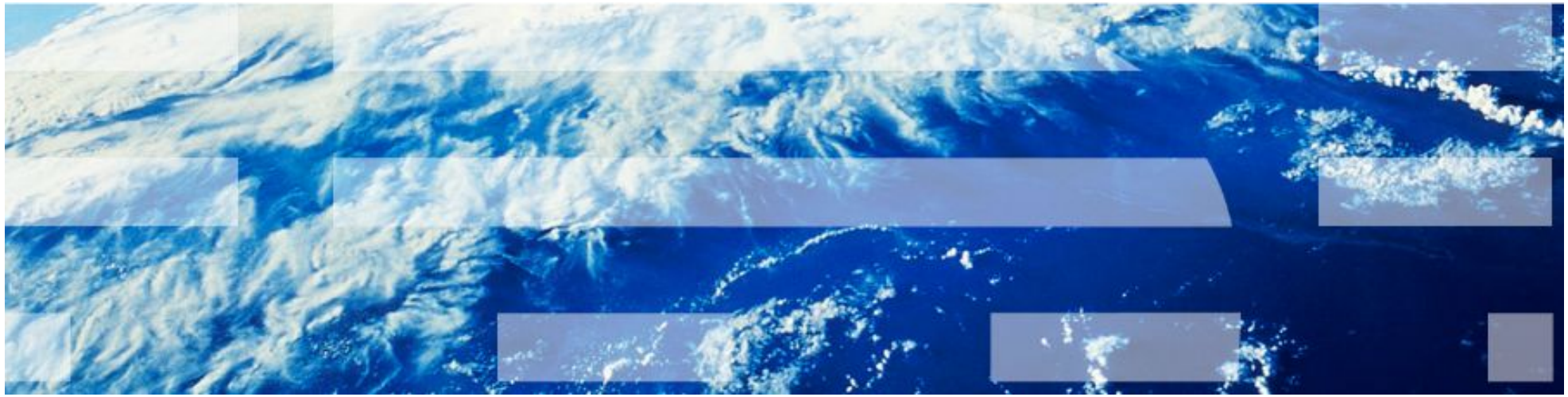
E6893 Big Data Analytics:

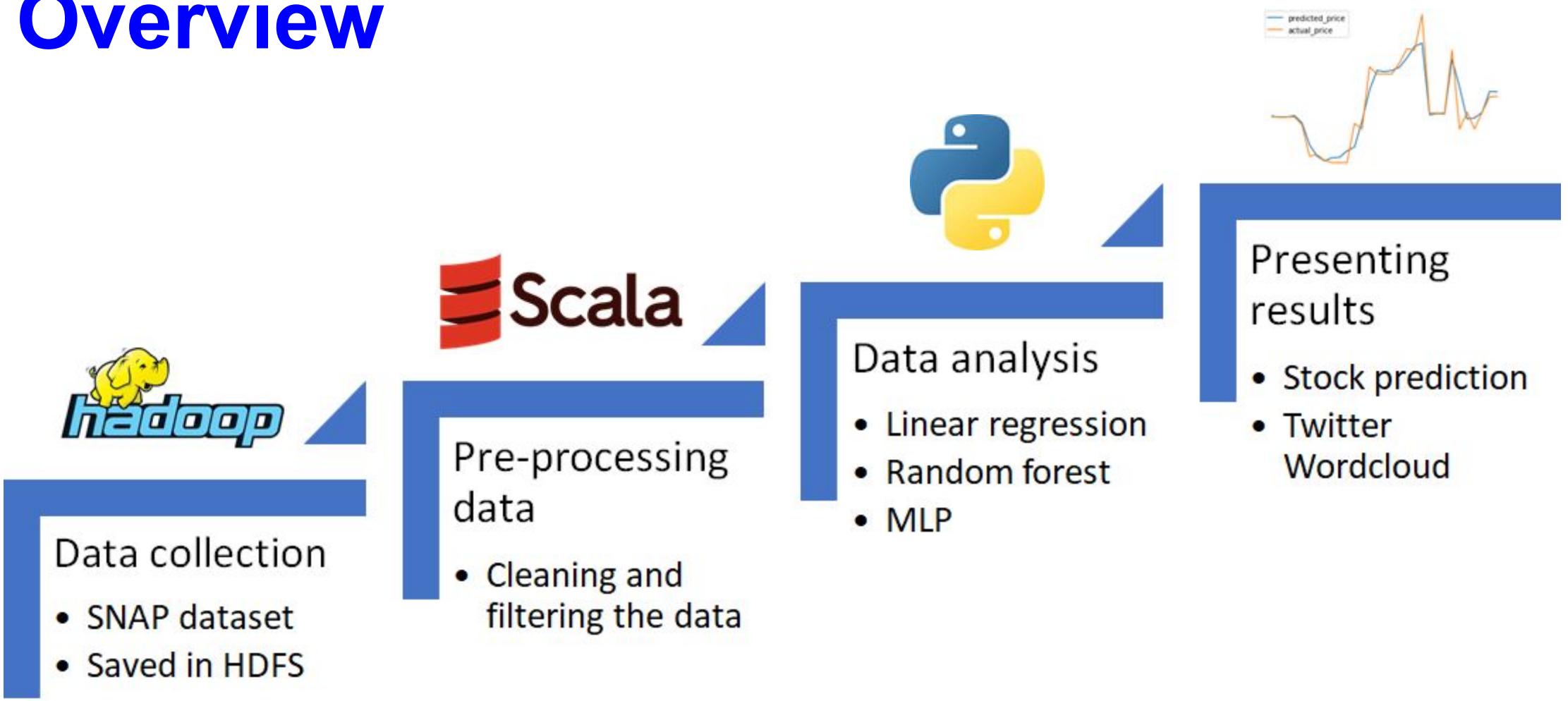
Prediction of Stock Trend with Media Sentiment Analysis

201812-30

JILIANG MA (jm4750)
LONG JIAO (lj2463)
QINYUAN WEI (qw2264)



Overview



Technologies Used

- **Tools**

  Hadoop

  Spark

  Jupyter Notebook

  sklearn(MLP, Random Forest, Linear Regression)

  D3

  Node.js

Dataset

- Historical stock price dataset
  - Microsoft,Nasdaq
  - Yahoo Finance
- Twitter7 dataset --from Stanford Large Network Data Collection(SNAP)
  - **25GB**
  - From June to December in 2009
  - **476 million** tweets collected
  - Includes **17,069,982** users, **476,553,560** tweets, **181,611,080** URLs, **49,293,684** Hashtags and **71,835,017** retweets.

Algorithm

- Data Pre-processing:
  - Read data: (scala)
    - load into HDFS
    - read in spark:
  - Data parsing & filtering: (scala)
    - only keep tweets and dates
    - filter all data by target trademarks
  - Combine data by date: (python)
  - Merge tweets with stock prices:
    - fill the blank data

```
T       2009-11-17 21:14:24
U       http://twitter.com/julianapanek
W       Today's weather is good
```

```
scala> dfTweets.show
+--------+--------------------+----------+---------+
|    Date|               Tweet|Stock code|trademark|
+--------+--------------------+----------+---------+
|20090611|Microsoft Outlook...|      MSFT|Microsoft|
|20090611|#Java Interoperab...|      MSFT|Microsoft|
|20090611|"Morro" é o nome ...|      MSFT|Microsoft|
|20090611|@methedivine well...|      MSFT|     xbox|
```

	Date	Tweet
0	20090701	Yahoo CEO: We Have Nothing To Say About Micros...
1	20090702	Microsoft's \Pink\" smartphone to be Microsof...
2	20090703	RT @Bob_do: Microsoft Changing Users' Default...

	Date	Tweet	Prices
0	20091001	Very funny Microsoft! Now can I use the Windo...	24.88
1	20091002	Played against getonyourkneesJR and he had a ...	24.96
2	20091003	Comment on Windows 7 RC ISO Official Microsof...	24.96

Algorithm

- Data processing and analysis
  - Sentiment Analysis

    We use vader_lexicon in NLTK to execute the sentiment analysis, get the polarity for the Tweets of each day, which is how much positive, negative, neutral the Tweets are.

	Date	Tweet	Prices	Negative	Neutral	Positive
0	20090701	I only use my credit card online, recent trans...	24.040001	0.058	0.798	0.144
1	20090702	Microsoft's \Pink\" smartphone to be Microsof...	23.370001	0.042	0.842	0.116
2	20090703	RT @Bob_do: Microsoft Changing Users' Default...	23.370001	0.071	0.825	0.104
3	20090704	RT @arturogoga La publicidad de Microsoft, ca...	23.370001	0.038	0.811	0.151
4	20090705	Get Rich on Microsoft Search engine Bing http...	23.370001	0.044	0.836	0.119

Algorithm

- Data processing and analysis
  - For data processing, the parameters in the machine learning include polarity of the Tweets, the close stock price of pre-day, the close Nasdaq Composite Index of pre-day.
  - We use Linear Regression, Random Forest and MLP in sklearn in Python to process the data, and compare the result of these three algorithm.

Results

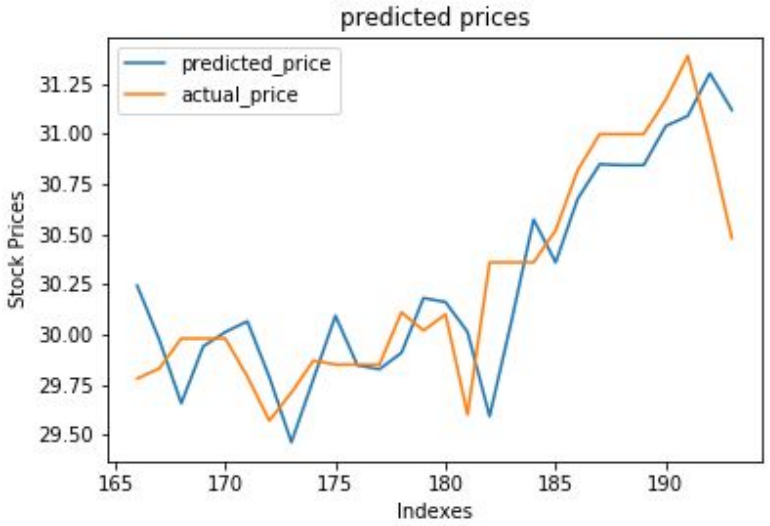
- Data visual
  - Draw word cloud of each month
  - Represent what people like to talk about Microsoft and its trademarks at that time

Word Cloud for December

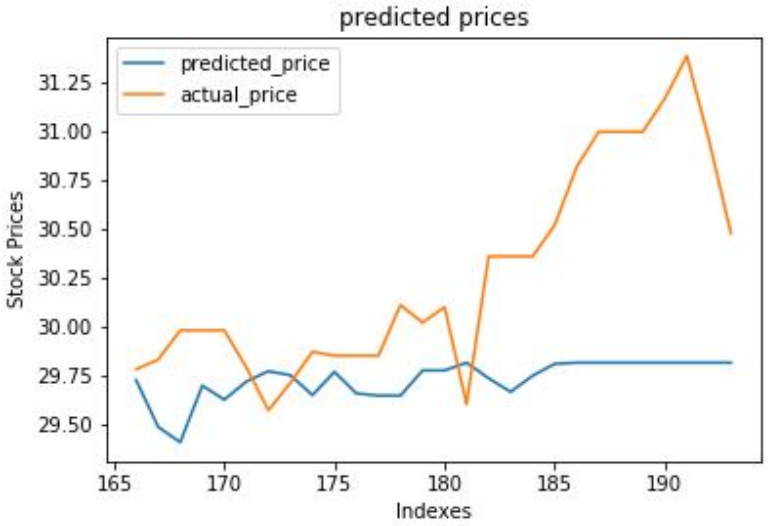
Results

- Predict result (using 6-11th months as training set, 12th month as testing set)

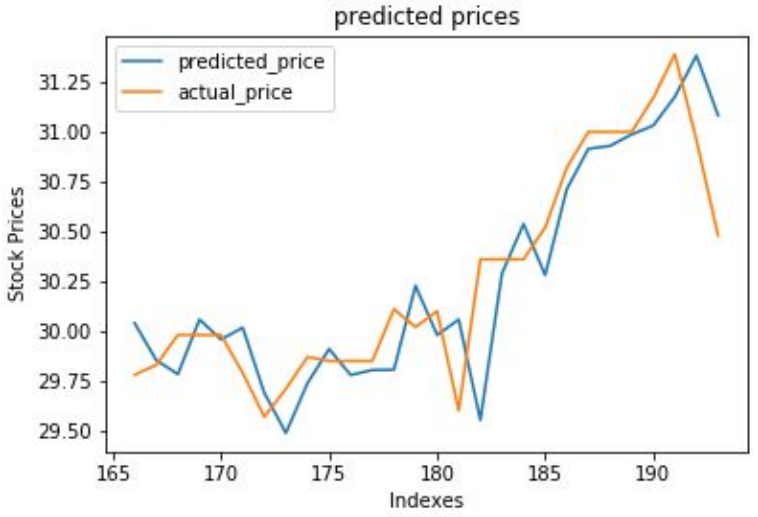
Linear Regression



Random Forest



MLP



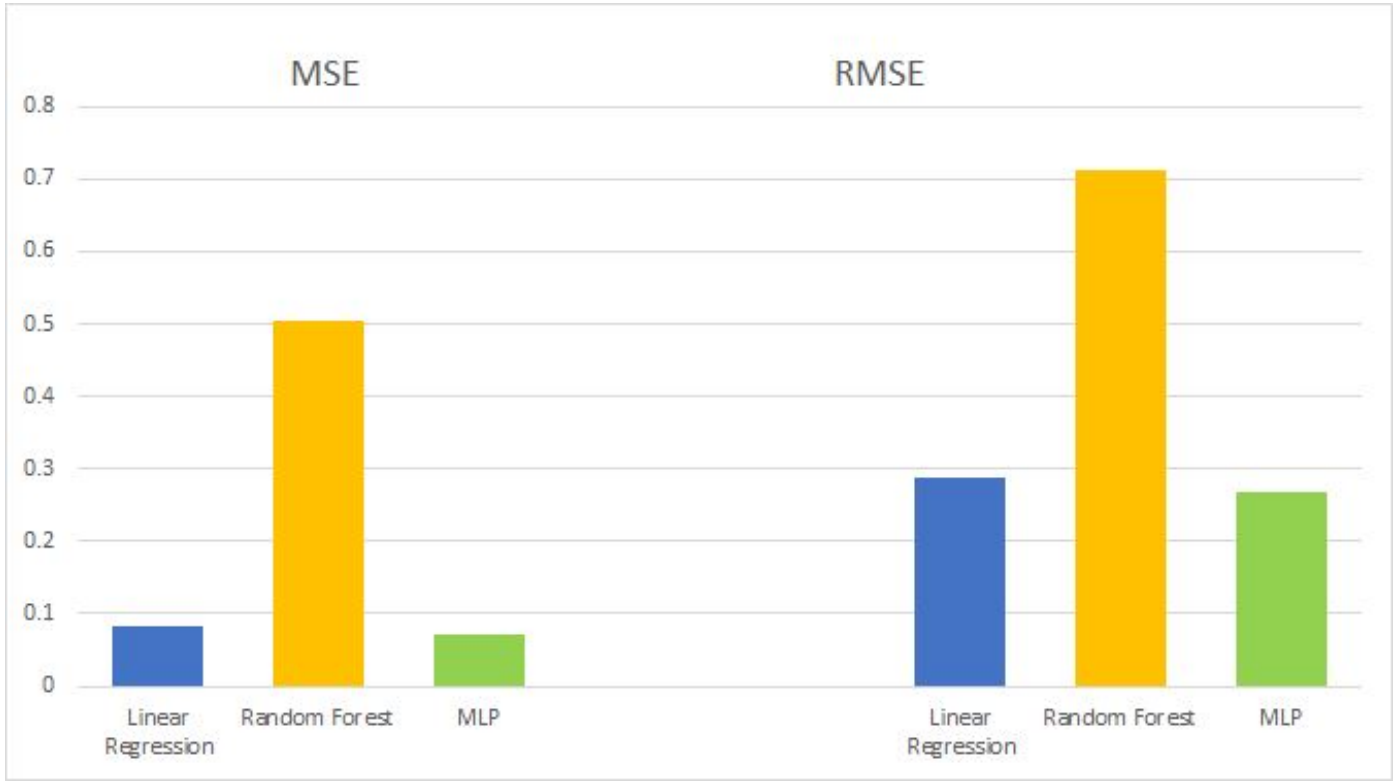
Results Evaluation

- Root Mean Square Error:

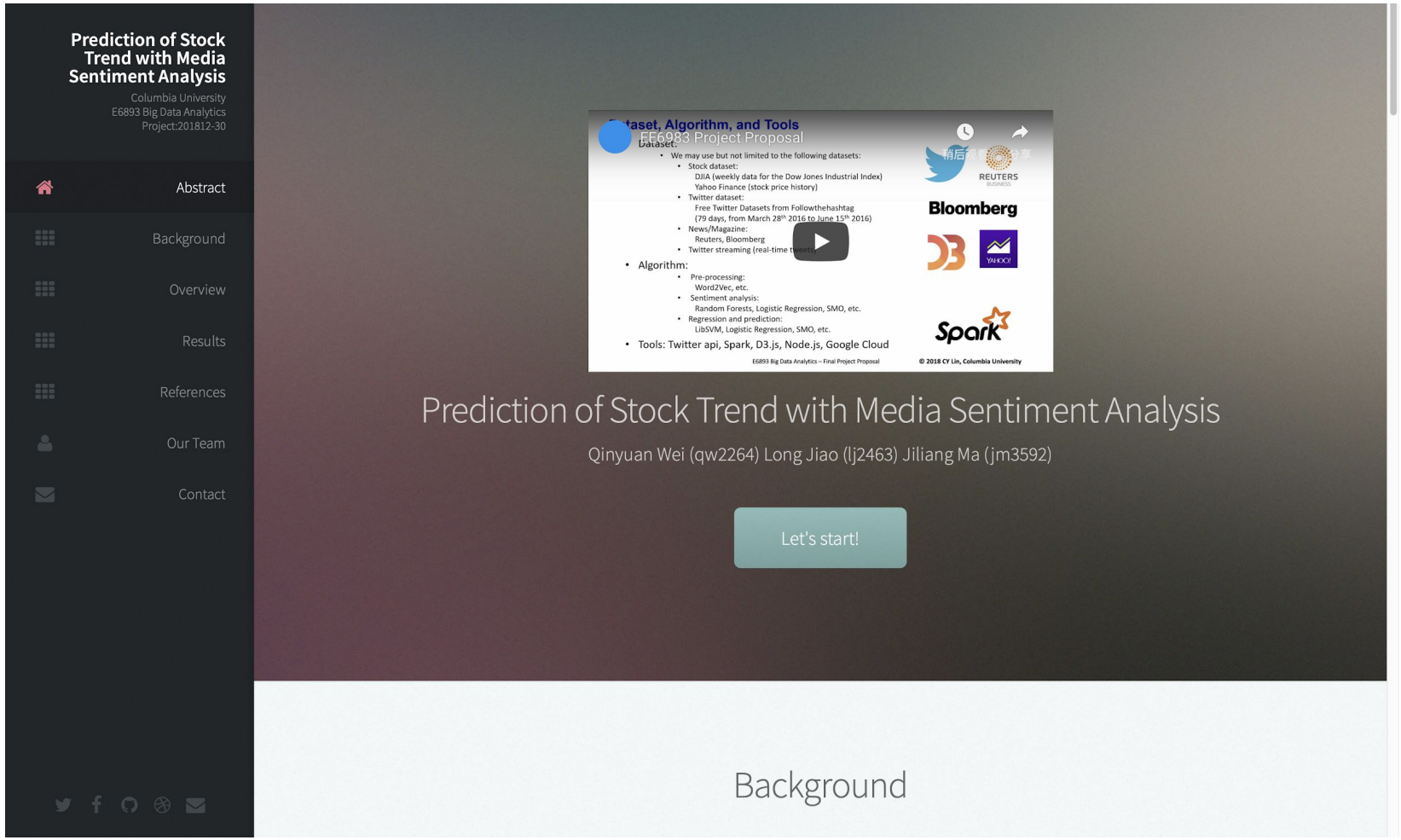
$$\mathrm{RMSD}(\mathbf{v}, \mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|v_i - w_i\|^2}$$

- Mean Square Error:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

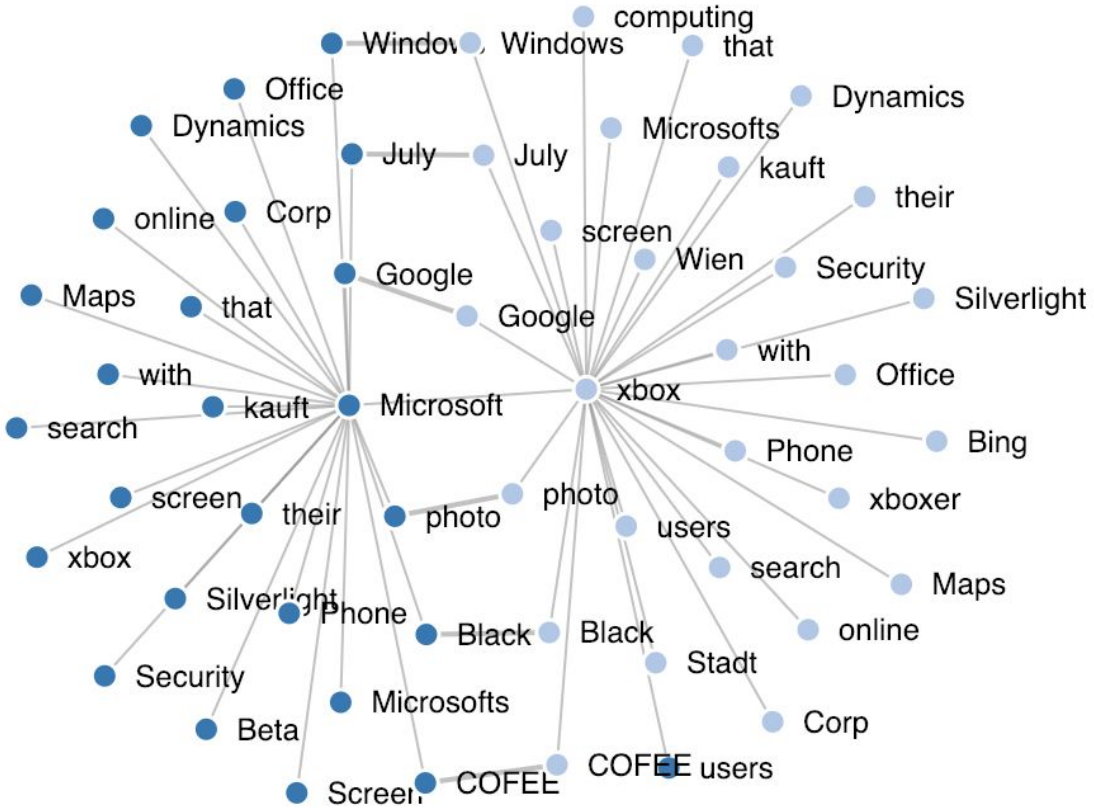


Website



Website

Website



Demo

Future work

- Test on more companies/market prices
- Use twitter api to gather more data and test (recent years data)
- More indexes to prove the model is efficient

References

1. 476 Million Twitter Tweets, J. Yang, J. Leskovec. Temporal Variation in Online Media . ACM International Conference on Web Search and Data Mining (WSDM '11), 2011. https://snap.stanford.edu/data/twitter7.html

2. Skuza, Micheal. Romanowski, Andrzej.(2015). Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction DOI: 10.15439/2015F230 http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7321604&tag=1

3. Pagolu, S.Venkata. Reddy, N.Kamal. Panda, Ganapati (2016). Sentiment analysis of Twitter data for predicting stock market movements DOI: 10.1109/SCOPES.2016.7955659 http://ieeexplore.ieee.org/document/7955659/?part=1

Thanks!