*Jeffrey Tharsen tharsen@uchicago.edu*

*“Text Analysis for Non-Western Scripts”*

*RCC Workshop — Summer 2019*

***Sources:***[*http://home.uchicago.edu/~jcarlsen/TA4NWS.zip*](http://home.uchicago.edu/~jcarlsen/TA4NWS.zip) *(781 MB)*

[*https://github.com/rcc-uchicago/text-analysis-for-non-western-scripts*](https://github.com/rcc-uchicago/text-analysis-for-non-western-scripts)

1. **Fonts with wide Unicode coverage (in Sources : Fonts)**

Cyberbit.ttf

Arialuni.ttf

Noto: <https://www.google.com/get/noto/>

1. **OCR and Basic Tools for Textual Analysis**

*Where can I get digital texts?*

**OCR** (paper 🡪 digital plaintext): Tesseract 4, ABBYY FineReader 14, Adobe ($$)

**Online repositories:**

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

HTRC Bookworm Search : <https://bookworm.htrc.illinois.edu/develop/>

Wikisource : <https://wikisource.org/>

Gutenberg : <http://www.gutenberg.org/>

***Basic Text Analysis Frameworks:***

*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

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)

*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**