Topic Modeling

Lihua Pei (Neo)

Analysis Donald Trump Twitters (@readDonaldTrump)

Data Describe

1. 2009/5/20 - 2017/4/20
2. 32826 Twitters
3. The mean length of twitters is 109.69 words.

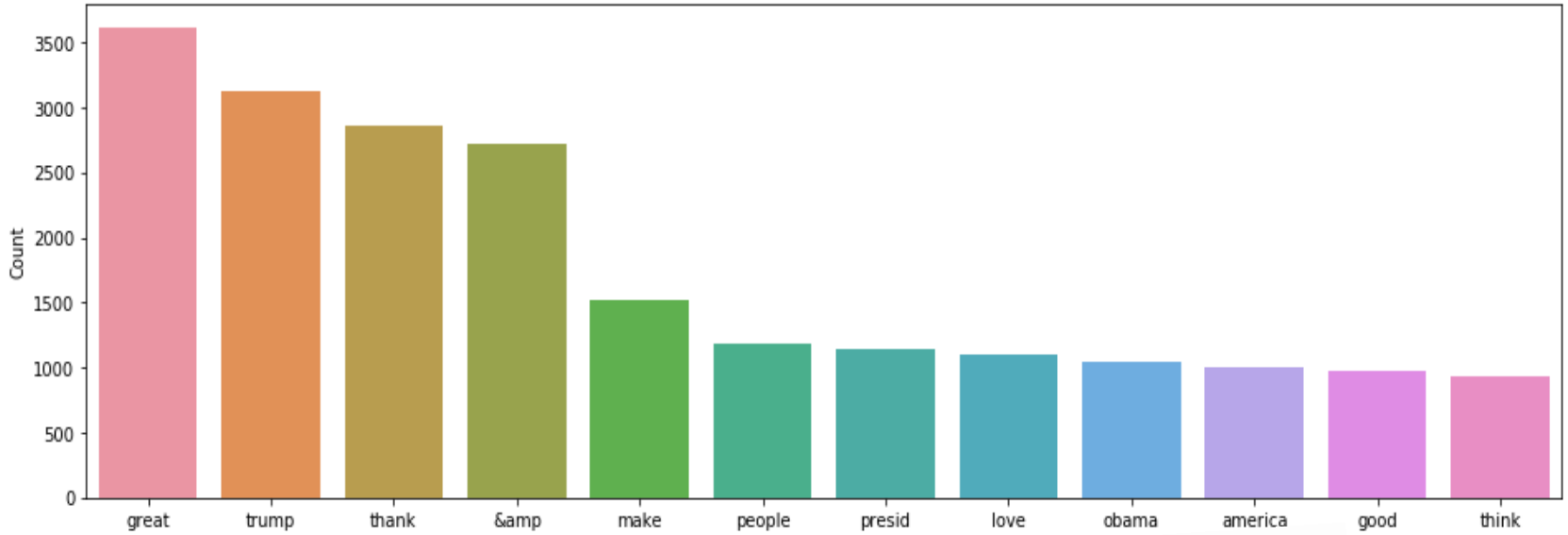
Preprocessing

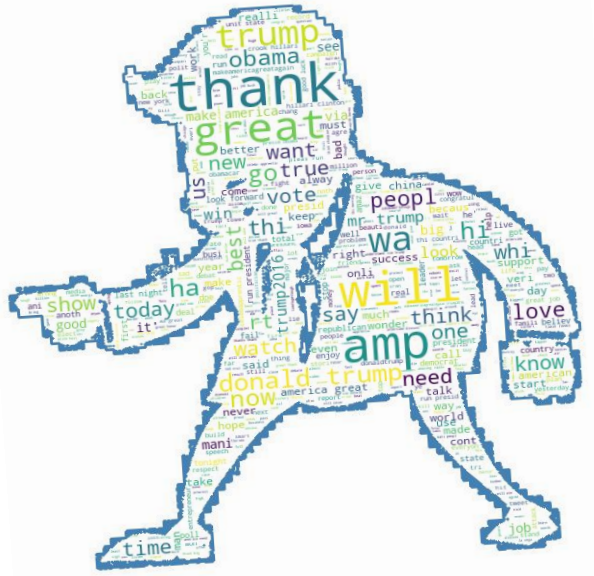
(Data Cleaning)

- 1. Remove the special elements of twitters: “https” (link), “RT”(retweet), “@” , “#” (hashtag), “&”(&amp) --- regular expression.
- 2. Normalization: .lower()
- 3. Tokenization: split()
- 4. Stop Word: NLTK—stop words
- 5. Stemming and Lemmatization: PorterSteemer().stem()

Basic Analysis

bag of words
BOW





Unsupervised Learning

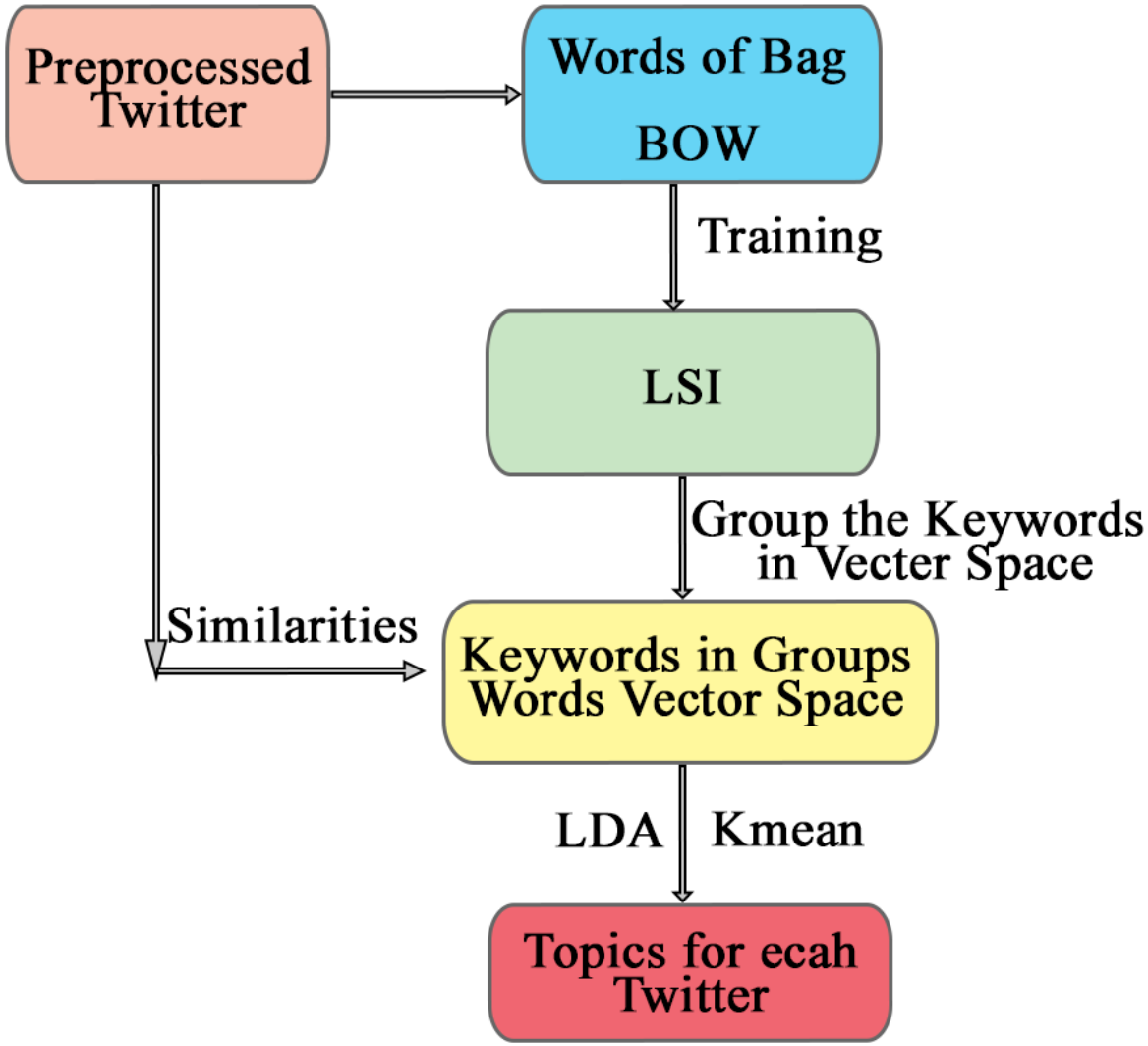
Using LSI and LDA

Analysis Donald Trump Twitters (@readDonaldTrump)

- 1. LSI (Latent Semantic Indexing):
- (**LSI**) is an **indexing** and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text.

- 2. LDA (Latent Dirichlet Allocation)
- (LDA) Latent Dirichlet allocation is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

Processing

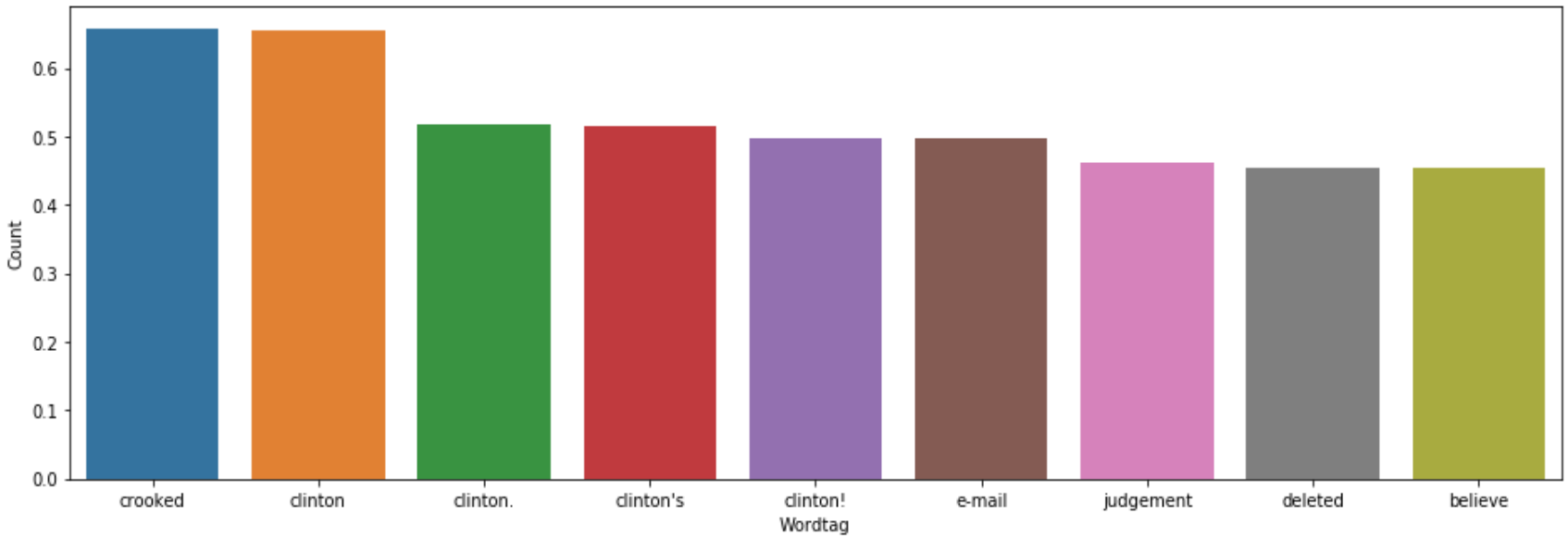

Preprocessed Twitter
Words of Bag
BOW
Training
LSI
Group the Keywords in Vecter Space
Similarities
Keywords in Groups Words Vector Space
LDA
Kmean
Topics for ecah Twitter

Grouped Keywords

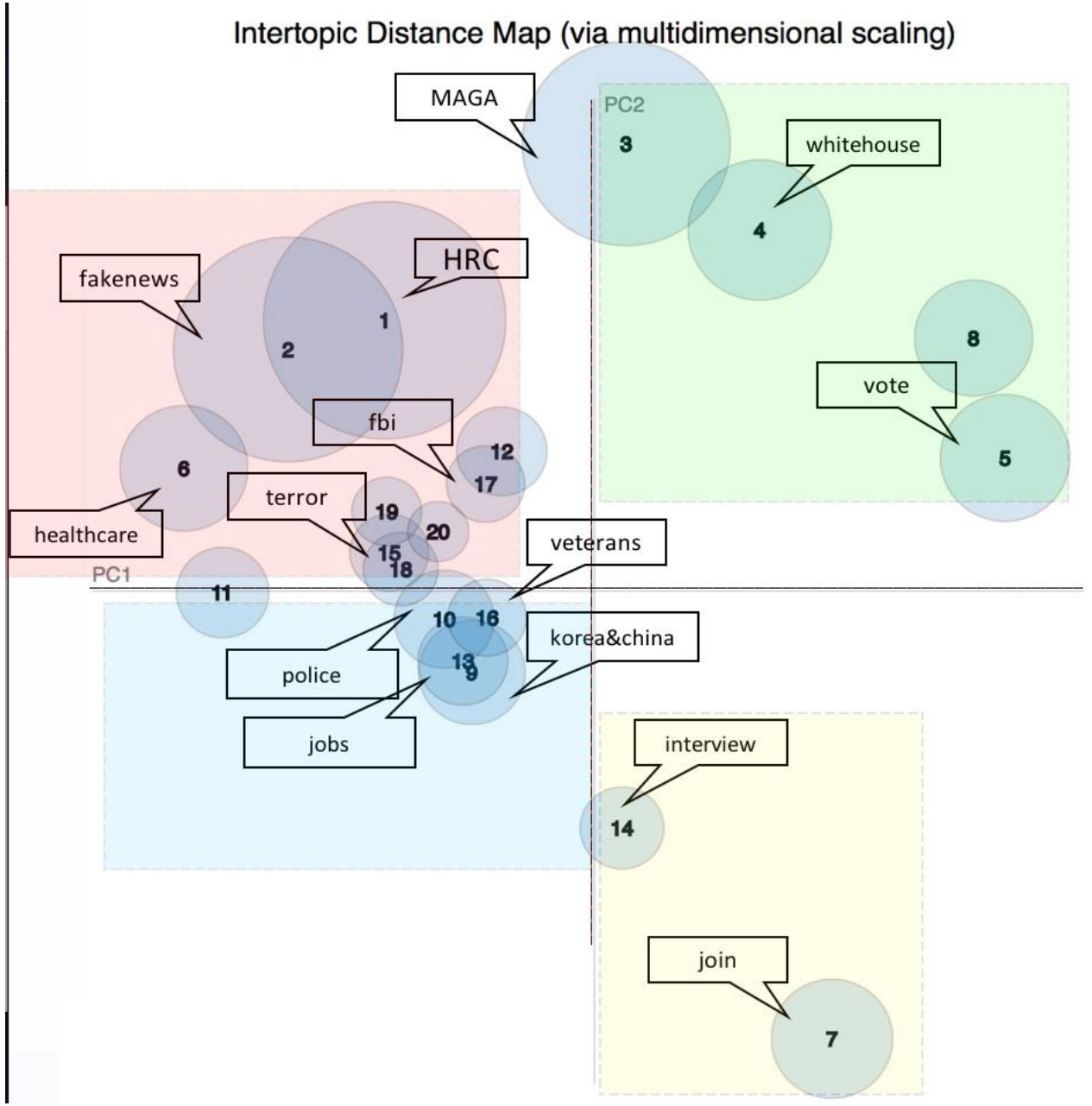
```
LSI:
Topic1
['thanks.', '"', 'great', 'thank', 'you.', 'true', 'trump', 'donald', 'you!', 'president']
Topic2
['thank', 'you.', 'you!', '"', 'great', '#trump2016', '#makeamericagreatagain', 'america', 'thanks.', 'make']
Topic3
['great', '"', 'america', 'make', 'you.', 'thank', 'thanks!', 'again!', 'trump', 'president']
Topic4
['thanks!', 'great', 'america', 'again!', 'make', 'you.', 'true', '#trump2016', '&amp;', '#makeamericagreatagain']
Topic5
['great', '"', 'again!', 'america', 'make', 'trump', 'true!', 'run', 'donald', 'thanks!']
Topic6
['#trump2016', '#makeamericagreatagain', 'you.', 'you!', 'again!', 'america', 'great', 'make', 'thank', 'new']
Topic7
['true!', '"', 'trump', 'again!', 'run', 'donald', 'president', 'great', 'america', 'make']
Topic8
['you!', '#makeamericagreatagain', 'you.', '#trump2016', 'thank', '#americafirst', 'america!', 'america', 'again!', '
make']
Topic9
['great!', 'thanks', 'run', 'president', 'good', 'please', 'luck.', '"', 'donald', 'trump']
Topic10
['thanks', 'great!', 'good', 'luck.', 'run', 'president', 'please', 'luck!', 'america', '"']
Topic11
['run', 'trump', 'please', 'thanks', 'donald', 'president', 'via', '&amp;', '"', 'new']
Topic12
['trump', 'donald', '&amp;', 'thanks', 'via', 'get', 'again!', 'people', 'america', 'like']
Topic13
['love', '"', '&amp;', 'president', 'run', 'obama', 'via', 'great', 'please', 'trump']
Topic14
['good', 'luck.', 'thanks', 'luck!', '&amp;', 'enjoy!', 'president', 'interview', 'new', 'interviewed']
```

Word Embedding

- The most similar words to [‘hillary’]
- 1. crooked - 0.65804
- 2. Clinton – 0.6548
- 3. e-mail – 0.4972



Inter-topic Distance



Thank You !

Lihua Pei (Neo)

Analysis Donald Trump Twitters (@readDonaldTrump)