*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*](http://home.uchicago.edu/~jcarlsen/AdvText.zip) *(453 MB)*

[*https://github.com/rcc-uchicago/TextViz*](https://github.com/rcc-uchicago/TextViz)

1. **General Platforms for Visualizing Data (for inspiration)**

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

*D3* (Javascript) Gallery : [github.com/d3/d3/wiki/Gallery](https://github.com/d3/d3/wiki/Gallery)

*bokeh (Python)* */ rbokeh* *(R)* : [bokeh.pydata.org](https://bokeh.pydata.org/en/latest/) / [hafen.github.io/rbokeh](https://hafen.github.io/rbokeh/)

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

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

* 1. ***Voyant* *Tools* :** [**voyant-tools.org**](http://voyant-tools.org/) **(word frequencies, word clouds, KWIC)**

*Python commands* (NLTK: Text object; collocations, KWIC, word frequencies) :

**Basic Text analyses.ipynb**

* 1. **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*](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*](https://spacy.io/usage)

SpaCy NER tags : [*https://spacy.io/usage/linguistic-features*](https://spacy.io/usage/linguistic-features)

*TAPoR* Tools : [tapor.ca](http://tapor.ca/home)

HathiTrust Research Center : [analytics.hathitrust.org](https://analytics.hathitrust.org/)

HTRC Bookworm (Ngram search) : <https://bookworm.htrc.illinois.edu/develop/>

*Visual Text Explorer* : [edoc.uchicago.edu/vte](http://edoc.uchicago.edu/vte/vte.php) “simultaneous close and distant reading”

1. **Tools for Stylometry (HCA Dendogram & *k-means* PCA)**
   1. ***LEXOS* (Comparative Stylometry : Dendrogram + PCA) :** [**lexos.wheatoncollege.edu**](http://lexos.wheatoncollege.edu/)
   2. *Python-based Stylometry* : **Stylometry\_HCA.ipynb** , **Stylometry\_PCA.ipynb**
2. **Tools for Topic Modeling + Word2vec**
   1. *MALLET Topic Modeling* : [mallet.cs.umass.edu](http://mallet.cs.umass.edu/)

**TopicModelingTool.jar** : standalone Java-based application for Topic Modeling

* 1. *Python-based Topic Modeling* (via the gensim library, NLTK + SpaCy) :

**Topic Modeling (gensim LDA + NLTK + SpaCy)\_Shakespeare.ipynb**

**Topic Modeling evaluations\_Shakespeare.ipynb**

* 1. *Python-based Word2vec* *& TF-IDF* (gensim) : **Word2Vec all\_Shakespeare.ipynb**

**Word2Vec TF-IDF Shakespeare.ipynb**

1. **Tools for Text Reuse**
   1. **Philologic4 :** [**http://anomander.uchicago.edu/text-pair/**](http://anomander.uchicago.edu/text-pair/)

[**https://textual-optics-lab.uchicago.edu/**](https://textual-optics-lab.uchicago.edu/)

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