

Stock Market Prediction Using Twitter Data

Team members:

- Jayden Tran 16213471
- Kavın Kumar Arumugam 16262979
- Alper Erel 16254091

Motivation

The stock market is considered a complicated and nonlinear system. Now stock market prediction is recognized as an attracting point for financial investors. The historical price is considered as the main factor to predict the stock market trend. Historical data may be unstructured and need special handling on storing and processing.

The purpose of this project is to analyze the stock market data and get general insight on this data through visualization to find stock behavior and value at risk for each stock.

Significance

When it comes to investing in stocks, it is important that the investor is capable of conducting a thorough analysis. Technical analysis will allow us to do the process of forecasting future price movements based on the past price movements within the stock data. It will be very helpful for the investors to make financial decisions of buying, holding, or selling stocks. Although it is impossible to make 100% accurate predictions, it can definitely help investors anticipate the future.

Objectives

Stock Market Analysis (on Big Data Hadoop). This project is based on Big Data analysis of Stock Market. The daily commodity rates of various company shares are collected and are analyzed with the help of query method. One can easily have a market watch for any day he/she wants to look at falling in the year 2016. The user can find out his profit/loss for the share he/she owns with the help of current price rate of that share stored in our database. One can also compare different shares' highs and lows with respect to the market position. This project aims at providing simple and easy analysis of the Stock Market as per the user's requirement. The analysis result can be obtained in the form of tables, graphs and pie charts. The user gets a choice to choose the method of his analysis based on the script he selects. Relational structured data has been taken in order to complete this analysis task.

Features

Technologies:

- Java
- MySql
- Sqoop
- Hive
- R-base

1. Visualization of stock data

Dataset:

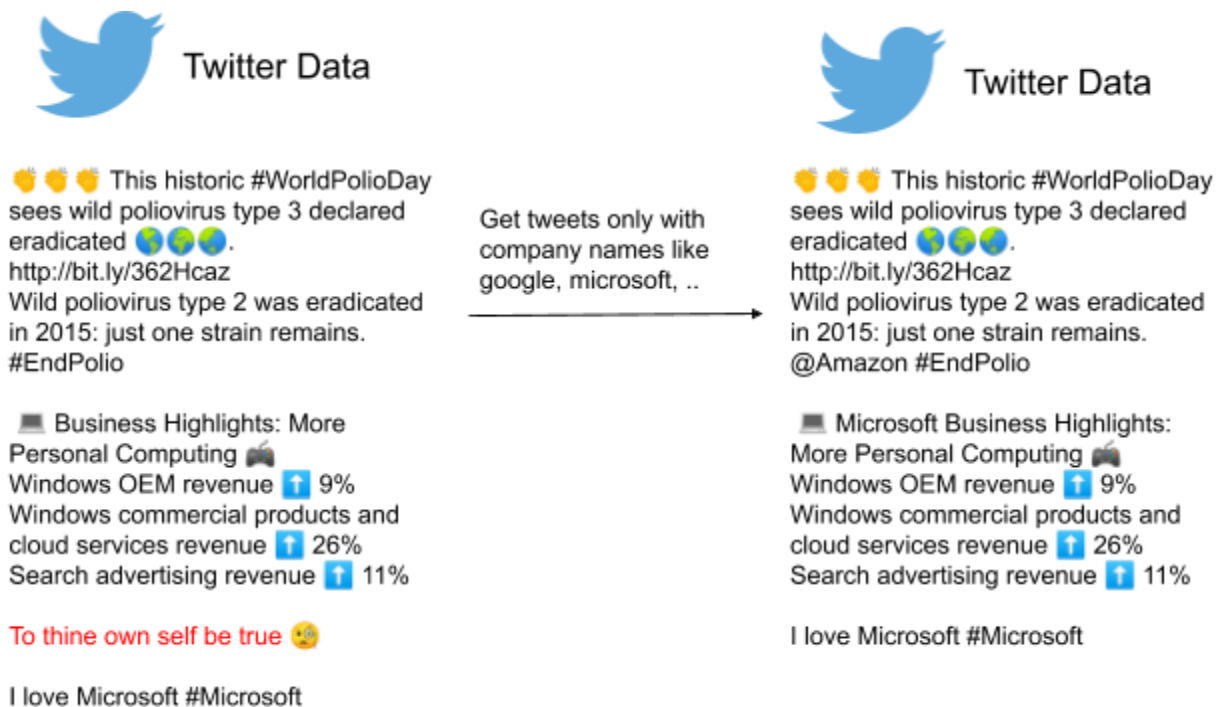
We have collected dataset for companies

Amazon	
Alphabet	

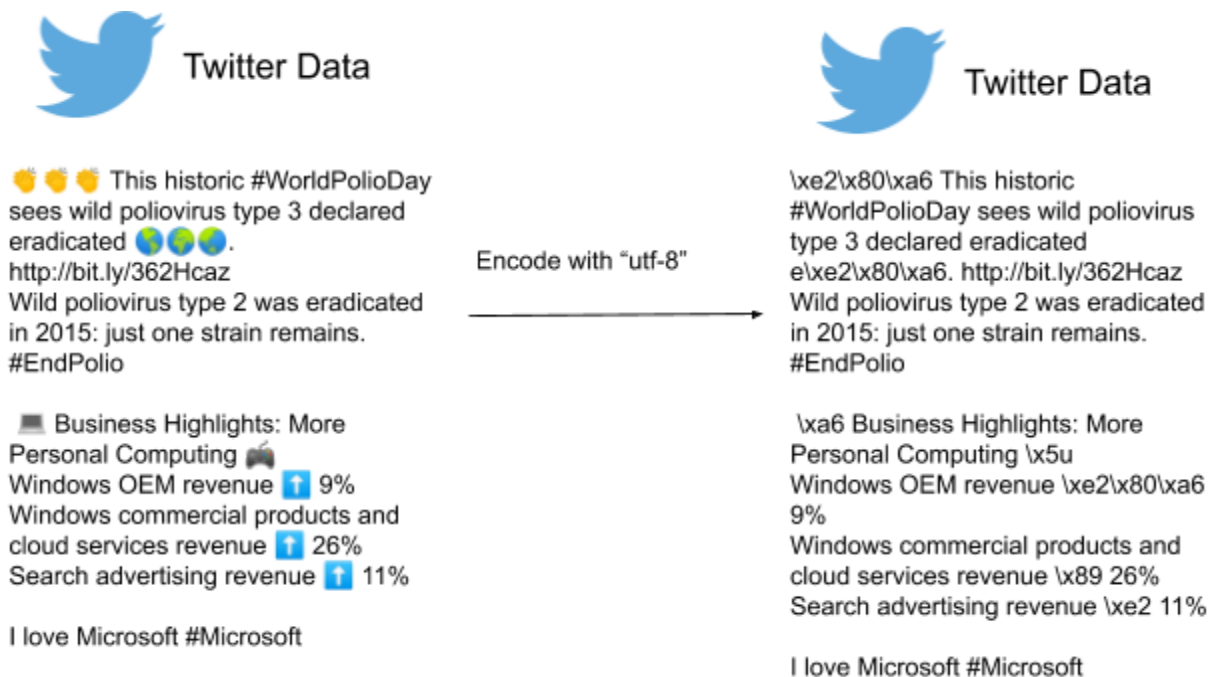
Facebook	
Intuitive Surgical	
Netflix	
Microsoft	
iRobot	
AT&T	
Verizon Communications	
Google	

Detail design of Features

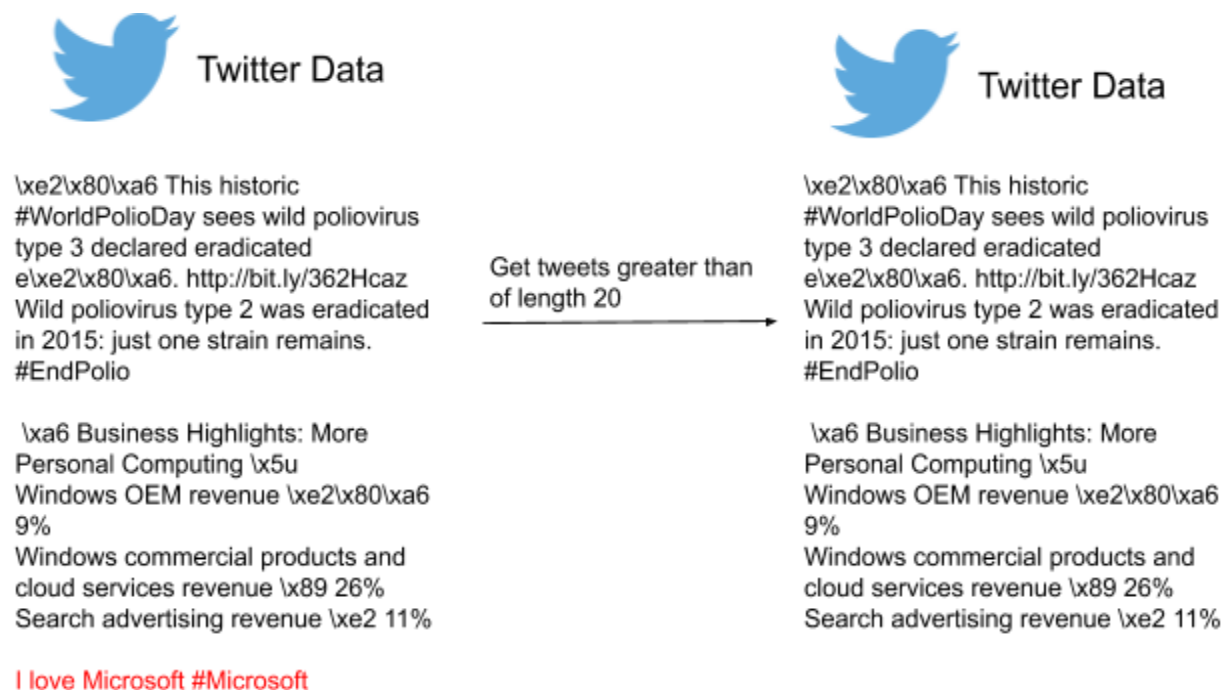
Step 1: Get tweets only with names like google and microsoft



Step 2: Encode the tweets with "utf-8"



Step 3: Remove tweets lesser than 20 words



Step 4: Classify the tweets and take only tweets related finance, business, news and science



Twitter Data

This historic
#WorldPolioDay sees wild poliovirus
type 3 declared eradicated
http://bit.ly/362Hcaz
Wild poliovirus type 2 was eradicated
in 2015: just one strain remains.
#EndPolio



Categorize
tweets using
google nlp

/Finance

filter

/Arts

/News

Business Highlights: More
Personal Computing
Windows OEM revenue
9%
Windows commercial products and
cloud services revenue 26%
Search advertising revenue 11%

Don's impact on the beginnings of
Silicon Valley and the entire tech
industry will not be soon forgotten.
Thoughts are with his family and the
entire Sequoia team. May he rest in
peace.



Twitter Data

This historic
#WorldPolioDay sees wild poliovirus
type 3 declared eradicated
http://bit.ly/362Hcaz
Wild poliovirus type 2 was eradicated
in 2015: just one strain remains.
#EndPolio

Business Highlights: More
Personal Computing
Windows OEM revenue
9%
Windows commercial products and
cloud services revenue 26%
Search advertising revenue 11%

Analysis

Implementation

Create Hive table for Microsoft file, and load the Microsoft.csv into Microsoft table.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ hive  
Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.13.0.jar!/hive-log4j.properties  
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.  
hive> CREATE TABLE Microsoft (Create_at String, id_str STRING, text STRING, user_id int, user_name string, user_screen_name string, user_location string, user_url string, user_description string, place string, entities_hashtags string, entities_url string, entities_user_mentions string) row format delimited fields terminated by ',' stored as textfile;  
OK  
Time taken: 3.215 seconds  
hive> load data local inpath '/home/cloudera/Downloads/Microsoft.csv' into table Microsoft;  
Loading data to table default.microsoft  
Table default.microsoft stats: [numFiles=1, totalSize=97788053]  
OK  
Time taken: 1.434 seconds  
hive> select * from Microsoft limit 10;  
OK  
b'2019-10-26 20:34:27' b'1188192186623414272' "b'RT @jimsciutto: Given Pentagon  
n\xe2\x80\x99s decision Friday to choose Microsoft over Amazon NULL b'318082  
7364' b'Elaine Guthrie' b'ElaineEguthrie1' "b'Fort Collins CO" b  
'None' b'Trauma Nurse' b'None' b'[]'
```

Show first 10 row:

```
hive> select * from Microsoft limit 10;  
OK  
b'2019-10-26 20:34:27' b'1188192186623414272' "b'RT @jimsciutto: Given Pentagon  
n\xe2\x80\x99s decision Friday to choose Microsoft over Amazon NULL b'318082  
7364' b'Elaine Guthrie' b'ElaineEguthrie1' "b'Fort Collins CO" b  
'None' b'Trauma Nurse' b'None' b'[]'  
b'2019-10-26 20:34:21' b'1188192161185046528' b'Congratulations #Microsoft fo  
r winning Pentagon\xe2\x80\x99s historic cloud-computing contract of worth USD 1  
0b \xf0\x9f\x91\x8d\xf0\x9f\x8f\xbc' NULL b'Iftikhar Alam' b'imifti  
kharalam' b'Lahore' b'https://t.co/nDREFBj9W6' b'Journalist | R  
eligion.Politics. Indo-Pak. Agriculture. @diplomat_APAC @theprintindia @nayadaur  
pk @the_nation' b'None' "b'["text': 'Microsoft' 'indices': [16 26]]]"  
b'2019-10-26 20:34:21' b'1188192159591256076' "b'RT @MSFTResearch: Ada is a co  
llaboration by architectural designer @jennysabin and Microsoft Research N  
ULL material inno\xe2\x80\xa6" b'138140384' b'GK' b'gktweets101' "  
b'London England'" b'None' b'Everything else' b'None'  
b'2019-10-26 20:34:14' b'1188192130663104515' "b'RT @jimsciutto: Given Pentagon  
n\xe2\x80\x99s decision Friday to choose Microsoft over Amazon NULL b'167450  
0558' b'Karen Babineau' b'airlift1300' "b'Florida USA" b'None'b  
'No lists! #TheResistance #TRUMPRUSSIA #trumpdossier #ImpeacTrump I am quiet til  
l I have something to say. Love my cat. wish I could travel & meet more people.'  
b'None' b'[]'  
b'2019-10-26 20:34:13' b'1188192127945101312' "b'RT @HAMSTER_Corp: ACA NEOGEO  
PUZZLE BOBBLE is now available on Windows 10 PC ! It's an action puzzle game re  
leased by Taito in 1994. Bub an\xe2\x80\xa6" NULL b'\xe3\x81\x97\xe3\x81\x  
8a\xe3\x82\x93\xe3\x83\x91\xe3\x83\x91' b'sions_papa' b'\xe5\xa4\xa7\xe9\x98\x  
aa\xe5\xba\x9c' b'https://t.co/68XXYKXzRx' b'1967\xe5\xb9\xb4\xe8\xa3\xbd\x  
e3\x81\xa7\xe3\x81\x99\xe3\x80\xe2\xe6\xb5\xe8\xa1\xe8c\xe3\x81\x99\xe3\x82\x  
8b\xe3\x81\xa8\xe8\x88\xe5\x91\xb3\xe3\x81\xaa\xe3\x81\x8f\xe3\x81\xaa\xe3\x
```

Show column into a table form with column name = true:


```

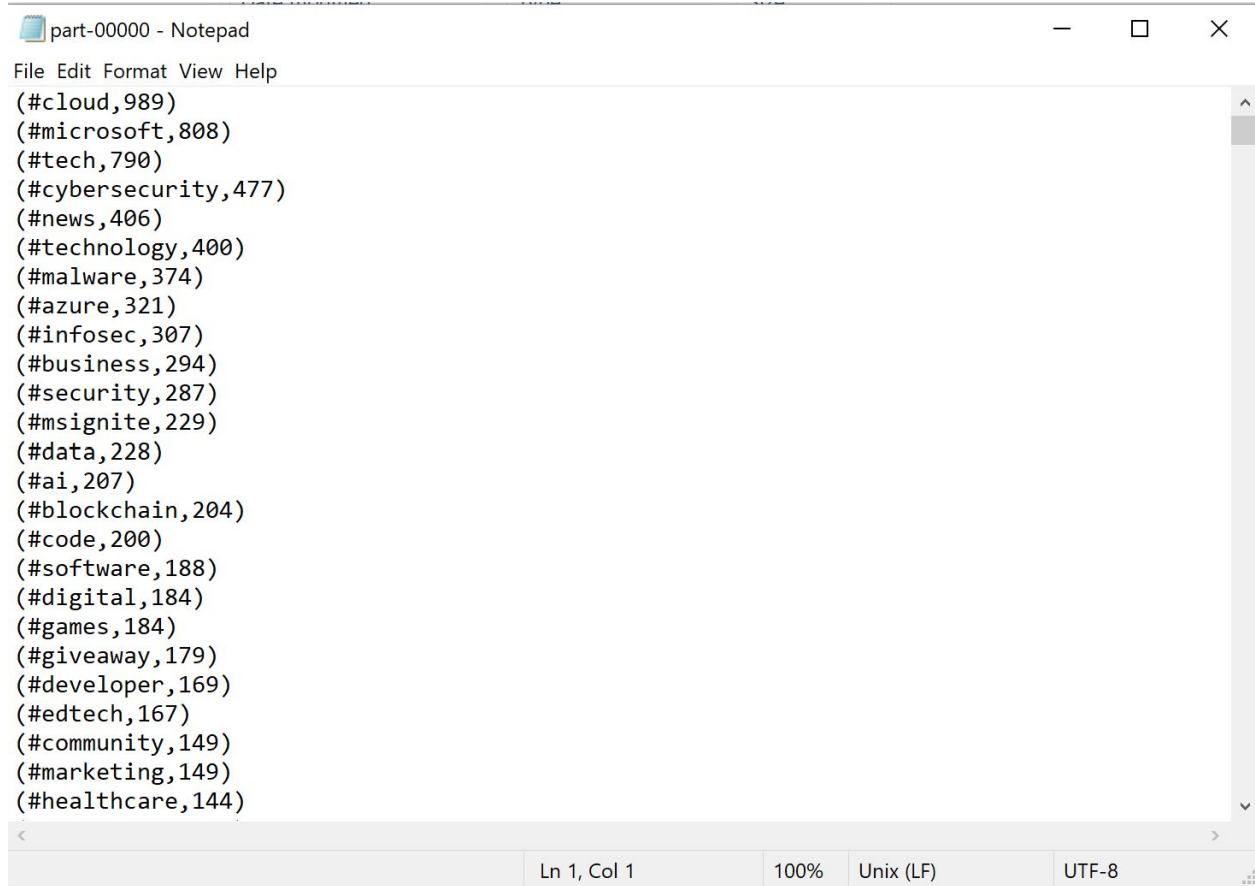
hive> Select Create_at, id_str, text, user_id, user_name, user_screen_name FROM Microsoft ORDER BY "Create_at" limit 10;
Query ID = cloudera_20191027131919_0db51146-f468-4209-b6b0-2a1ce1217998
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1572203653193_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1572203653193_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1572203653193_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-10-27 13:19:35,278 Stage-1 map = 0%, reduce = 0%
2019-10-27 13:19:45,000 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.65 sec
2019-10-27 13:19:56,593 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.98 sec
MapReduce Total cumulative CPU time: 8 seconds 980 msec

OK
create_at      id_str      text      user_id user_name      user_screen_name
b'2019-10-26 20:33:57' b'1188192059229904899' "b""Who else out there is missing out on watching raw NULL
nxt      ROH
b'2019-10-26 20:33:58' b'1188192066070822913' b'Microsoft beats Amazon to win the Pentagon\xe2\x80\x99s $10 bil
lion JEDI cloud contract https://t.co/0hJFMNk0Xa @Verge' NULL b'\xc5\x81lukasz Wiz\xc5\x82a' b'LukaszW
izla'
b'2019-10-21 13:26:24' b'1186272522955874304' "b""RT @ParallelsRAS: We hope to see you at #MSIgnite in 2 weeks!
Find the Parallels team at booth 2626 where we'll be showcasing Parallels RAS\xe2\x80\xa6"" NULL b'K1 4mos
on' b'k14mo'
b'2019-10-26 20:34:02' b'1188192082927579141' b'RT @TheAmyCode: Working in the cloud infra industry is complica
ted. I get that you won\xe2\x80\x99t have control over how the company uses software.\xe2\x80\xa6' NULL b
'Dami\xc3\xadn Garc\xc3\xada S.' b'Damian Garcias'
b'2019-10-26 20:34:04' b'1188192089592487938' b'@matthewsmall So the JEDI staff who were fortuitously employed
by AWS ... \n\nDo they keep their new jobs ... Or join microsoft?' NULL b'Damo' b'MajorDamo'
b'2019-10-26 20:34:13' b'1188192127945101312' "b""RT @HAMSTER Corp: ACA NEOGEO PUZZLE BOBBLE is now available o
n Windows 10 PC ! It's an action puzzle game released by Taito in 1994. Bub an\xe2\x80\xa6"" NULL b'\xe3\x8
1\x97\xe3\x81\x8a\xe3\x82\x93\xe3\x83\x91\xe3\x83\x91' b'sions_papa'
b'2019-10-26 20:34:14' b'1188192130663104515' "b'RT @jimsciutto: Given Pentagon\xe2\x80\x99s decision Friday to
choose Microsoft over Amazon NULL b'1674500558' b'Karen Babineau'
b'2019-10-26 20:34:21' b'1188192159591256076' "b'RT @MSFTResearch: Ada is a collaboration by architectural desi
gner @jennysabin and Microsoft Research NULL material inno\xe2\x80\xa6"" b'138140384'
b'2019-10-21 13:26:34' b'1186272569030303744' b'Win Power BI Swag with Community Kudopalooza! #PowerBI https:/
/t.co/H2Zxv1x0TK' NULL b'Katie Novotny' b'KatrinaNovotny'
b'2019-10-26 20:34:27' b'1188192186623414272' "b'RT @jimsciutto: Given Pentagon\xe2\x80\x99s decision Friday to
choose Microsoft over Amazon NULL b'3180827364' b'Elaine Guthrie'
Time taken: 34.599 seconds, Fetched: 10 row(s)
hive> █

```

Preliminary Results

- Upon evaluating the Microsoft dataset, we found out the most popular hashtags which define Microsoft including, but not limited to:
 - #cloud
 - #tech
 - #cybersecurity etc.
- These are the top words that people talk about that describes Microsoft.



```
part-00000 - Notepad
File Edit Format View Help
(#cloud,989)
(#microsoft,808)
(#tech,790)
(#cybersecurity,477)
(#news,406)
(#technology,400)
(#malware,374)
(#azure,321)
(#infosec,307)
(#business,294)
(#security,287)
(#msignite,229)
(#data,228)
(#ai,207)
(#blockchain,204)
(#code,200)
(#software,188)
(#digital,184)
(#games,184)
(#giveaway,179)
(#developer,169)
(#edtech,167)
(#community,149)
(#marketing,149)
(#healthcare,144)
Ln 1, Col 1 100% Unix (LF) UTF-8
```

Project Management

References/Bibliography

- 1) <https://m.benzinga.com/article/9602734>
- 2) <https://www.investopedia.com/terms/s/stock-analysis.asp>
- 3) <https://cleartax.in/s/stock-market-analysis>

