Intro to Text Analysis



Bay Area Women in ML/DS

Nov 2016

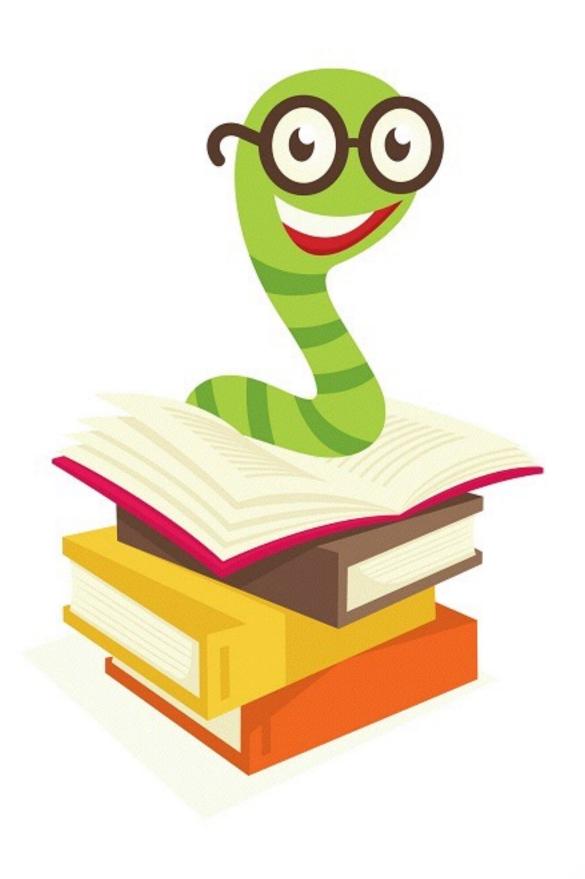
Overview

jobs support campaign bad sanders republican wow pennsylvania virginia @foxnews president

Cruz america country carolina media #imwithyou obama join og special money #votetrump kasich speech isis wall totalwatch interviewed win york lyin newsnice totalwatch enjoy bernie totally people bernie ber

- Intro to Text Analysis
- Feature Engineering for Text Analysis
 - How to convert raw text data to feature rich data ready for ML
 - Bag of Words & TF-IDF
- Topic Modeling & LDA
- Text Analysis Software in R & Python

Terminology



- Stemming
- Tokenization
- Document Term Matrix (DTM)
- Bag of Words
- TF-IDF
- Word2Vec
- Topic Modeling

Text Analysis in R



- tidytext by Julia Silge & David Robinson
 - http://tidytextmining.com
- quanteda by Kenneth Benoit
- tokenizers by Lincoln Mullen
- tm (Text Mining)
- topicmodels (Topic Modeling)
- · Ida (Topic Modeling)

Text Analysis in Python



- nltk (Natural Language Toolkit)
- scikit-learn
 - http://scikit-learn.org/stable/tutorial/ text_analytics/working_with_text_data.html
- gensim (Topic Modeling)
 - https://www.analyticsvidhya.com/blog/ 2016/08/beginners-guide-to-topic-modelingin-python/