Text Analysis / Mining

1. Get text data
   1. Direct download existing datasets
   2. API
   3. Scrape webpage
2. Tokenization
   1. Doc – sentences – words
   2. Remove
      1. Punctuation
      2. Spaces
      3. Numbers
      4. Tags, css, js
   3. Lower case
   4. Remove stopwords (SJ is watching the blackboard. ‘is’ and ‘the’ are stopwords).
      1. (nltk, corpus)
   5. Stemming
   6. Bags of words
      1. Phrases
      2. Common combinations
   7. **TF**(Term Frequency) / **IDT**(Inverse Document Frequency)
3. Analysis
   1. objectives