Course Expectations

Module 1 – Introduction
  • NLP Versus Text Analytics
  • The Libraries Used in the Course
  • The Data Sets
  • Working Definitions

Module 2 – Tokenization
  • Introducing Tokenization
  • Word and Sentence Tokenization
  • Specialized Tokenization Example - Tweets
  • N-Grams
  • Part-of-Speech (POS) Tagging
  • Hands-On Lab #1

Module 3 – Token Normalization
  • Case Folding
  • Stopword Removal
  • Stemming
  • Lemmatization
  • Stemming Versus Lemmatization
  • Hands-On Lab #2

Module 4 – Vector Space Model
  • Document Vectors
  • Bag of Words (BoW)
  • Optimizing BoW
  • Adding N-Grams
  • Controlling Dimensionality
  • The Vector Space Model
  • Hands-On Lab #3
Module 5 – TF-IDF
- Vector Space Model Limitations
- Introducing Term Frequency-Inverse Document Frequency (TF-IDF)
- The TF-IDF Calculation
- Adding N-Grams
- Controlling Dimensionality
- TF-IDF and the Vector Space Model

Module 6 – Grouping Documents
- Techniques for Grouping Documents
- Cosine Similarity
- Clustering Documents
- Clustering Documents with K-Means
- The K-Means Algorithm
- Euclidian Distance
- K-Means Caveats
- Hands-On Lab #4

Module 7 – Classifying Documents
- Introducing Document Classification
- The Naïve Bayes Algorithm
- How Naïve Bayes Learns
- Predicting Document Classes with Naïve Bayes

Module 8 – Additional Resources

Hands-On Lab #5