Review Stat. 653

Prof. Eric A. Suess

April 28, 2021

Review

What have we learned about studying Text Mining?

Lot of new things about how to work with text data.

Review

- Bag-of-Words Model
- Tokenizers
- Corpus and Document Term Matrix
- Stemming
- Word Counts
- TF-IDF
- Sentiment Analysis
- n-grams
- Pairwise Correlation of words
- Topic Modeling
- Parts of Speech

Review

All of these ideas are as useful as computing means, standard deviation, correlations, t-tests, regression, etc. for numeric data.

You are now prepared to work with the other half of the data that is out there in the world!

Review

We have studied Unsupervised Learning techniques for text based data.

Sentiment Analysis is very useful for learning about the sentiment in documents.

Review

We have studied Unsupervised Learning techniques for *clustering* text based data.

Topic Analysis is very useful for learning about the different topics discussed in documents.

Review

We have studied Supervised Learning techniques for *classifying* text based data.

Naive Bayes and **Logistic Regression with lasso/regularization** are very useful for predicting which class documents are in.

Review

There are a lot of other R packages that can be used for Text Mining.

- Rvest
- Quanteda
- Text2vec
- Spacy
- Rtweet

Review

There are a lot of Python packages that can be used for Text Mining.

- NLTK
- Textblob
- SciKit Learn
- Beautiful soup
- Gensim
- Spacy
- CoreNLP
- Pattern
- Polyglot
- Twint

Review

There are growing opportunties to work doing Text Mining. This is a very interesting new field to work in and there are a growing number of excellent tools available to pursue such work.