Lab post-hoc "syllabus"
Sociology 220: Computational Text Analysis – Spring 2018
Zach Wehrwein

1. A (re) introduction to Python
   (a) Jupyter notebooks
   (b) Package management
   (c) Python data structures
   (d) Loops, list comprehensions, user-defined functions, lambda functions
   (e) Control flow
2. Web-scrapping with Beautiful Soup
   (a) Structure of HTML
   (b) Basic pandas
   (c) Web-scrape of Old Bailey Online
   (d) Pickles and json files
3. APIs, data storage options
   (a) Solutions to scrape of UCSB SOTU speeches website
   (b) Twitter API
   (c) streaming vs REST examples
   (d) Writing to CSV files
4. Pre-processing.
   (a) The Grimmer and Stewart "standard" bag-of-words options
   (b) The $2^7 = 128$ choices to be made
5. Data wrangling, descriptives, dictionary methods
   (a) Document-term matrix and more Pandas
   (b) NLTK for contextual language, word frequencies
   (c) spaCy for named entity recognition and parts of speech tagging
   (d) Simple count dictionary methods / "sentiment" analysis
6. Baby introduction to machine learning
   (a) Logic of ML and cross-validation
   (b) Introduction to scikit-learn package.
   (c) Logistic "regression"
   (d) Naive Bayes
   (e) SVMs
   (f) Visualizing t-SNE, ROC-AUC, and confusion matrices with yellowbrick package.
7. Clustering algorithms and topic models 1
   (a) k-means clustering
   (b) Hierarchical clustering
   (c) Vanilla LDA
   (d) Visualizing LDA with pyLDAvis
8. Topic models 2
   (a) Perplexity and log-likelihood of training data

(b) Tuning the hyper-parameters of topic models
(c) Plotting vanilla topic longitudinally
(d) LDA vs NMF (in scikit-learn)

9. Topic models 3
   (a) "Toy" topic model done by hand
   (b) Introduction to gensim library
   (c) Vanilla LDA in gensim
   (d) Dynamic topic models in gensim
   (e) Visualizing DTMs
10. Word embeddings 1
   (a) XML files
   (b) word2vec
   (c) word2vec visualization
11. Word embeddings 2?
   (a) doc2vec
   (b) lda2vec
   (c) co-occurence networks?
12. A strong cocktail?