Jeffrey Tharsen tharsen@uchicago.edu
"Advanced Methods for Textual Analysis: Stylometry, Topic Modeling and Text Visualizations"
RCC Workshop — Spring 2019

Sources: *http://home.uchicago.edu/~jcarlsen/AdvText.zip (453 MB)*
https://github.com/rcc-uchicago/TextViz

I. General Platforms for Visualizing Data (for inspiration)

Tableau : tableau.com/academic/teaching *(Sign up for an educational account)*

D3 (Javascript) Gallery : github.com/d3/d3/wiki/Gallery
bokeh (Python) / rbokeh (R) : bokeh.pydata.org / hafen.github.io/rbokeh

II. Basic Tools for Textual Analysis (+ Part-Of-Speech & Named Entity Recognition)

Where can I get digital texts? = Online repositories; OCR (paper → digital plaintext)

a. ***Voyant Tools* : voyant-tools.org (word frequencies, word clouds, KWIC)**

Python commands (NLTK: Text object; collocations, KWIC, word frequencies) :
Basic Text analyses.ipynb

b. **POS & NER : stanford-postagger-3.7.0.jar , stanford-ner-3.7.0.jar**

List of POS tags:
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Python (SpaCy) POS & NER : **POS-tagging and Lemmatization in SpaCy.ipynb**
NER in SpaCy.ipynb

SpaCy installation instructions: https://spacy.io/usage

SpaCy NER tags : *https://spacy.io/usage/linguistic-features*

TAPoR Tools : tapor.ca

HathiTrust Research Center : analytics.hathitrust.org
HTRC Bookworm (Ngram search) : https://bookworm.htrc.illinois.edu/develop/

Visual Text Explorer : edoc.uchicago.edu/vte "simultaneous close and distant reading"

III. Tools for Stylometry (HCA Dendogram & *k-means* PCA)

a. ***LEXOS* (Comparative Stylometry : Dendrogram + PCA) : lexos.wheatoncollege.edu**

b. *Python-based Stylometry* : **Stylometry_HCA.ipynb** , **Stylometry_PCA.ipynb**

IV. Tools for Topic Modeling + Word2vec

a. *MALLET Topic Modeling* : mallet.cs.umass.edu
TopicModelingTool.jar : standalone Java-based application for Topic Modeling

b. *Python-based Topic Modeling* (via the gensim library, NLTK + SpaCy) :
Topic Modeling (gensim LDA + NLTK + SpaCy)_Shakespeare.ipynb
Topic Modeling evaluations_Shakespeare.ipynb

c. *Python-based Word2vec & TF-IDF* (gensim) : **Word2Vec all_Shakespeare.ipynb**
Word2Vec TF-IDF Shakespeare.ipynb

V. Tools for Text Reuse

a. **Philologic4 : http://anomander.uchicago.edu/text-pair/**
https://textual-optics-lab.uchicago.edu/

PCA + BLAST (for genomic/literary sequence analysis) : pvierth.herokuapp.com