Jasmine Wong

GENERAL STRUCTURE

For my final data file, I plan to pull the top 5%* (rounded up) of authors from each of my 5 datasets and put them all in one spreadsheet alongside their count weight (calculated relative to their respective dataset). The final dataset should have about 1520* records, with 4 data points for each record: first name, last name, dataset, and the author's fractional weight in that dataset (calculated by dividing the author's count in that dataset by the total number of counts in that dataset). The dataset will be arranged by fractional weight, descending from greatest weight to smallest.

Although I collected birth/deathdate data, I don't intend to include it in this final dataset because that element is less consistent across datasets. Ideally I could create another dataset that organizes the authors by the periods in which they wrote, but I don't think I have time to create that dataset in addition to my primary dataset of interest.

DIAGRAM (demonstrating structure only)

last_name	first_name	dataset	fraction-total (author count / total dataset count)
Surname1	Given1	eebo	0.5
Surname1	Given1	estc	0.4
Surname2	Given2	estc	0.3
Surname3	Given4	open-syllabus	0.2
Surname1	Given1	project-gutenberg	0.1

CONCERNS

I've been a bit concerned that I won't have time to completely clean all of the datasets, especially some of the larger ones. Now that I've chosen to only collect the top 5%* of authors, I can focus on those sections in each dataset and expedite the cleaning process. Ideally I'll go back and clean the rest of the datasets later, but within the scope of this project I think I need to compromise and focus on the records I'm collecting for my final data file.

I'm a bit disappointed I won't be able to go much further with the analysis beyond collecting the top authors and putting them in a file, but I also realize that's beyond the scope of the project and far beyond the scope of what I have time to do.

I'm not sure if my dataset diagram follows the tidy dataset guidelines – I'm still having trouble wrapping my head around transforming the datasets properly. I think my dataset diagram follows those guidelines, but I'm happy to change it if it doesn't follow tidy standards.

Jasmine Wong

TO-DO LIST:

- Finish cleaning datasets, focusing on the top 5%*
- Extract relevant records from cleaned datasets and aggregate to form final dataset (using Python? manually?)
- Conduct analysis?
- Form additional dataset (time permitting)?

* I'm not sure if there's a limit to how big our final dataset can be / guideline on how big it should be. The scaling for different percentage slices would be as follows:

- 5% = 1520 records
- 10% = 3040 records
- 20% = 6081 records
- 25% = 7601 records

Do you have a recommendation on what slice I should take?