Instructions for creating page-level training data:

(Originally written by Mike Black, with additions by Boris Capitanu and Ted Underwood)

Things you need:

- Java 1.6 or 1.7 on your machine -- probably already there, and if not, a campus machine will have it.

- A metadata seed file (metatable.txt).
- Allowable genre codes file (codes.ini).
- A volume folder containing the text of books to be tagged
- An .arff file listing the books in that folder.

Installation and First Time Use:

To install, place the jar file in the directory where you'd like the browser to create its database and configuration files. The first time that the program is executed, it will prompt users for a metadata seed file and prompt them to name the database that will be created from it. (A name like "Database" is fine.) This database is not going to contain the pagemaps created by your tagging; rather, it contains underlying metadata for the larger 470,000 volume collection we're working in.

The browser will then populate the database from the seed file. This can take anywhere from 5 to 20 minutes depending on the speed of the drive you're running the program from. Currently, there is no progress monitor (and on a Windows machine, there will be no icon to indicate the program is still running). So some patience and trust is required here. The browser window will load once the database is populated. When you exit the program, it should write a configuration file that will keep track of the database and make note of which file was used to populate it. So the next time you load the program things will be faster.

IMPORTANT: The browser must be exited from the window close button (red dot in the upper left on Mac, 'x' in the upper right on Windows). Forcing it closed using hotkeys or the MacOS application menu will not trigger the close listener that resets the database to its default state or update the configuration file. If not closed properly, the database may become unstable, although closing it properly in future sessions should cause the database to reset correctly. (In other words, if you forget to close it properly, run it and then close it the correct way immediately. Things should reset).

Loading a set of volumes to be mapped:
Each time we distribute a set of volumes to be mapped, we'll give you a folder (named for instance "subset1") and a list of volumes in that folder (named for instance "subset1.arff"). The .arff suffix is related to the Weka data mining toolkit, but for our purposes this is just a list of volumes.

Before you can map individual volumes, you'll need to click "Load" at the top of the browser window and select the .arff file for the folder you want to map.

Then you need to click "Load source records" in the middle of the browser window. At this point a list of volumes should appear which matches the files in your subset folder.

You can select individual volumes to map.

Working with Page Maps:
In order to use the page mapper, users will need a list of allowable genre codes and labels. Upon first loading the page mapper, users will be prompted for a directory containing the volume text files for a mapping session. This choice will be remembered until the Genre Browser is closed. Users will also be prompted for a genre codes file. This choice will be remembered and stored in the browser's configuration file for future sessions. All completed page maps will be stored in the browser's local directory in a subdirectory "pagemaps/" (created by the browser as needed).

This last point is important because as you get new subsets the list of maps in pagemaps will keep growing unless you "move things out." Probably when you ship a set of maps off to Shawn you'll want to "archive" the old maps by renaming the /pagemaps folder something like /mapsforset1. Then when you get a new subset the browser will create a new /pagemaps folder to store the new maps.

When you open a new volume, page zero will be displayed (it's usually blank). You can select a code for each page from a pull-down menu; then click "store code" to label the page. Generally, blank or ambiguous pages at the front of a volume are "front matter," and blank or ambiguous pages at the back of a volume are typically "back matter," unless they're material added by the library (date due slip, etc.) for which we use "libra." You can use the "scan" button to add the same code to a whole range of pages. In theory the scanner should stop if the page format changes dramatically, but it may not always stop at the back. You should probably check the back of the book manually to make sure the last few pages are correctly categorized.

Here's a list of the genre codes available in the browser. I know there will be a lot of ambiguities in practice: cases where you're not sure what code to assign or where you perhaps feel a need for additional codes. Record any interesting anomalies you discover, and we'll have meetings to discuss them. But also feel comfortable

guessing: each page is going to be evaluated by three people, and we've got a system for ironing

Genre codes

At the page level, we want to choose <u>one</u> of the codes below for a given page. When we move up to characterizing whole works, classification is going to be a more fluid process, where works can belong to multiple genres to different extents. But at the page level, we want to use classification to segment works, so it's "choose one and only one of the following codes."

Structural codes (like "front matter," "table of contents," etc.) are fairly straightforward. When you get into literary genres there's more squishiness here, because I've often provided specific codes (verse drama and prose drama) as well as a more general code (just "dra.") Make the classification as specific as you can *without having to turn a single work into a mosaic of many different codes.* E.g., if a drama mixes verse and prose, don't try to tag each page appropriately (*vdr* / *pdr*); just tag the whole thing as *dra.*

When I use these codes to train classifiers, I'm going to be aware that "verse drama" and "prose drama" are *parts of* a larger category "drama." The process is designed to echo Google's process for filtering out spam advertisements, which includes classifiers trained to recognize specific narrow kinds of spam as well as a more general classifier that tries to cover the whole boundary (with less precision).

Can be applied to volumes or pages (for the moment, we're characterizing pages):

non	nonfiction prose that doesn't belong to the more specific categories below, and isn't a "preface" located before the table of contents.
aut	autobiography
bio	biography
trv	nonfiction about travel
ora	orations and sermons (that seem to have been actually delivered orally)
let	personal letters (that were actually written as correspondence, not rhetorical "open letters" to the Bishop on his new policy of etc. Also, not epistolary fiction. *Let* is actually a pretty rare category.)
	---
fic	prose fiction
epi	epistolary fiction (written in the form of correspondence)
	---
poe	Nondramatic verse. If a long series of pages is clearly a narrative poem, use "nar." If it's clearly a collection of lyrics, use "lyr." But if it's ambiguous or it keeps going back and forth every three pages, just use "poe." Verse that's part of a drama meant to be performed should not be tagged "poe." In categorizing page ranges, it's okay to apply this code when the pages are actually a mosaic of verse and prose footnotes. We'll get those

	footnotes later. You can also apply it to poems inside works of prose fiction, or inside nonfiction, if a given page is more than 50% verse.
lyr	lyric poetry
nar	narrative poetry
clo	A "closet drama" -- poetry in the form of dramatic dialogue that doesn't look like it was ever really intended to be performed. The line between closet drama and verse drama meant to be performed is blurry; context and title pages provide a clue, but you may also have to guess.
	---
dra	drama (could be verse or prose, or mixed)
vdr	drama written entirely or almost entirely (90%) in verse
pdr	drama written entirely or almost entirely (90%) in prose

Additional codes that can only be applied to pages:

title	title page
pref	preface, dedication, so-called "advertisement," or other prose in front matter (*only use this code for prose that is located before the table of contents; an introduction after the table of contents is just "non")*
impri	A page following the title page that lists the authority for publication. E.g., "Entered according to act of Congress in the year …"
bookp	A library bookplate, usually in the front of a book. Generally, if the words "University" or "Library" appear on a page without much other text, you're looking at a bookplate.
libra	Text added by the library other than a bookplate -- especially, for instance, a due date slip or library information at the back of the volume.
toc	Table of contents
subsc	List of subscribers, or any page that's mainly a list of names.
front	any front matter not otherwise categorized
index	Index
notes	Endnotes to the work; for instance notes that follow a poem. These may not necessarily always be at the back of the book.
gloss	A glossary.
bibli	Bibliography.
ads	A publisher's catalog of titles, or other ads.
errat	Errata slip.
back	Back matter not otherwise categorized.
argum	Prose argument preceding a poem. You can also use this for other short pieces of prose associated with / mixed into verse or drama. (Except endnotes, which should be tagged *notes*.)