Journal Report 9
11/11/19-11/15/19
Abby Kim
Computer Systems Research Lab
Period 2, White

Daily Log

Monday October 28

So... There is an important thing that I forgot about that I realized when I was reading The Namesake, and that is the idea of subjects and objects. For example if character X makes character Y cry, the action "crying" is associated with both characters, but has entirely different connotations. Luckily one output file (book.id) has tags for nouns in the forms of "agent" (subject) and "patient" (object).

Tuesday October 29

Read in useful stuff (agent/patient, names, words) from the book.id file and did a similar thing to when I was working with the tokens file, just adding another bucket to account for dividing agent/patient words.

Thursday October 31

Since I took care of the agent/patient thing alright, I started playing around with dialogue. I kind of ignored it before since I didn't know what to do with it, but I created a new dictionary for name:[quote1, quote2...] and then applied the same process I did for the whole text. Not sure if it's useful, but I'll come back to it later.

Timeline

Date	Goal	Met
Two weeks ago	Figure out how to run other books with the pipeline	correct :)
Last week	Dealt with objects and subjects	yessir
This week	Graph the data points generated from the previous week and see what it looks like. Also clean up code and making it entirely command line based	N/A
Today plus 1 week	Train model on new books (from same time period) and normalize writing styles	N/A
Today plus 2 weeks	Curate lists of books using	N/A
Winter Goal	(Vague) Have data for 100? books per era and some cute graphics explaining it (Incorporate findings from annotated book file, normalize writing styles, what they say)	N/A

Reflection

This Week - The git hub jaunt made me realize I really ought to clean up my code so it's run entirely from the command line (which is easy to do but I have been putting it off). Speaking of things I've been putting off: graphing. I don't think it's a big deal that I'm not doing that yet because I'm still being productive and I'm more productive when I want to do what I'm doing. The whole subject/object thing (explained Monday) gave me "quite a fright" (that phrase is used wayyyyy to much in the 1700s) but not hard to fix at all.

Normalizing Writing Styles - Another thought I had, and I've decided to deal with this by just training the model on books from the same time period. Not a perfect science but generally language develops in a way that two people from the same time period are more likely to write similarly than two people from different time periods. Another way I could maybe deal with this is taking percentages. (Ex: number of "action words" when the character is subject (agent)/total number of "action words") I'll probably do both actually, it doesn't make sense not to.

Selecting novels - I think I'm going to find the top 100 most popular novels for each era. I want them to be popular in their time period rather than what people now read the most since those books reached the biggest audience during their time. This will probably skew the data to represent the majority opinion so I'll also find 100 books representing a sort of "counter" movement (Ex. The American Woman's Home) or at least something just acknowledging that not everyone wrote the same things. The idea behind doing it this way is to trace "general America" and also the American feminist movement and how it evolved.