Workset 4: digitization

Goals

1. Improve your understanding of gaps in the digitized record by finding some;
2. Gain firsthand experience with digitization practices by bringing a text through from physical to digital form.
3. (If so inclined) expand on regular expression knowledge for cleaning texts.

Part 1: Inventorying non-digitized sources

1. Reflecting on what we've read, what you know from past experience, what we've seen at the library, but *without doing any Internet searches*, write down three *printed texts* that you think may have escaped the dragnet of digitization.[1] At least one should be a book; I'll be happy if all three are, but you can also choose magazine articles, broadsides, newspaper editions, and so forth.

2. Check (figuratively and literally) whether your three texts exist in digitized, analogue, or physical forms in the following places. Put, let's say, "A" for analogue, "D" for digital, and "N" for neither.

Library	Source 1	Source 2	Source 3
Northeastern			
Boston Public			
Any library in eastern MA			
Any library in Worldcat			
Hathi Trust			
Google Books			
Internet Archive			
A domain-appropriate engine			
Google primary search engine			

3. During Thursday office hours, someone will set up a Google docs spreadsheet to track these works. Post a citation of some sort (don't worry about formatting) for the texts that you are confident don't exist online, but which *do* exist in a Boston-area library or of which you have a physical copy.

Part 2: Digitization

Take responsibility for one of the texts identified in step 3. (Obviously you have first dibs on your own). This step will be much easier if you work with others. (Only two people should need to go to the BPL, or Tufts, or wherever.) I'm happy for any three people to share a single book.

4. Scan it to an image format. If it's a whole book, you can do just a couple chapters, or split up with someone else.

1. By printed texts, I mean, affixed to paper by a printing press of some sort. By "escaped the dragnet," I mean: physical copies exist in public collections, but there aren't downloadable copies on the Internet.

5. Perform OCR on the image. You can use Adobe Acrobat, Bill Turkel's instructions, or some instructions for OS X from the last time this class was offered.
6. Export the OCR'ed text as a `.txt` format; upload to the course web page as a *page* (not a post) as a subcategory of readings.

Choose one of the following two.

7. **Either…** Clean it up until you've fixed 50 mistakes in the main text or the whole document is spotless, whichever comes first.
8. **… or** Come up with some regular expressions a la Bill Turkel's post to help clean up *general* OCR errors. We can set up some web site to automatically apply it to everyone's texts. To take this tack, you can either write regular expressions, or just point out regular errors in the texts that might be solved through regular expressions and persuade someone else to do it. This latter one is obviously harder. But if start to get a decent list, I'll make a webpage that bundles them all together onto which anyone (our course or not) can paste texts to get corrections. If you want to do this, start letting people, including me, know.

Suggestions

I would very much like someone to load in the following: Publishers' Weekly, March 1956, pp. 1274-78, "Editing at the Speed of Light." That's an article about our angry priest from the web page.