Digital Sinology & Digital East Asia Workshop
tharsen@uchicago.edu
April 25, 2015

Handout for "Tokenization and Transformation: Turning Texts into Databases"

I. Whence the Text? (aka "Know thy Source Data")

II. Regular Expressions: Search and Replace

a. Always work with a **copy** of your source file
b. Working in Word
c. "Wildcards" and Text Editors: Full regular expression capabilities
d. Preparing the text/data for export
   i. Markup and layout strategies (see also section IV below)
   ii. Flat text – save as **UTF-8 (Unicode)**
   iii. File formats: Tab-delimited, CSV or …

III. Word → Flat Text File → Excel

a. Export to plain text (≠OLE)
b. Excel Strategies:
   i. Importing
   ii. Arranging the data
   iii. Generating probabilities and statistics
   iv. Visualizing the data (D3: https://github.com/mbostock/d3/wiki/Gallery)

IV. Natural Language Processing Strategies, Part-Of-Speech, Treeviews and Text Mining

a. Auto-tagging Parts of Speech: Tokenization
   i. Sinica Tagged Corpus Treeview: Intro http://turing.iis.sinica.edu.tw/treesearch/
   ii. Treeview Example: http://turing.iis.sinica.edu.tw/treesearch/treeview.exe?sTID=42486
   iii. Tagged Corpus Example: http://app.sinica.edu.tw/cgi-bin/kiwi/dkiwi/kiwi.sh?ukey=79872003&qtype=11&tid=0&swTag=1

b. Tagging and Markup strategies
   i. MARKUS http://dh.chinese-empires.eu/beta/
   ii. China Biographical Database Project (CBDB): http://isites.harvard.edu/icb/icb.do?keyword=k16229&tabgroupid=icb.tabgroup144476

c. Creating custom markup procedures

d. Text-Mining: Upscaling and Downscaling
   i. Google Books : n-gram Viewer https://books.google.com/ngrams
   ii. Bookworm http://bookworm.culturomics.org/
   Hathi Trust Bookworm Viewer http://bookworm.htrc.illinois.edu/

e. Digital Philology: Philologic4
   http://artflsrv01.uchicago.edu/philologic4/shakespeare_demo/