**Daily News Headline for Stock Market Prediction**

Team #25

Members: Anastasia Gkousioudi, Wanying Ye, Cindy Chen, Tanya Gonzalez, and Swadha Rai

**Data Sourcing**

Name of Dataset 1: Top 25 historical news headlines from 2008-06-08 to 2016-07-01

It includes the date and news headlines from [Reddit World News](https://www.reddit.com/r/worldnews/?hl=). They are ranked by Reddit users’ votes, and only the top 25 headlines are considered for a single date. This dataset includes 73608 records and 2 columns (one for the date and one for the headline).

Size: 9.1 MB

Source: <https://www.kaggle.com>

Direct link: <https://www.kaggle.com/datasets/aaron7sun/stocknews?ref=hackernoon.com>

Name of Dataset 2: Dow Jones Industrial Average (DJIA) historical data from 2008-08-08 to 2016-07-01

It includes the date, opening and closing value, high and low value, the adjusted closing value, and the volume of stocks that have been traded that day. It consists of 1989 records and 7 columns (one for the date and 6 for each stock value).

Size: 167 KB

Source: <https://www.kaggle.com>

Direct link: <https://www.kaggle.com/datasets/aaron7sun/stocknews?ref=hackernoon.com>

Name of Dataset 3: Nasdaq Composite historical data from 2008-08-08 to 2016-07-01

(see description for Dataset 2)

Size: 194 KB

Source: <https://finance.yahoo.com>

Direct link: [https://finance.yahoo.com/quote/^IXIC?p=^IXIC&.tsrc=fin-srch](https://finance.yahoo.com/quote/%5EIXIC?p=%5EIXIC&.tsrc=fin-srch)

Name of Dataset 4: S&P 500 historical data from 2008-08-08 to 2016-07-01

(see description for Dataset 2)

Size: 161 KB

Source: <https://finance.yahoo.com>

Direct link: [https://finance.yahoo.com/quote/^GSPC?p=^GSPC&.tsrc=fin-srch](https://finance.yahoo.com/quote/%5EGSPC?p=%5EGSPC&.tsrc=fin-srch)

Name of Dataset 5: Combined DJIA and news headlines dataset

Combined DJIA and news headlines datasets into one, with 1989 rows and 29 columns. Each row corresponds to a single day. The first column of the new dataset is the date, the next column is the label and the remaining 25 columns contain the news headlines ranging from “Top 1” to “Top 25”. As label, we define the performance of the DJIA index on a specific day and denote with a value of:

* “1” if the adjusted closing value is equal to or greater than the opening value, and
* “0” if the adjusted closing value is less than the opening value.

Size: 5.6 MB

Source: <https://www.kaggle.com>

Direct link: <https://www.kaggle.com/datasets/aaron7sun/stocknews?ref=hackernoon.com>

**Data Schema**

Dataset 1: Top 25 historical news headlines from 2008-06-08 to 2016-07-01

| Field | Type | Description |
| --- | --- | --- |
| date | STRING | mm/dd/yy |
| news | STRING | Top 25 news headlines from Reddit World News |

Dataset 2: Dow Jones Industrial Average (DJIA) historical data from 2008-08-08 to 2016-07-01

| Field | Type | Description |
| --- | --- | --- |
| date | STRING | mm/dd/yy |
| open | FLOAT | Opening value |
| high | FLOAT | Daily high value |
| low | FLOAT | Daily low value |
| close | FLOAT | Closing value |
| volume | FLOAT | Daily trading volume |
| adj\_close | FLOAT | Adjusted closing price which considers corporate actions such as dividend payouts, stock splits, or the issuance of more shares |

Dataset 3: Nasdaq Composite historical data from 2008-08-08 to 2016-07-01

| Field | Type | Description |
| --- | --- | --- |
| date | STRING | mm/dd/yy |
| open | FLOAT | Opening value |
| high | FLOAT | Daily high value |
| low | FLOAT | Daily low value |
| close | FLOAT | Closing value |
| volume | FLOAT | Daily trading volume |
| adj\_close | FLOAT | Adjusted closing price which considers corporate actions such as dividend payouts, stock splits, or the issuance of more shares |

Dataset 4: S&P 500 historical data from 2008-08-08 to 2016-07-01

| Field | Type | Description |
| --- | --- | --- |
| date | STRING | mm/dd/yy |
| open | FLOAT | Opening value |
| high | FLOAT | Daily high value |
| low | FLOAT | Daily low value |
| close | FLOAT | Closing value |
| volume | FLOAT | Daily trading volume |
| adj\_close | FLOAT | Adjusted closing price which considers corporate actions such as dividend payouts, stock splits, or the issuance of more shares |

Dataset 5: Combined DJIA and news headlines dataset

| Field | Type | Description |
| --- | --- | --- |
| date | STRING | mm/dd/yy |
| Label | INTEGER | 1: adj close value >= open value  0: adj close value < open |
| Top1, Top 2, … Top 25 | STRING | News headlines from Reddit World News |

Dataset 6: Combined DJIA, Nasdaq Composite, S&P 500 and news headlines dataset

This dataset includes the labels for Nasdaq Composite and S&P 500. As it was mentioned before, as label, we define the performance of the each index on a specific day and denote with a value of:

* “1” if the adjacent closing value is equal to or greater than the opening value, and
* “0” if the adjacent closing value is less than the opening value.

| Field | Type | Description |
| --- | --- | --- |
| date | STRING | mm/dd/yy |
| label\_DJIA | INTEGER | 1: adj close value >= open value  0: adj close value < open |
| label\_Nasdaq | INTEGER | 1: adj close value >= open value  0: adj close value < open |
| label\_SP500 | INTEGER | 1: adj close value >= open value  0: adj close value < open |
| Top1, Top2, … Top25 | STRING | News headlines from Reddit World News |

**Data Preprocessing**

Step 1: Remove all irrelevant characters or punctuations

* Since the headlines are crawled from Reddit, there are many irrelevant punctuations such as “ ‘ or \b. We need to delete them because they are not useful for modeling.

Step 2: Convert all characters to lowercase

* We want to treat “Man” the same as “man”.

Step 3: Lemmatization

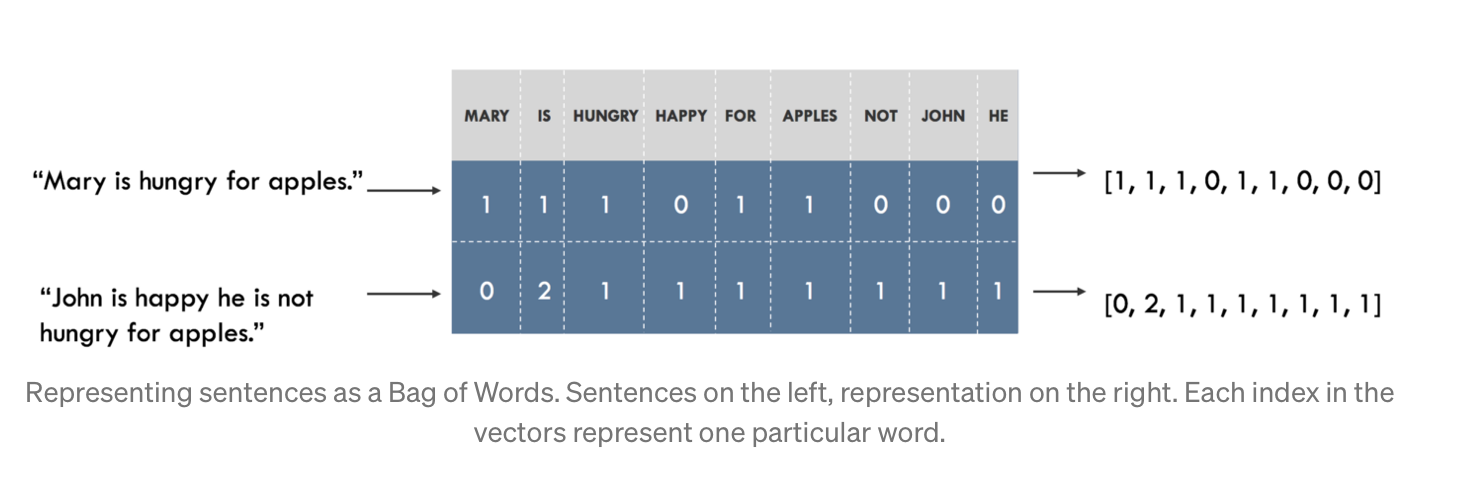
* Convert to root form: eg. Reduce words such as “am”, “are”, and “is” to a common form such as “be”.
* Plural to single: eg. Change “countries” to “country”.
* We tried both the NLTK and Spacy package and it seems that Spacy can handle lemmatization better!

Step 4: Remove step words

* Remove words with no actual meaning such as “the”, “a”, “this”.

Step 5: Convert the text representation to numerical measurement

* Use countvectorizer (bag of words) to convert our text data to a sparse matrix with counts (most of them are zeros).
  + We build a vocabulary of all the unique words in our headline dataset, and associate a unique index to each word in the vocabulary. Each sentence is then represented as a list that is as long as the number of distinct words in our vocabulary. At each index in this list, we mark how many times the given word appears in our sentence. An example is attached below (Figure 1).

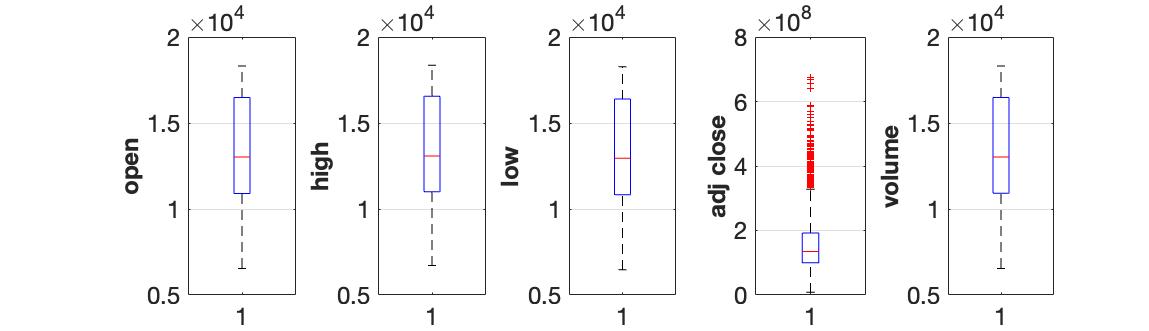


*Figure 1. Conversion of text data using countvectorizer.*

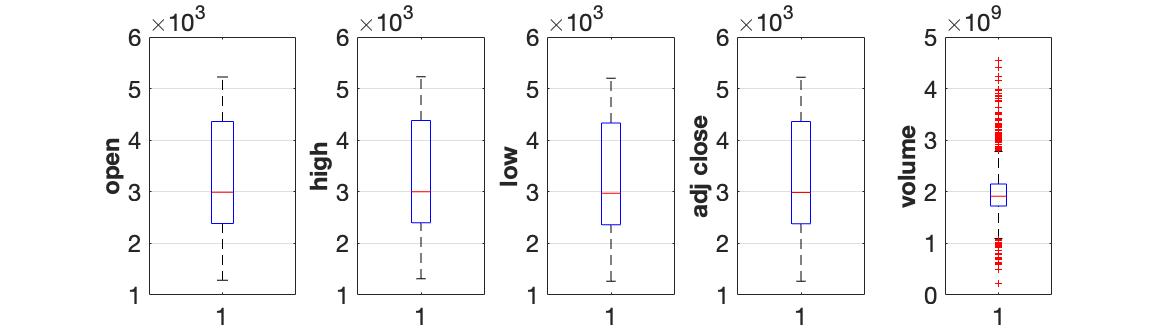
**Exploratory Data Analysis (EDA)**

1. Box Plots

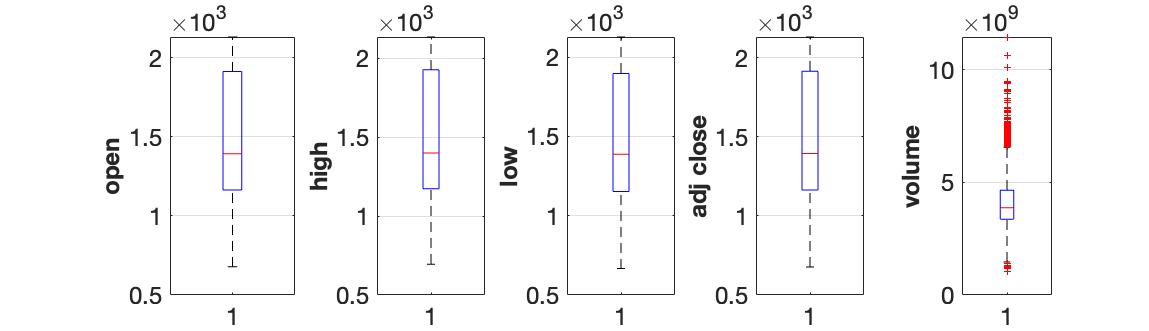
**DJIA**

****

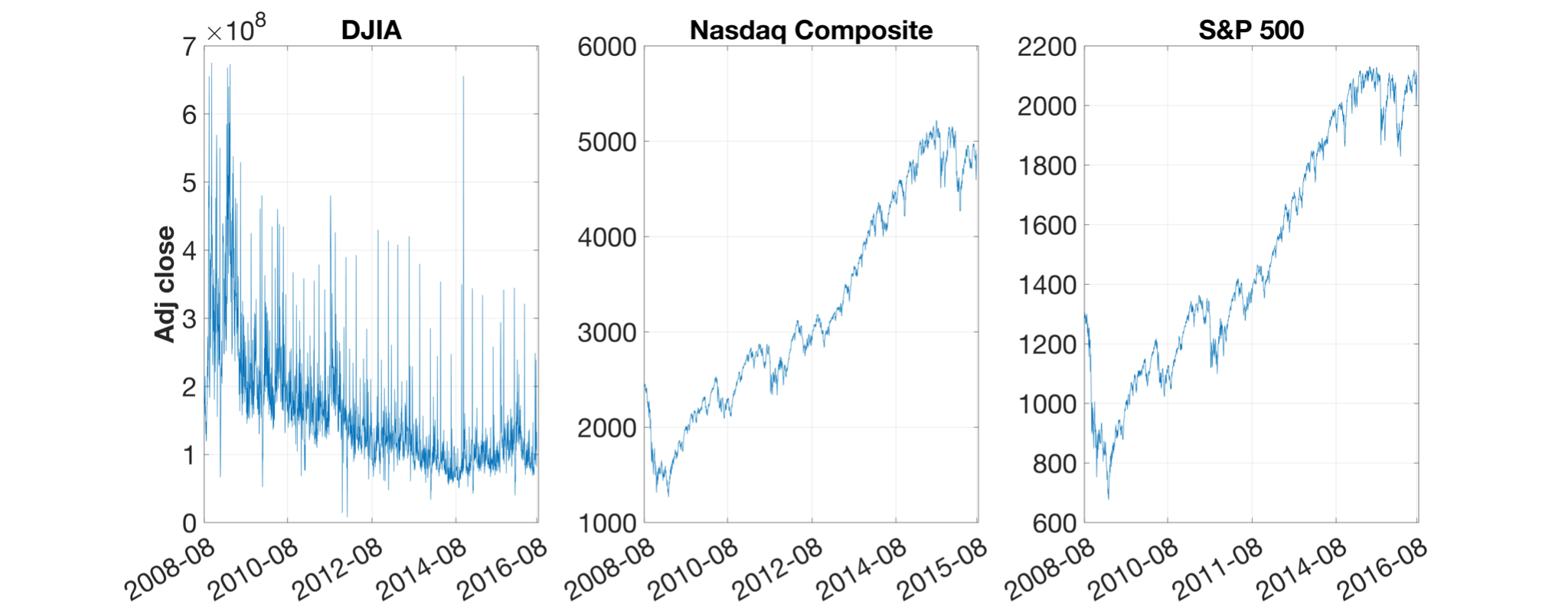
**Nasdaq Composite**

****

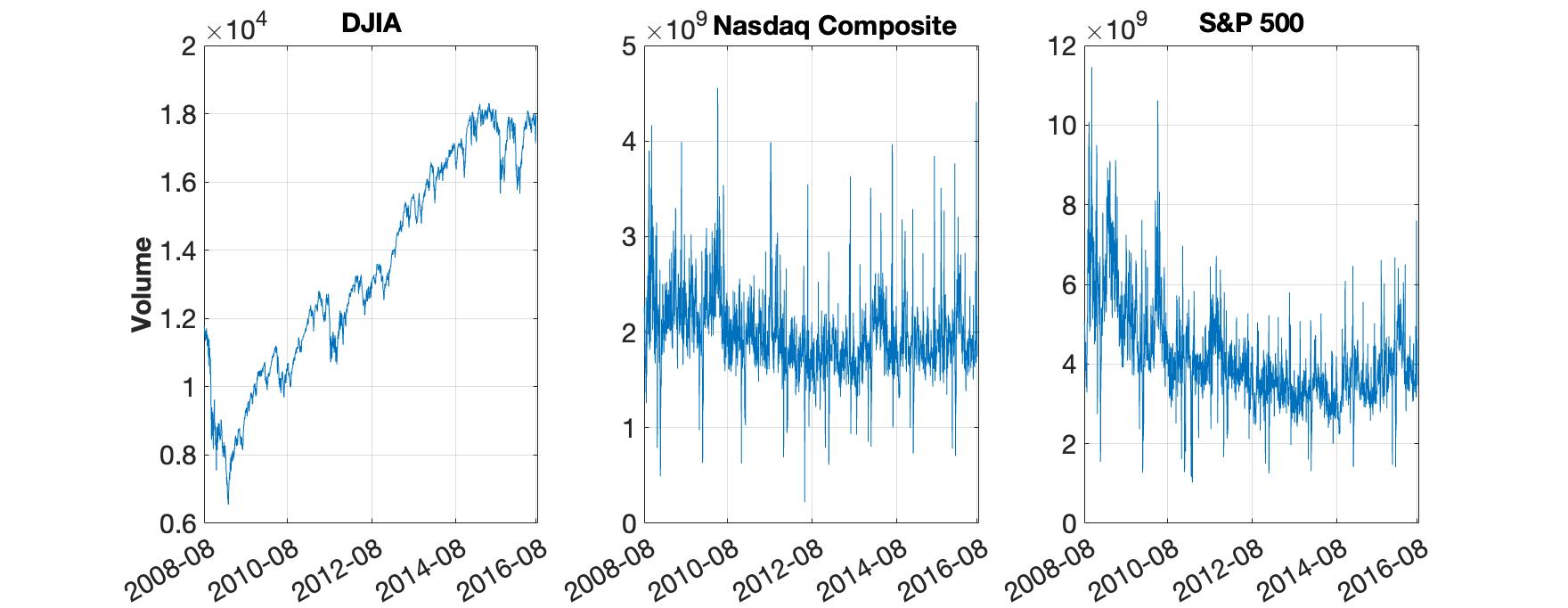
**S&P 500**

****

1. Adj close vs time



1. Volume vs time



1. Heatmap for every index

|  |
| --- |
|  |
|  |

1. Heatmap between the three indices

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

1. Wordcloud

