LDA – MALLET Demo

School of Information Studies
Syracuse University

LDA

- Latent Dirichlet Allocation: “a generative probabilistic model for collections of discrete data such as text corpora” (Blei, Ng, Jordan, 2003)
  - Generative: unsupervised
  - Probabilistic: uses probabilities
  - Discrete: data can be categorized into classes

- Most popular library: Gensim

LDA: MALLET

- MAchine Learning for LanguagE Toolkit
- "MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text" (website)
- Implementation of LDA

Demo

- Jupyter Notebook with code for running Gensim's LDA and MALLET
  - Including pre-processing steps
- Visualizing results with pyLDAvis
- Creating a csv file with list of documents and cluster labels assigned to them
- Basic code to run Gensim and SpaCy