Workset 6: Text analysis

For the second unit as well, I want to try a kind of group workset. First, we'll divide into three or four groups of four or so around a specific corpus and problem; then we'll talk about it.

I also want to emphasize again–this will not necessarily be easy, and much may be completely impossible if you don't come to office hours! Please do! Don't wait for your group to meet before you come see me!

The goal here is threefold:

1. Leverage teamwork so you can experience what it's like to undertake a text-analysis project, seeing where the major pitfalls and questions are without each person needing to do all the work that might entail on their own.
2. Engage in a discussion of what some interesting questions might be *from a particular corpus*.
3. Start trying out some forms of group collaboration before the project proposals are due in early November.

Instead of e-mailing the workset over the next week, we'll take forty minutes at the start of class on October 30th (in two weeks) to go over each of the three groups. So this will be a sort of a grouped sprint[1] towards a preliminary mini-works-in-progress conference.

The deliverables at the end will be:

1. Someone should be prepared to talk for 10 minutes, probably with slides. Ten minutes won't be long enough to describe everything you've done–be succinct!
2. About 100-150 words on paper describing who did what.

Possible Corpora

Each group should choose a different method

Possible corpora include:

1. Wordcount data from JSTOR data for research
2. Wordcount data from some subset of the Hathi Trust
3. Texts from the Internet Archive
4. Newspaper articles from the 19th century United States
5. Some smaller but fairly *comprehensive* corpus, like the one Mullen and Funk use.
    - For example, I've done some work with the State of the Union Addresses.
    - There are Hansard collections of parliamentary speeches in the UK or Australia over two hundred years.
    - Or build off of some of the documents scanned last week. You could try to generalize about OCR quality, say, using ngrams counts.

[1] In real life, the only actual grouped sprint in the world is the three-legged race. Ideally you'll look like the 1-mile relay team handing off perfectly, but don't worry if this ends up with your whole group face down in the mud, as long as you can explain how.

Possible methods

Each group should choose a different method unless you've got a great justification

- **Topic Modeling** You should be well familiar with it at this point. Actually running the model can be a bit of a pain, but there are detailed step-by-step instructions in a variety of places online, most notably the Programming Historian. Does not require a large corpus.
- **Word Vectors** This is similar to topic modeling in some ways: a hot topic from machine learning that's identifies structural categories of language use in texts. But it principally focuses on just finding new elements of vocabulary that are similar to other words. You could use this to define a topic in the texts, you could use to uncover patterns in language use, or you could use it to find words that show a certain orientation: it provides a route towards answering weird questions like "where do foods fall on the salty-sweet spectrum in 19th century newspapers" or SAT-style analogies like `man:king::???:queen`. Requires a pretty large corpus.
- **Text Reuse** Just like Mullen and Funk. If you can think of another type of texts where it might work well, this could be an easy chance to re-use Mullen's code, which tends to be clear and well-documented.
- **Named-entity extraction** Like Blevins in a couple weeks, you can use the Stanford Natural Language Toolkit to extract lists of names and places out of corpora of texts. Once you have the list, someone in the group can plot the locations on a map, as well. This will seem from a distance like it's easier.
- **Sentiment analysis** This remains a weirdly hot topic in English department structural analysis of texts. Try to track the existence of position or negative language in a corpus. There are solid tools for sentiment analysis out there: one good one is Matt Jockers R package `syuzhet`. If you can think of a dataset where sentiment matters. An surprisingly open question in digital humanities research right now is: are there differences in the sentiment trajectories between different genres?
- **Other** Get in touch with me if you have a corpus that you think might benefit from some other method.

Sample projects:

- Write a regular expression that captures 100 words around every use of the names of presidential candidates in elections between 1892 and 1920. Sentiment analyze them. Are there patterns of when positive language, negative language, or non-emotional language is used relative to the life-cycle of the campaign?
- Topic model a different field from JStor data than Underwood and Goldstone did.
- Figure out how to load the data from the Hall, Jurafsky, and Manning article and trace how a few different words change their context. (This might be hard, so you'll have to come by!)

Roles

You'll have to figure out how to divide stuff up for this, which is why the set is not due for two weeks.

But some potential roles are:

- The project manager.

- The corpus builder.
- The corpus cleaner.
- The algorithm runner.
- The result inspector.
- The talker.