead to market movement and correctly determining such keywords need more situational awareness which can be hard to achieve with machine learning. Hence the aim of the task is to determine if such model can be trained and if this task is possible.

Business goals

1. Determine if a model that can predict stock market movement based on news headlines can be trained
2. Train such a model

Business success criteria

1. Insight if a model that can predict market movement based on headlines can be trained
2. If possible, implemented

Assessing your situation

Inventory of resources

- Stock market data (809MB)

- Worldnews reddit headlines (82MB)

Requirements, assumptions, constraints

- Regulize the data by date (i.e headlines for weeks and price movement based on simple moving average for 2 weeks data)
- For stock data etf's data or bigger stocks grouped by category should be used to reduce the data points.
- Smallest time range to be analyzed should not get bigger than 1 week.
- Time constraint = deadline of the course project

Risks and contingencies

- Stock market movement and keywords do not correlate, therefore predictions cannot be made.
- Data is too narrow and market movements lag behind/ahead headlines therefore connections do not seem apparent.

Defining your data mining goals

Goals

- Deliver a model that given a set of keywords outputs most probable securities/sectors which this headline will impact.
- Get a list of best indicators for every security or sector.
- Get a list of the keywords with the highest correlation in market fluctuation

Success criteria

- The model has to be at least 60% accurate.

Data Understanding

Gathering data

We don't need to gather our own data as we are using pre-made datasets.

Data Requirements

We require our data to have news headlines as well as accurate dates. We also need stock market information within the range of dates that our news articles were written in (circa 2008-2017).

Verify data availability

The data is usable.

Define selection criteria

We are going to use the whole reddit worldnews dataset. Stocks data is spread into separate files. We aren't certain yet whether or not we will need all of it. We most probably won't get use out of the author and over_18 fields of the reddit dataset. Similarly the time_created field is redundant since we have the date_created field. On the stocks dataset we won't exclude any fields for now.

The data was acquired from 2 seperate Kaggle datasets. Since we're using pre-made datasets, they are already formatted for us and there were no problems in reading the data into pandas dataframes.

Stock data

The stock data is firstly separated into ETFs and Stocks. For the dates(string), earliest and latest vary between stocks.. Each row of data encompasses a single day. Open(float) and Close(float) being the value of the stock when the market was opened and closed on that day. The High(float) and Low(float) signify the highest and lowest value of the stock on that day. Volume (int) is the number of stocks that were traded on the day.

Headline data

Each row of the reddit worldnews dataset is a single headline. Time_created(int) is the unix timestamp of when the post was created. Date_created(string) is the date formatted like %Y-%m-%d. Up_votes(int) and down_votes(int) are the amount of times the given post was upvoted or downvoted. The title(string) is the title of the post. Over_18(bool) whether or not the post is restricted to accounts whose owners are 18 years or older. Author(string) the reddit account name of the poster. Subreddit(string) the subreddit this post was made on.

Most important to us will be to see whether or not there is a correlation between the dates of certain articles being posted and the stock price fluctuating, this we can do as we have dates for both datasets.

Exploring data

The data doesn't seem to have quality problems. We are considering classifying stocks based on their industry, to see if some industries' stocks are affected by headlines more than others.

Verifying data quality

The data is good enough to support our goals. Only possible problem we might have is that we might or might not need to classify stocks by industry depending on if we can detect any correlation without classifying stocks.

Planning your project

Tasks:

1. Making the used data more readable (currently the stock market data is currently contained within separate .txt files which should be made more readable by first transferring the data to .csv files and then compacting the files to one dataset)
    - Python and pandas to read in the data and compact it into one file/dataframe (not more than 5h)
2. Cleaning up the data (keep only the data from years 2015 and 2016 as that's the only time where we have both the stock market data and the world news data)
    - Python (pandas) data cleaning tools (not more than 10h)
3. Determining the best way to find a correlation between keywords in r/worldnews headlines and US stock market price changes. (not more than 15h)
4. Applying the methods not more than (30h)
    - sklearn will be of help here
5. Analysis and conclusions (not more than 30h)
    - Looking at the results and drawing conclusions based on those: Did we manage to achieve our initial business and data mining goals

Division of work between group members

- We will try to spread the work between all the members at all task steps so everyone remains on the same page but we will keep in mind the minimum 30h of work per student.