Project-1
Text Mining

Jiannliang Tsay
NYU-Cybersecurity
CS-GY6923
jlt245@nyu.edu

Introduction

- Twitter is an online social networking service that enables users to send and read tweets, short messages of 140-characters. Over 300 millions monthly active users create over 500 millions tweets per day.
- This project shows my interests in recent election 2016 and practices my skill in text mining of Machine Learning.

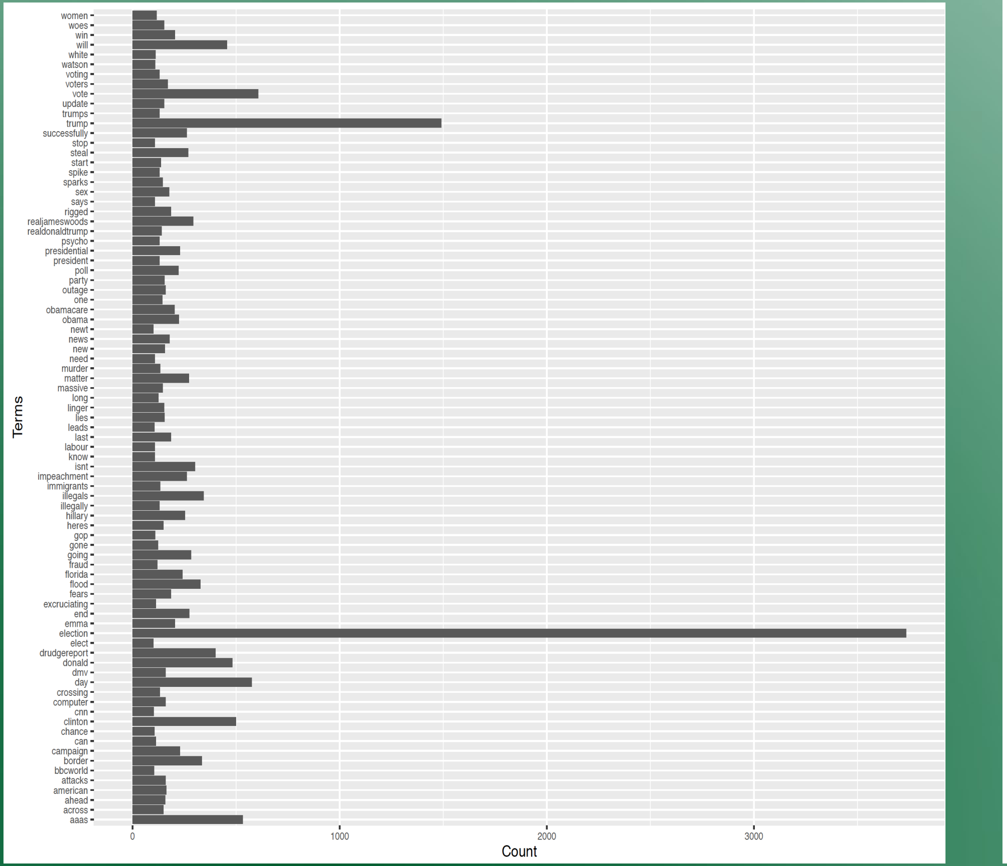
Techniques and Tools

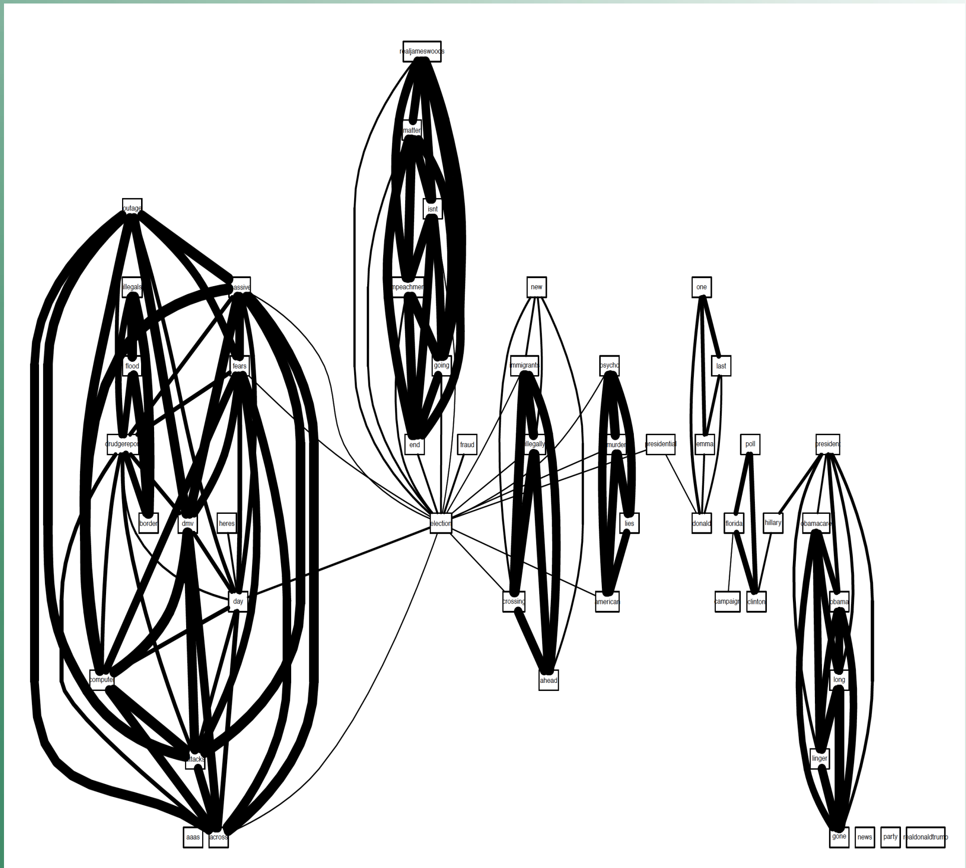
- Techniques
  - Text Mining
  - Topic Modeling (LDA)
  - Clustering (Spherical K-means)
  - Social network analytics
- R and its packages
  - twitteR
  - tm
  - topicmodels
  - skmeans

Outline

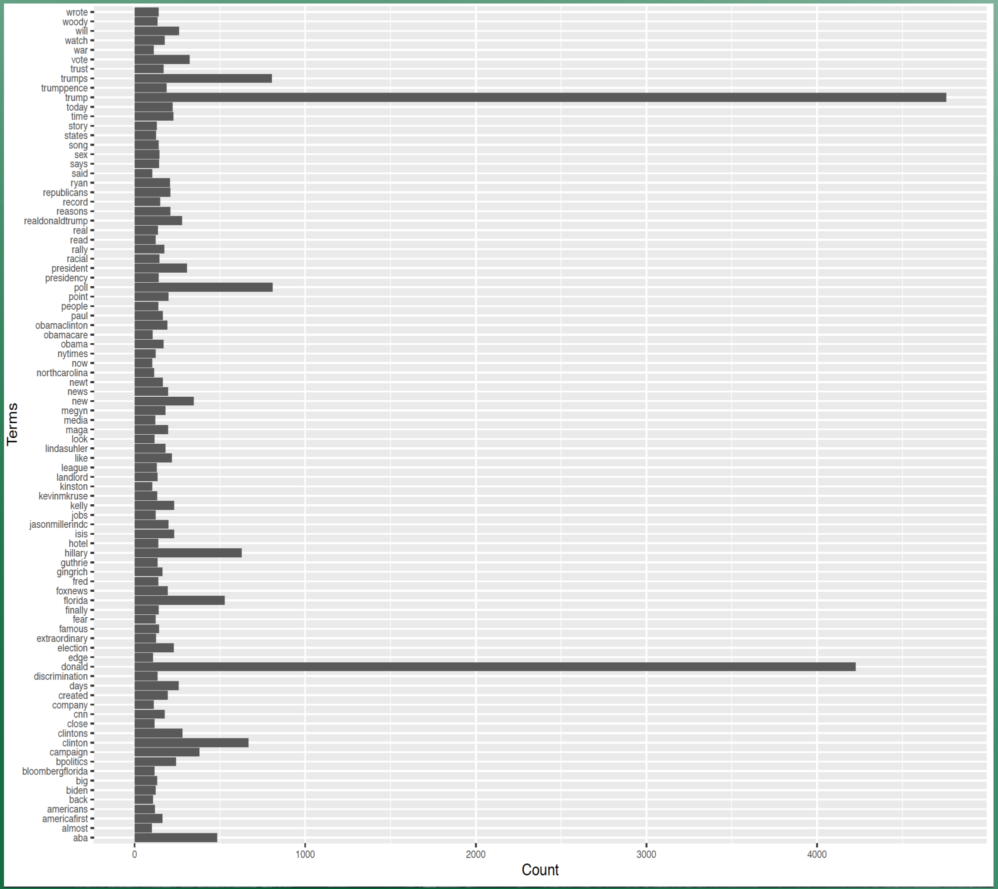
- Data Extraction from Twitter with R and twitteR package, search on “Election 2016”, “Donald Trump” and “Hillary Clinton” three datasets.
- Data Preprocessing with tm package, save the TermDocumentMatrix into files for each dataset.
- Analyze the term occurrence, association for each dataset with tm and wordcloud packages.
- Analyze the topics with LDA by using package topicmodels.
- Clustering the topics with skmeans and hclust packages.

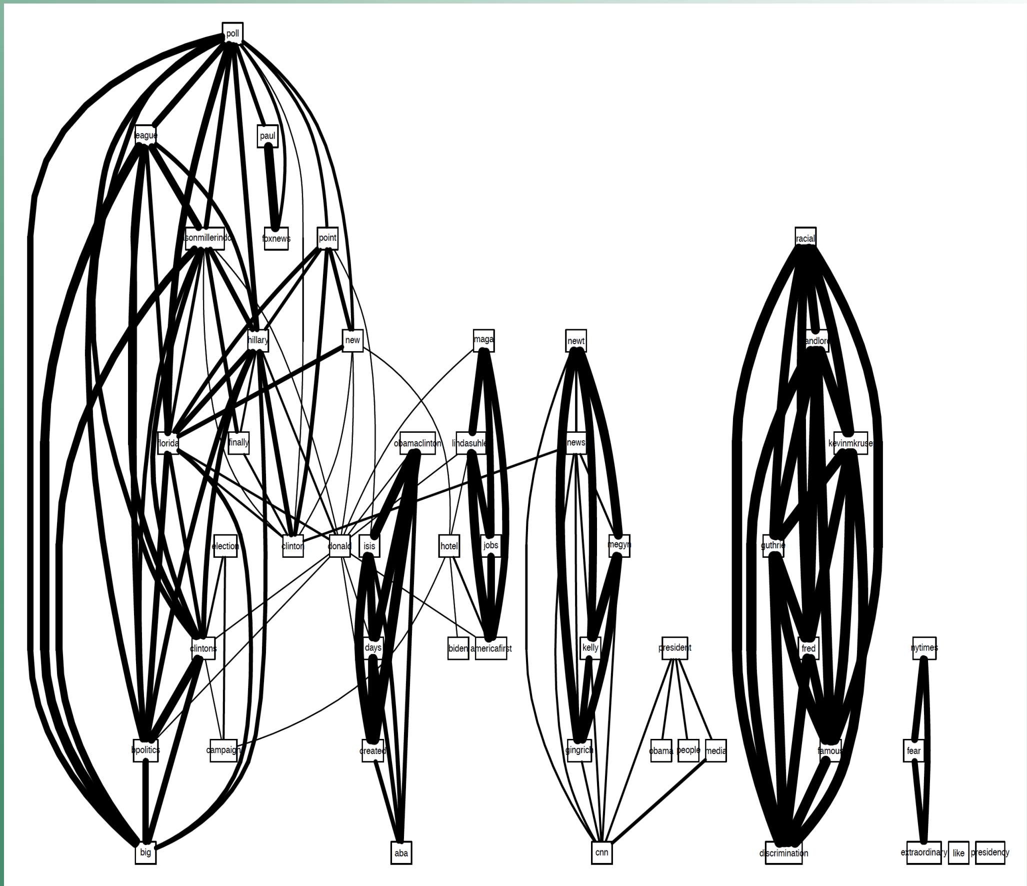
Term Frequency Analysis (Election 2016 Dataset)



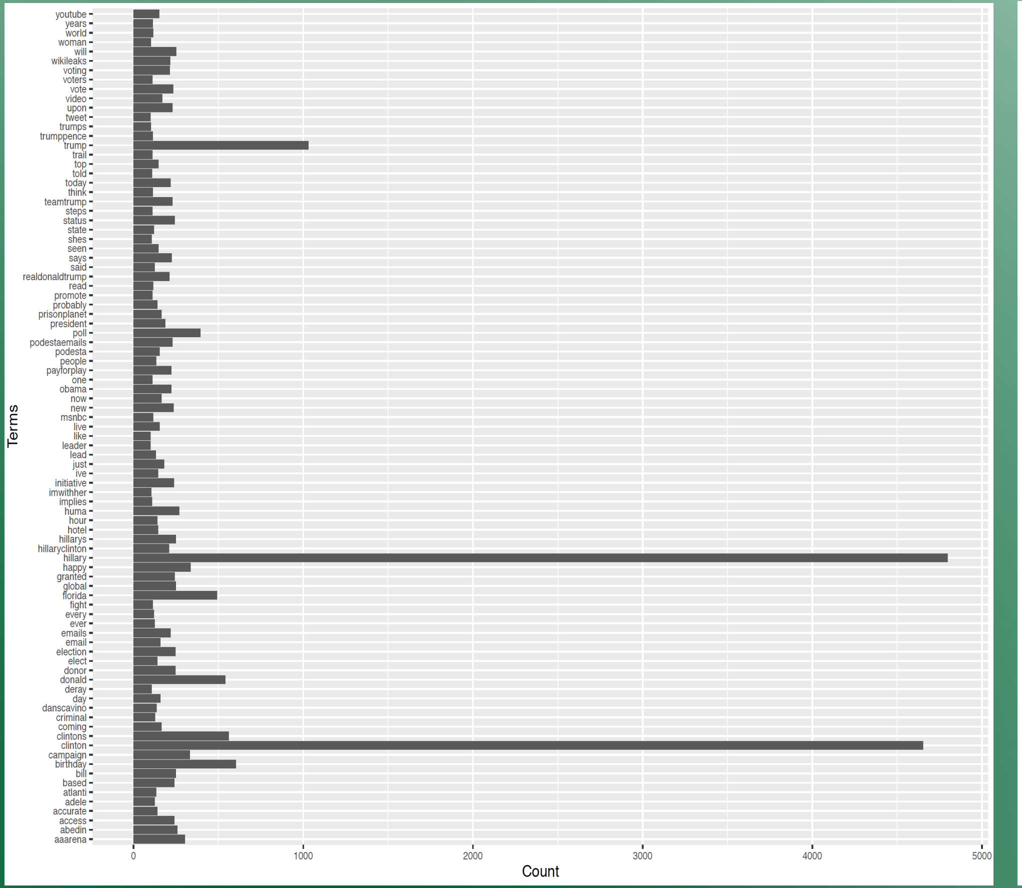


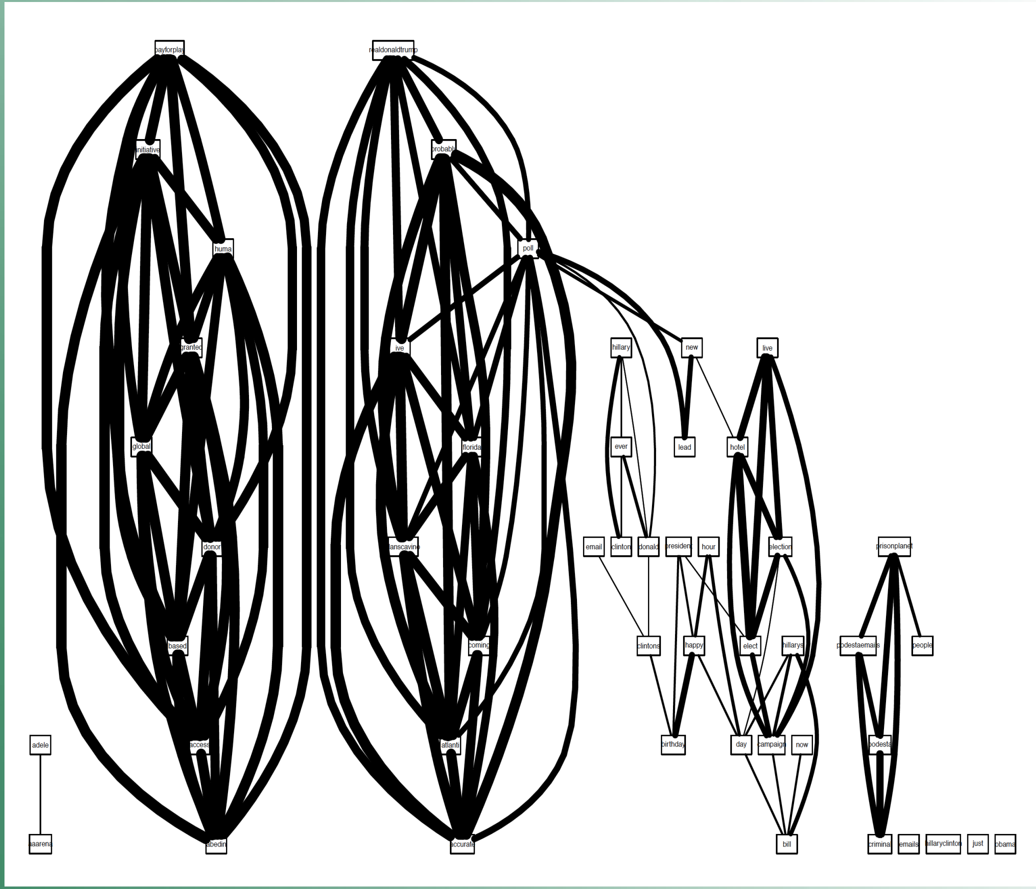
Term Frequency Analysis (Donald Trump Dataset)





Term Frequency Analysis (Hillary Clinton Dataset)

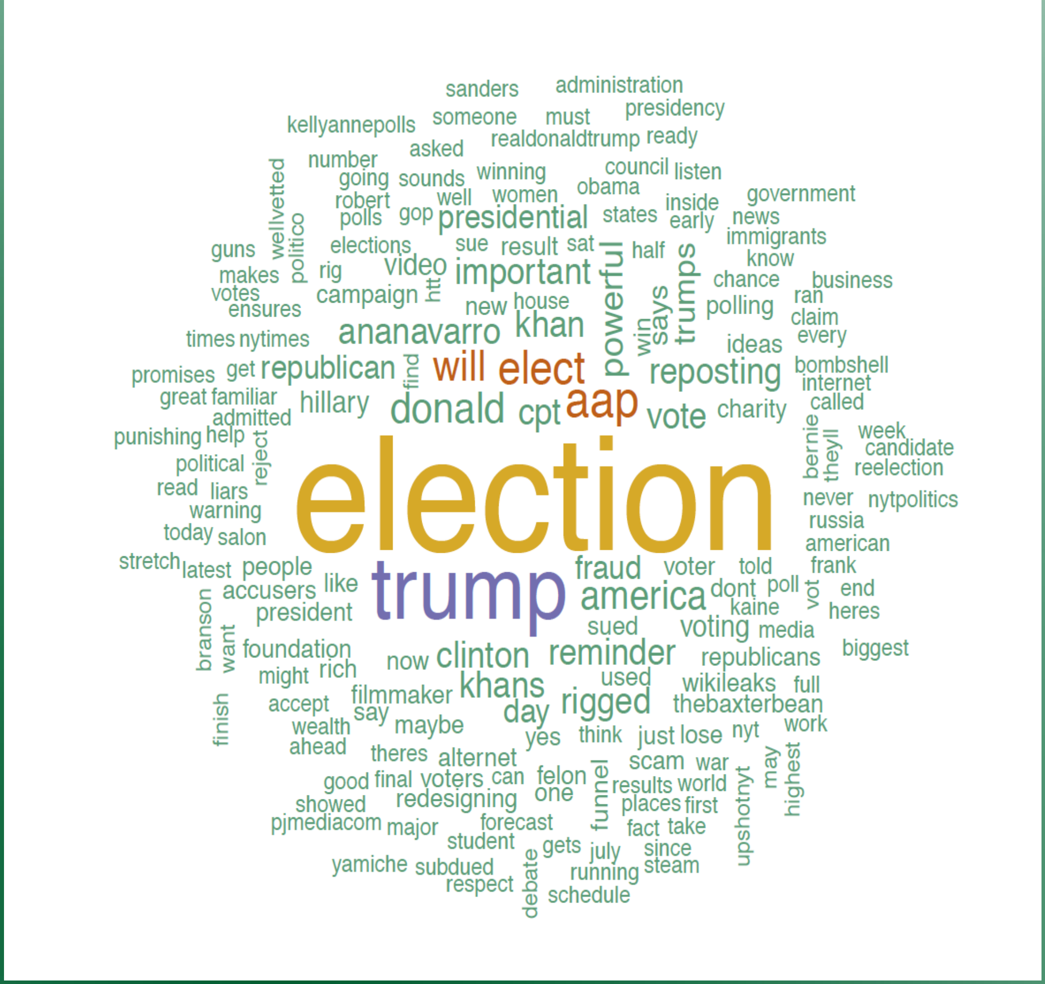


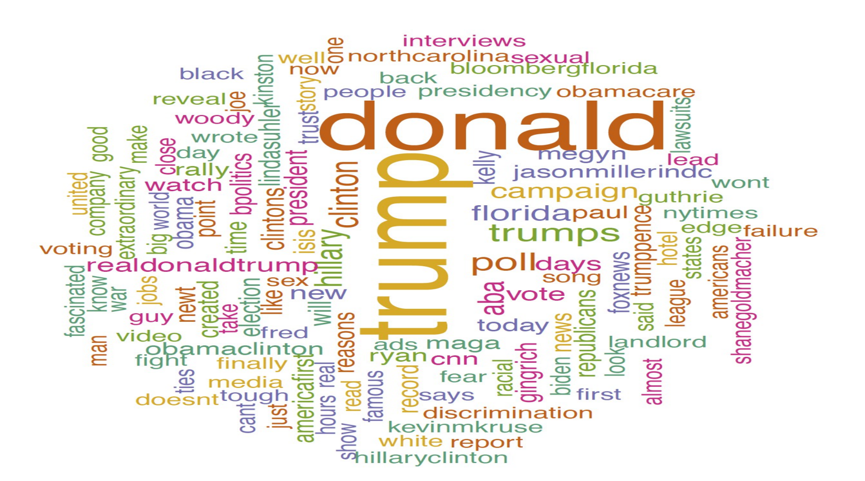


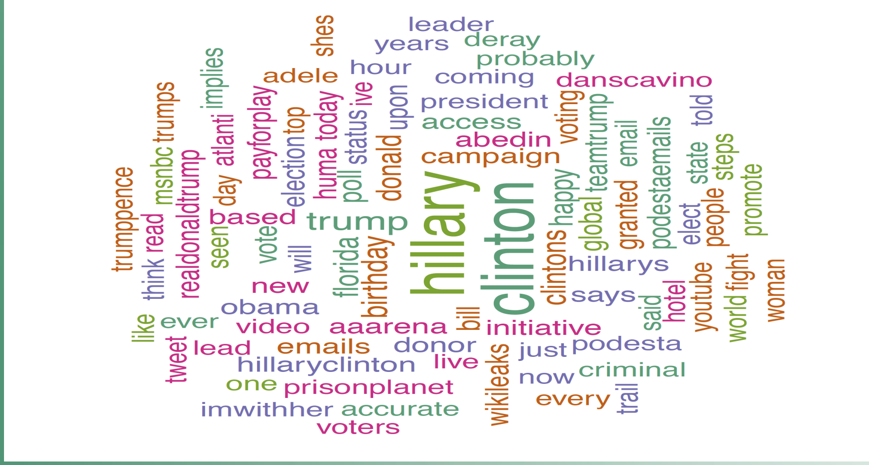
Term Frequency Analysis Summary

- For this Election, people talk more about “Trump” than “Hillary”, obviously “Trump” is controversial.
- If people talk about “Hillary”, seems they will discuss “Trump” as well. However it does not occur in “Trump” dataset.
- “CNN” seems a topic discussed for Trump. Trump criticizes the media is unfriendly to him.
- The topic of Florida Pool appears on both sides, seems it is a critical state to both of them.
- The topic of “Discrimination” is hot on Trump side.

Comparison of Wordclouds







Summary of wordcloud

- In this election, people talk about “Trump” more than “Hillary”.
- “Discrimination” and “Florida Poll” seem hot topics in Trump dataset.
- “Florida Poll” is discussed in Hillary dataset as well.
- Wordcloud charts pretty much reflect similar results as Term Frequency Analysis.

Topic Modeling

- This is my weakness of this project that I am unable to trace the tweets over a period of time.
-

Clustering