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
  • Stemming
  • 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

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.