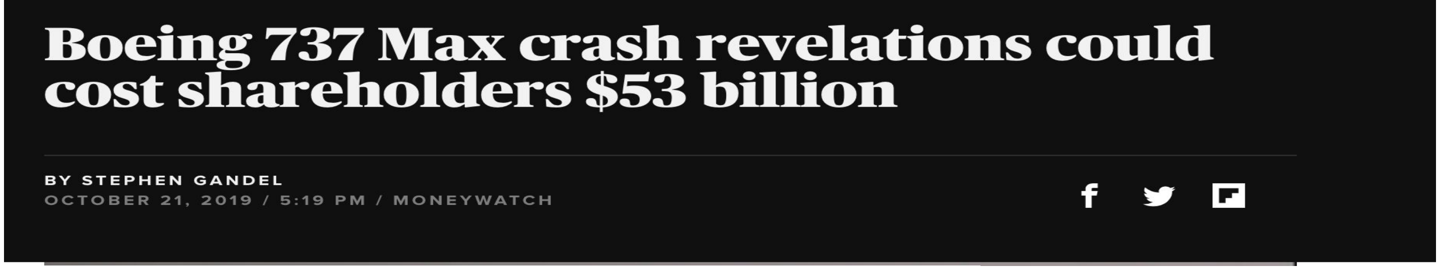
Stock market Prediction based on Daily News Headlines

Shaohang Hao, Weikun Hu,
Ji Peng and Ruiyu Zeng

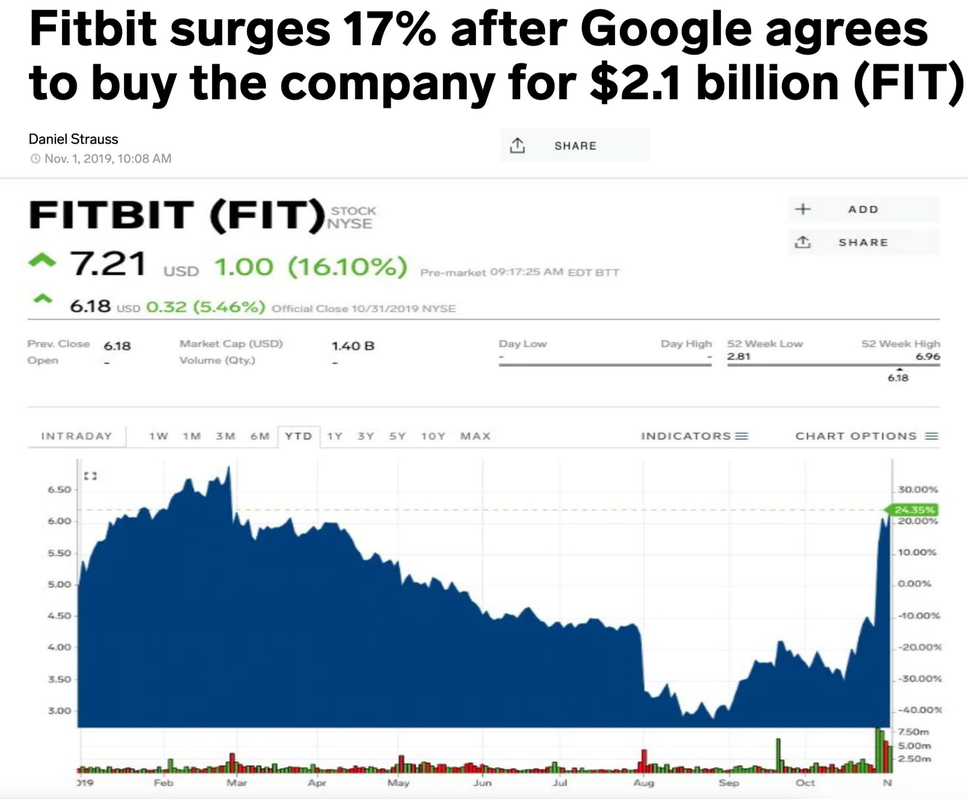
CSE 538

2019/11/06



Boeing 737 Max crash revelations could cost shareholders $53 billion

BY STEPHEN GANDEL
OCTOBER 21, 2019 / 5:19 PM / MONEYWATCH



Fitbit surges 17% after Google agrees to buy the company for $2.1 billion (FIT)

Daniel Strauss
Nov. 1, 2019, 10:08 AM

SHARE

FITBIT (FIT) STOCK NYSE

ADD
SHARE

7.21 USD 1.00 (16.10%) Pre-market 09:17:25 AM EDT BTT

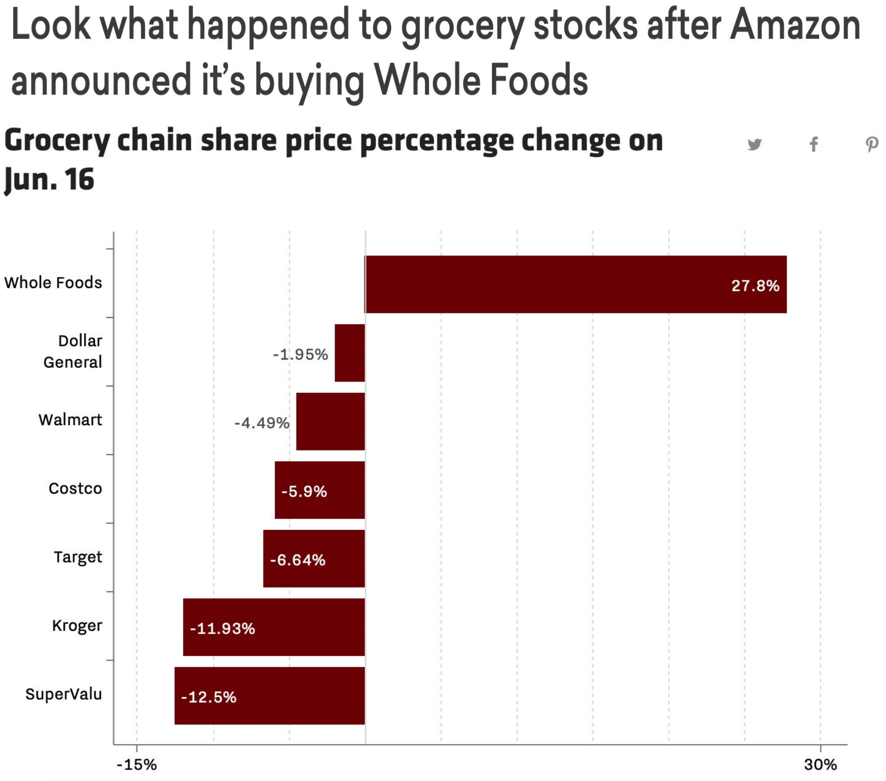
6.18 USD 0.32 (5.46%) Official Close 10/31/2019 NYSE

Prev. Close	6.18	Market Cap (USD)	1.40 B
Open	-	Volume (Qty.)	-

Day Low	Day High	52 Week Low	52 Week High
-	-	2.81	6.96

6.18

INTRADAY 1W 1M 3M 6M YTD 1Y 3Y 5Y 10Y MAX INDICATORS CHART OPTIONS



Look what happened to grocery stocks after Amazon announced it's buying Whole Foods

Grocery chain share price percentage change on Jun. 16

Background

Use Daily News to Predict Stock Market Performance

- **News data** was obtained from Reddit WorldNews Channel (/r/worldnews). Top 25 headlines were voted by reddit users for a single date. (Range: 2008-06-08 to 2016-07-01)
- **Stock data**: Dow Jones Industrial Average (DJIA) is used as the label to supervise model training . (Range: 2008-08-08 to 2016-07-01)
- Data from 2008-08-08 to 2014-12-31 will be used as Training Set, and Test Set is then the following two years data (from 2015-01-02 to 2016-07-01). This is roughly a 80%/20% split.
- AUC will be used as the evaluation metrics

Use Case

User:

People who buy stock

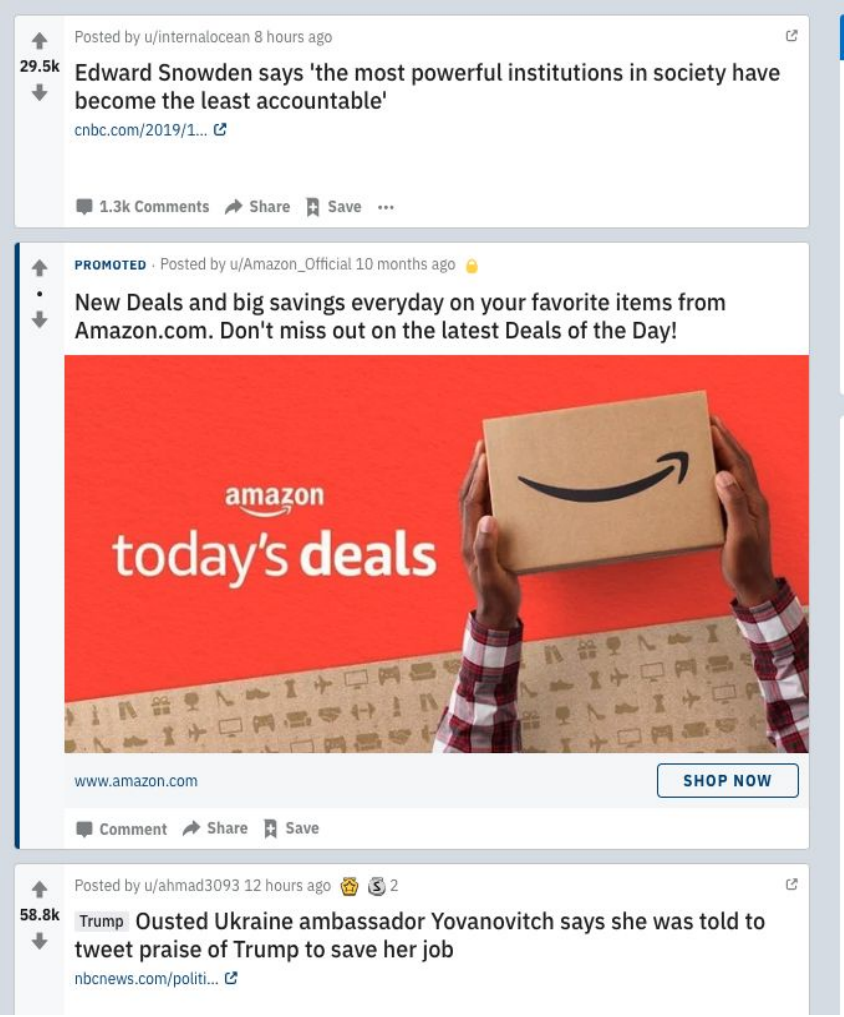
(i.e. investors and mutual fund managers)

Input:

Top 25 headlines of daily news on Reddit

Output:

DJIA (Dow Jones industrial average) is increased (1) or decreased (0) for a single day



29.5k

Posted by u/internalocean 8 hours ago

Edward Snowden says 'the most powerful institutions in society have become the least accountable'

cnbc.com/2019/1...

1.3k Comments Share Save ...

PROMOTED · Posted by u/Amazon_Official 10 months ago

New Deals and big savings everyday on your favorite items from Amazon.com. Don't miss out on the latest Deals of the Day!

www.amazon.com

SHOP NOW

Comment Share Save

58.8k

Posted by u/ahmad3093 12 hours ago 2

Trump Ousted Ukraine ambassador Yovanovitch says she was told to tweet praise of Trump to save her job

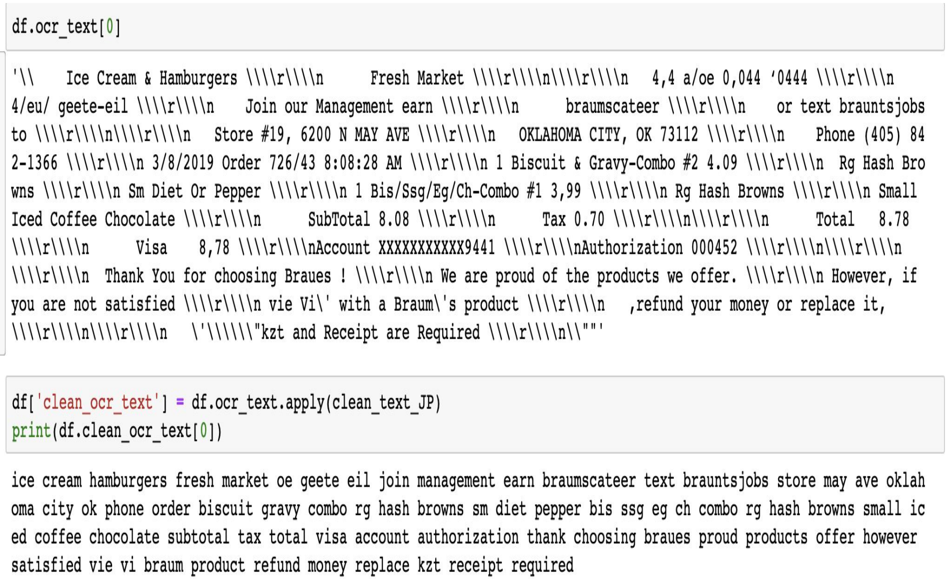
nbcnews.com/politi...

Example: News headline on Reddit

Python Libraries for Natural Language Processing

NLTK (Natural Language Toolkit)

- Preprocessing: tokenization, stopwords, stemming



```
df.ocr_text[0]
```

```
'\\    Ice Cream & Hamburgers \\\\r\\\\n      Fresh Market \\\\r\\\\n\\\\r\\\\n   4,4 a/oe 0,044 ‘0444 \\\\r\\\\n
4/eu/ geete-eil \\\\r\\\\n    Join our Management earn \\\\r\\\\n      braumscateer \\\\r\\\\n    or text brauntsjobs
to \\\\r\\\\n\\\\r\\\\n   Store #19, 6200 N MAY AVE \\\\r\\\\n   OKLAHOMA CITY, OK 73112 \\\\r\\\\n    Phone (405) 84
2-1366 \\\\r\\\\n 3/8/2019 Order 726/43 8:08:28 AM \\\\r\\\\n 1 Biscuit & Gravy-Combo #2 4.09 \\\\r\\\\n  Rg Hash Bro
wns \\\\r\\\\n Sm Diet Or Pepper \\\\r\\\\n 1 Bis/Ssg/Eg/Ch-Combo #1 3,99 \\\\r\\\\n Rg Hash Browns \\\\r\\\\n Small
Iced Coffee Chocolate \\\\r\\\\n      SubTotal 8.08 \\\\r\\\\n      Tax 0.70 \\\\r\\\\n\\\\r\\\\n      Total   8.78
\\\\r\\\\n      Visa    8,78 \\\\r\\\\nAccount XXXXXXXXXXX9441 \\\\r\\\\nAuthorization 000452 \\\\r\\\\n\\\\r\\\\n
\\\\r\\\\n  Thank You for choosing Braues ! \\\\r\\\\n We are proud of the products we offer. \\\\r\\\\n However, if
you are not satisfied \\\\r\\\\n vie Vi\' with a Braum\'s product \\\\r\\\\n   ,refund your money or replace it,
\\\\r\\\\n\\\\r\\\\n   \'\\\\\\"kzt and Receipt are Required \\\\r\\\\n\\""'
```

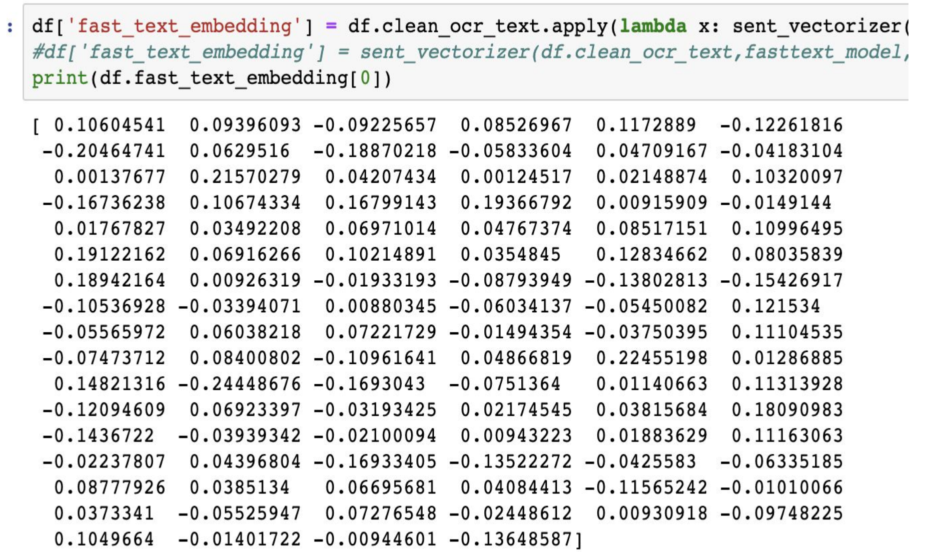
```
df['clean_ocr_text'] = df.ocr_text.apply(clean_text_JP)
print(df.clean_ocr_text[0])
```

```
ice cream hamburgers fresh market oe geete eil join management earn braumscateer text brauntsjobs store may ave oklah
oma city ok phone order biscuit gravy combo rg hash browns sm diet pepper bis ssg eg ch combo rg hash browns small ic
ed coffee chocolate subtotal tax total visa account authorization thank choosing braues proud products offer however
satisfied vie vi braum product refund money replace kzt receipt required
```

Scikit-learn

- Word embedding (Vectorization)

  TfidfVectorizer,CountVectorizer



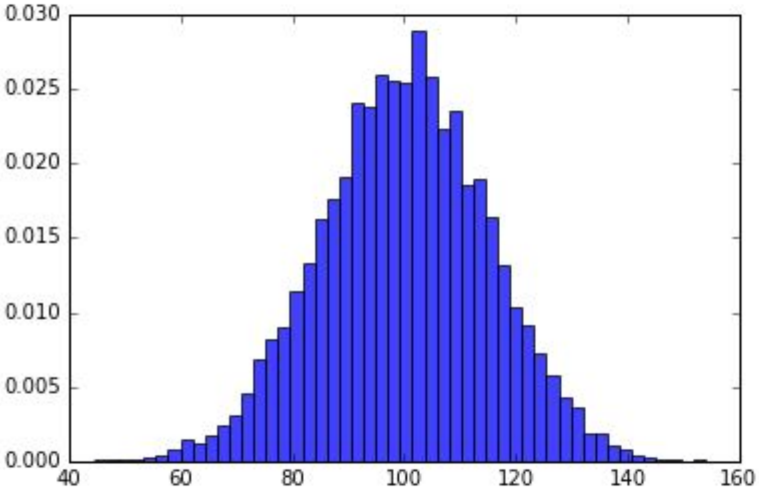
```
df['fast_text_embedding'] = df.clean_ocr_text.apply(lambda x: sent_vectorizer(
#df['fast_text_embedding'] = sent_vectorizer(df.clean_ocr_text,fasttext_model,
print(df.fast_text_embedding[0])
```

```
[ 0.10604541  0.09396093 -0.09225657  0.08526967  0.1172889  -0.12261816
 -0.20464741  0.0629516  -0.18870218 -0.05833604  0.04709167 -0.04183104
  0.00137677  0.21570279  0.04207434  0.00124517  0.02148874  0.10320097
 -0.16736238  0.10674334  0.16799143  0.19366792  0.00915909 -0.0149144
  0.01767827  0.03492208  0.06971014  0.04767374  0.08517151  0.10996495
  0.19122162  0.06916266  0.10214891  0.0354845   0.12834662  0.08035839
  0.18942164  0.00926319 -0.01933193 -0.08793949 -0.13802813 -0.15426917
 -0.10536928 -0.03394071  0.00880345 -0.06034137 -0.05450082  0.121534
 -0.05565972  0.06038218  0.07221729 -0.01494354 -0.03750395  0.11104535
 -0.07473712  0.08400802 -0.10961641  0.04866819  0.22455198  0.01286885
  0.14821316 -0.24448676 -0.1693043  -0.0751364   0.01140663  0.11313928
 -0.12094609  0.06923397 -0.03193425  0.02174545  0.03815684  0.18090983
 -0.1436722  -0.03939342 -0.02100094  0.00943223  0.01883629  0.11163063
 -0.02237807  0.04396804 -0.16933405 -0.13522272 -0.0425583  -0.06335185
  0.08777926  0.0385134   0.06695681  0.04084413 -0.11565242 -0.01010066
  0.0373341  -0.05525947  0.07276548 -0.02448612  0.00930918 -0.09748225
  0.1049664  -0.01401722 -0.00944601 -0.13648587]
```

Python Libraries for Visualization

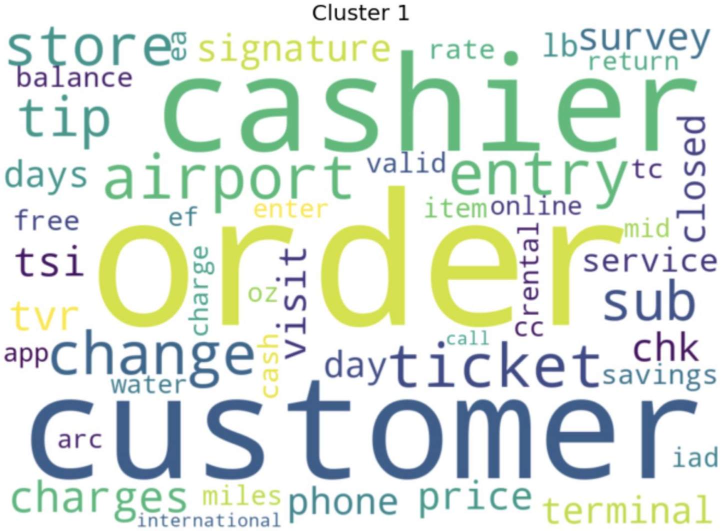
Matplotlib:

- Primarily used for 2D visual representation of data distribution



Word Cloud

- Size of each word indicates its frequency or importance.



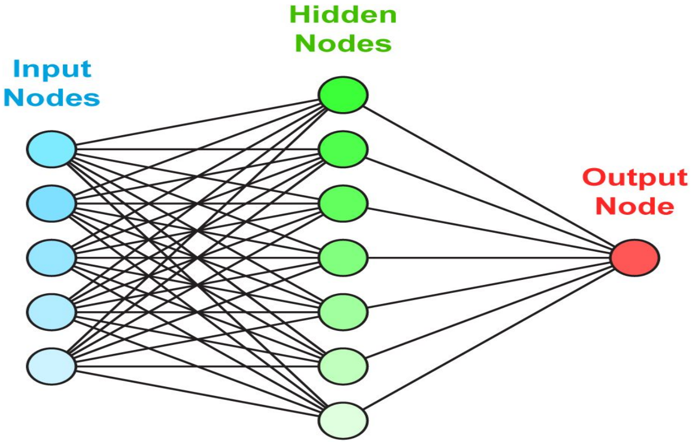
Python Libraries for Machine Learning and Deep Learning

Scikit-learn:

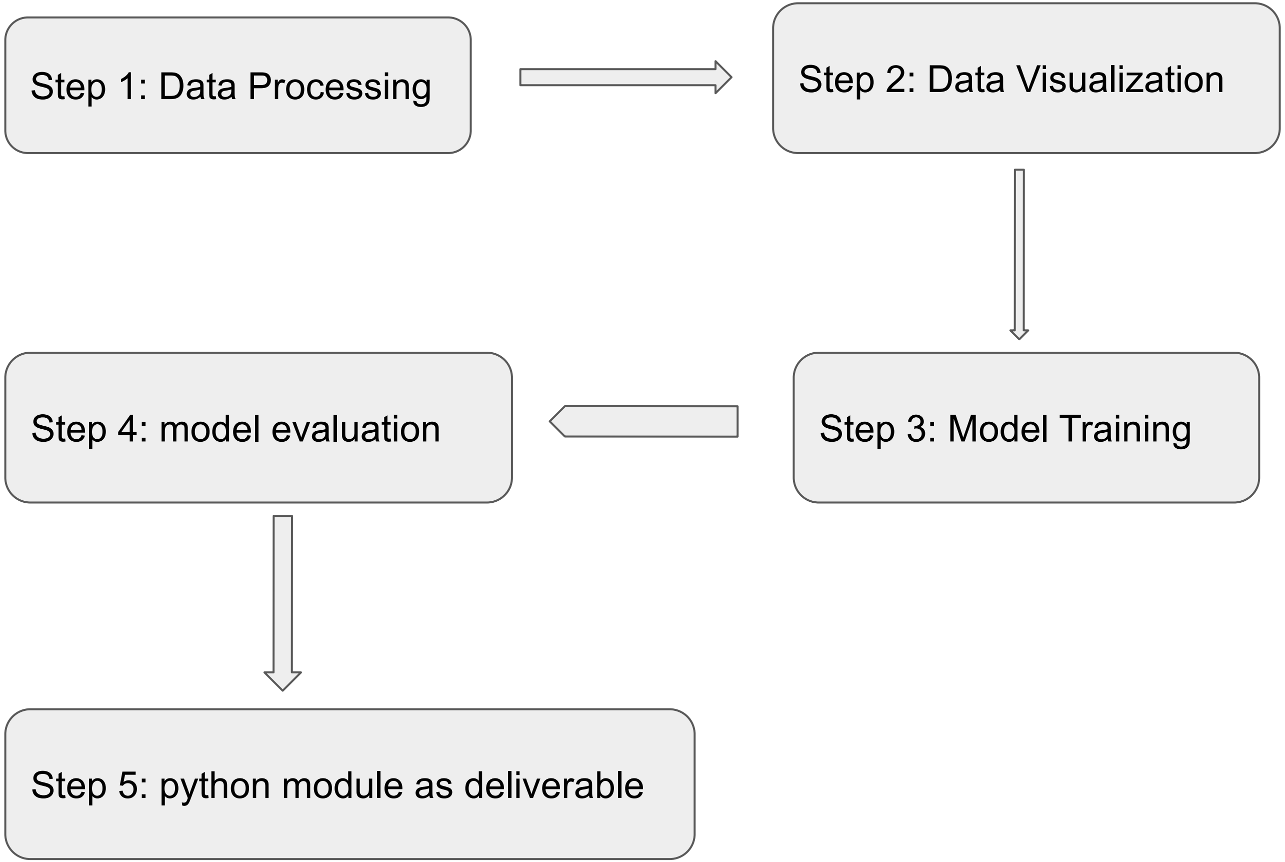
- Functions for building machine learning models (ex. Logistic regression, support vector machine, and random forest)

Keras:

- Python deep learning library
- Capable of running on neural networks (ex. Tensorflow)



Timeline of project



Step 1: Data Processing

Step 2: Data Visualization

Step 3: Model Training

Step 4: model evaluation

Step 5: python module as deliverable

Thank you!

Reference

Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved [Date You Retrieved This Data] from https://www.kaggle.com/aaron7sun/stocknews.

https://www.kaggle.com/lseiyjg/use-news-to-predict-stock-markets

https://www.kaggle.com/shreyams/stock-price-prediction-94-xgboost

Requirements

The technology review is a group presentation. It should be about 10-15 minutes in length. The presentation should address the following:

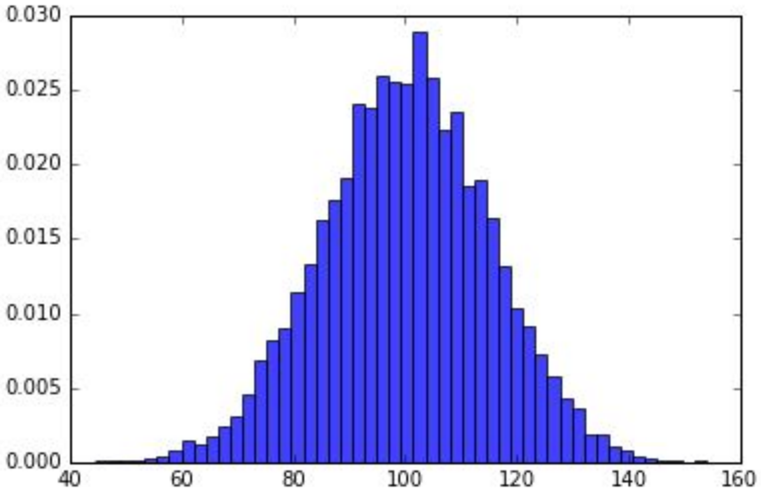
- Brief background on the problem you're solving to motivate a technology for which you need a python library (e.g., interactive maps).
- One slide desciption of a use case in which the technology is required.
- One slide that describes at least two python libraries that potentially address your technology requirement.
- For each libarary, show a simple example of using it to implement the use case described above. This means that you will need to install each of the python libraries and attempt to use them.
- One slide side-by-side comparisons of the technologies.

http://uwseds.github.io/projects.html

Python Libraries for Visualization

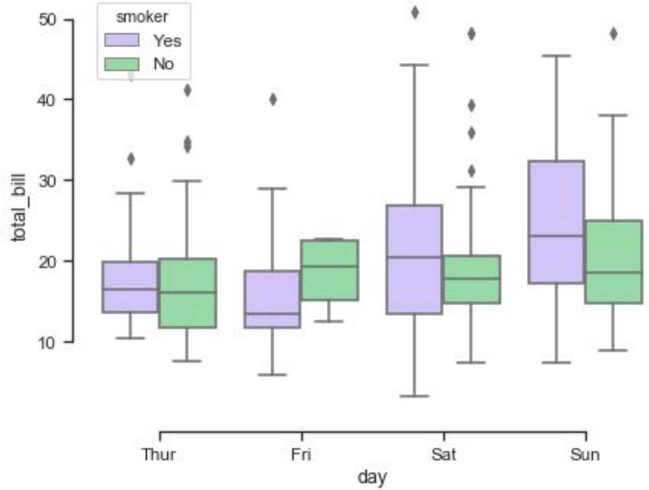
Matplotlib:

- Primarily used for 2D visual representation of data distribution



Seaborn:

- Can be used as a complement for Matplotlib
- More versatile and customizable than Matplotlib


smoker
Yes
No
total_bill
Thur
Fri
Sat
Sun
day