Sources: http://home.uchicago.edu/~jcarlsen/TA4NWS.zip (781 MB)
https://github.com/rcc-uchicago/text-analysis-for-non-western-scripts

I. Fonts with wide Unicode coverage (in Sources: Fonts)
   Cyberbit.ttf
   Arialuni.ttf
   Noto: https://www.google.com/get/noto/

II. 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
   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 (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

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

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