Real-time news aggregation and exploratory natural language processing

Micah Rosales

B.S. Computer Science, Yale University

Introduction

- News sites are independent, fragmented, and hard to analyze collectively.

News Aggregator

- Aggregate news articles from Tweets ~ 40,000 articles in 3 wks
- Follow < 50 Twitter Accounts
- Over 4000 sources of news (websites)

- Extracts URLs and performs basic NLP

Processing

- Preprocessing
  - Stemmming
  - Tokenizing
  - Stopwords
  - Dictionary limits
- Keywords
  - Rapid Automatic Keyword Extraction (RAKE),
  - Stanford Named Entity Recognizer (NER),
  - tf-idf keyword weighting
  - Latent Dirichlet Allocation (LDA)

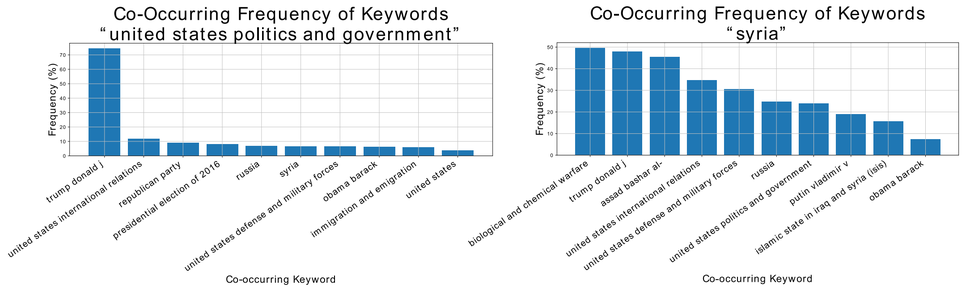

Figure 1. Co-occurrence of top terms frequency plots

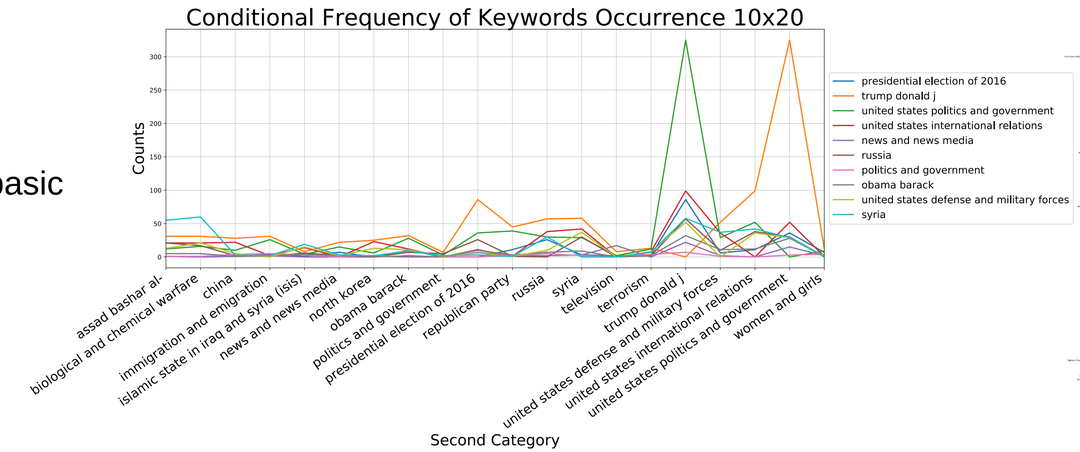

Figure 2. Conditional frequency plot of most common keywords

Cluster 2 words: North, Korea’s, nuclear, missiles, South, Carolina

	Keyword 1	Keyword 2	Keyword 3
0	US Defense and Military Forces	US Politics and Government	Missiles and Missile Defense Systems
1	US Politics and Government	US Navy	North Korea
2	Missiles and Missile Defense Systems	Kim Jong-un	North Korea
3	NCAA Basketball Championship (Men)	Basketball (college)	Gonzaga University

Figure 3. K-Means clustering result. Interesting result that it groups the NCAA basketball championship with the North Korea Missile stories. Articles mention S. Carolina governor Nikki Haley linking N. Korea and S. Carolina (where basketball championships were held).

Figure 5. Ward hierarchical clustering

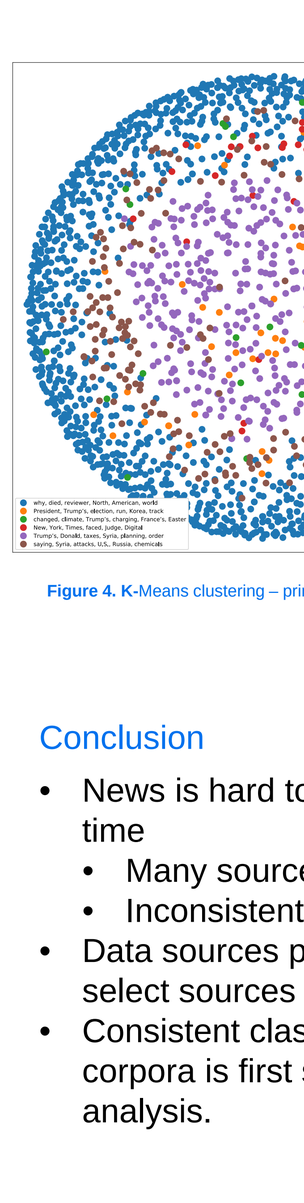

Figure 4. K-Means clustering – primary keywords

Conclusion

- News is hard to analyze in real time
  - Many sources
  - Inconsistent formats
- Data sources privilege few select sources
- Consistent classification across corpora is first step in news analysis.