Text Analytics & NLP

Giri Narasimhan

2019

Text Analytics

- Preprocessing
- Counting Terms
- Latent Semantic Analysis
  - Dimensionality reduction techniques via SVD
- Topic Modeling

Why is text analytics important?

- 80% of world’s data is in textual form
- 2.2 M books published every year
  - 66B words per year
- 10K daily newspapers
  - 10B words per year
- 100B texts on social media every year
  - 1T words per year
- 4000 hrs on YouTube
- Alexa, Siri commands

Preprocessing

- Text files can be in:
  - Text, RTF, HTML, XHTML, WP, pdf, docx
- Files may be encoded and in different languages
  - Need UTF8 or UTF16
- Documents contains strings
- Corpus is a collection of documents
- **Tokenization**
  - Convert strings to units of analysis, e.g., words

More preprocessing

- Filtering stop words (SMART)
- Zipf distribution of counts:
  - Kth most frequent term has frequency 1/k
- Remaining are called content words
- Normalization
  - Words may be capitalized
  - Same word different meanings (position tagging & lemmatization)
  - Singular vs plural words (stemming)

Counting

- Term-document matrix with entries:
  - **Bag-of-words** assumption – set matters, not order
  - Term presence/absence or frequencies in document
- Document Matrix with entries:
  - # of docs in which a term appears

Applications of Counting & TDMs

- **Document retrieval**: Search for a document given a set of search terms
- **Co-occurrence**:
  - Given that a term appears, find prob of occurrence of another term in same document
  - bigram matrix – freq of consecutive pairs of words
  - n-gram matrix – freq of sequence of n consecutive words
  - Many more … smoothing, skipgrams
- **Document similarity**: comparing vectors

Shakespeare

- 28K words from 43 volumes
- 12K singletons
- 500 or so stop words
- 10 most frequent words make up 21%
- In general, matrices are
  - Large, sparse and noisy

Latent Semantic Indexing (LSI)

- SVD decomposes M into product of 3 matrices
- Apply SVD to TDM
- The middle matrix gives you bases vectors
- By reducing the number of singular vectors, one can reduce dimensionality of the space
- Clusters in this space reflect doc similarity

Topic Modeling

- Basis vectors in SVD don’t have clear meaning
  - Find more meaningful decompositions
- Another concept of