Workset 5

This workset is about working with texts on the **small scale.** We'll engage with larger scale analysis and topic modeling text week.

Choose a small-ish set of texts to work with. This may be an individual work, several smaller ones, or something in between; it should be between 10,000 and 1 million words, but you could go down to 3,000 if there's something you really want to look at. Choose them based on the first questions below.

Voyant

1. Upload your texts as a single document to Voyant-tools.org. Just screw around for a little while with the results. Try to obey Steven Ramsay's suggestion to think about what the results will be before you look. What are the most common words going to be? Will there be trends?

2. Upload your texts as *multiple* documents to Voyant tools using some *level of address.* You might lump together a few of the corrected versions vs uncorrected of scans; compare two authors whose works you can find online; etc. What are some of the *differences* between the levels of text? The statistics will come out easily.

3. What is one obvious functionality that an online portal like Voyant should have but doesn't?

Our course as a corpus

3. I've uploaded a copy of all of our reading responses into an instance of my own text analysis platform somewhat similar to Voyant, Bookworm. In class we (should have) looked at your blog post usage of individual terms. What did you find? Here is a link that tells you *which individual words* distinguish your writing: are they explicable? Do you believe them? What's present here that isn't elsewhere?

4. Finally, I've pasted below a topic model trained on the paragraphs of texts. Give it a reading. Where does that model *work*? Where does it fail? How might it be useful on a much larger corpus of posts? How would you want to see it plotted?

Topic label	Words
media McLuhan Manovich medium visual ways encourage	325.0
text analysis texts topic modeling Rhody Blevins	1544.0
slaves impact potential writers cliometric figures free	290.0
form cultural fully reproductions objects aesthetics freedom	256.0
contributors menus Menu dishes menu NYPL transcription	190.0
project process crowdsourcing transcribe transcription documents transcribing	1165.0
Marathon archive Boston Bombing story contribute contributed	227.0
data Fogel Cross Engerman Haskell slavery evidence	1847.0
initiative Illinois Bentham site Omeka workers Bracero	270.0
museum Presidential living Foundation museums galleries Obama	317.0
History history http com org year University	1255.0
buildings large windows building prominent built th	484.0
maps map modern location needed historical control	221.0

Topic label	Words
tags project media images music video frames	322.0
interest old city Cape came Caesar Classics	242.0
data logs captain's website record role helping	231.0
Ruggles quantitative conclusions Engerman's United American studies	566.0
access OCR copy scanned 1930 ease hurdles	247.0
value reproduction film piece original Benjamin's lost	380.0
Ramsay evidence computing House humanistic relationship Dream	327.0
recorded ages person data African age ship	261.0
Greek fragments papyrus Ancient identifying Coptic Lives	277.0
digital humanities Cecire field Digital Humanities Liu	1224.0
book level books easy ultimate presented outside	401.0
galaxy galaxies Bush Web scientists transport memex	154.0
year floggings whippings watch math 0.7 days	209.0
women agriculture efficiency Southern ibid Northern 1975	360.0
data analysis Gibbs scholars Owens methods qualitative	549.0
video Trove translated crowd name articles Nazis	294.0
work use historical history digital sources works	3496.0
text highlight website created multiple idea different	213.0