Texts as Data: Problem Set

Ben Schmidt

March 5, 2015

I know I said in the syllabus that there wouldn't be a problem set over break.

Instead, work on building up your own data set of texts or numbers: post some summary statistics on the blog, if you haven't, and think about adapting one of the successful random-walk generators to work with your texts.

That can still be the case, if you wish:

What this will do is work you through the tasks of building term-document matrices where the documents aren't necessary full texts, but are the *contexts* around individual words. This will let you use any of the tools from our course not on just texts, but on *contexts* for a word.

1. Build one of the concordance frames we talked about last week on your own set of texts.
2. Choose a list of words you're interested in, and then reduce the frame down so that it only includes phrases that have your word in the middle position.
3. Use `tidyr`'s gather function or a *do* function to reduce that frame down so that it just has one column of keywords, and another column which is each word that appears within two of that word.
4. Reduce that down to some of the most common words: visualize as a heatmap using the `geom_tile` ggplot function.
5. Spread that out into a term-document matrix–plot some of the same things we looked at in the class section.