Week 7 Progress Report

DATA 450 Capstone

Andrew Nemkov

Week ending: 3/9/23

Time Log

Day	Time	# Hours	Task(s)	GH Commit(s)
Mon.	[5:20pm-8:15pm] & [10:45pm-12:30am]	[4.66]	[further data creation issues and testing]	Link 1
Tue.				
Wed.				
Thu.	[5:10pm-8:00pm]	[2.83]	[fixed creation issues with memory and processed half of Baki Dou 2]	Link 1
Fri.	[9:10am-11:30am]	[1.83]	[finished all data processing of 3 manga series]	Link 2
Sat.				
Sun.	[4:30pm-5:00am] & [6:45pm-8:00pm]	[1.75]	[organized files and started question 1]	Link 3

Total # of hours worked this week: 11.07

What you accomplished (or attempted) this week:

This week has been extrmely significant in my project progress. I was finally able to fix all my issues regarding data processing and creation. The problem was during OCR model code running, my computer kept running out of memory and the code kept either crashing

or stopping abruptly. I was able to handle this by reducing the bathc size of my code and pushing back calculation of word frequencies till the very end of data creation. After around a total of 6 hours including issues, with the speed of around 800 images processed per hour, I was able to successfully create my dataset with around 49,000 rows of information.

In addition, I began working on my first question delving into analysis of OCR complexity. I began by creation of my scatter plot for this analysis and am currenly workign on refining this graph before moving on to other parts.

Any setbacks/roadblocks you experienced:

I had a lot of difficulty fidgetting and figuring out data creation and computer memory optimization, but after a good amoutn of waiting and brute force approach I was able to correctly put together my data to be used for this analysis. Currently I have a light issue of my scatter plot for question 1 looking a little strange and not very informative, so I am takking steps to improve the visulization.

What you plan to work on next:

For the next week I plan to continue working on my question 1 and completing all 3 parts of this complexity analysis. The other two visualizations will involve the box plot and heat map of this portion of the project analysis. After this completion, I will work on creatingm my basic presentation for the in-person class on Thursday 03/13.

Are you on track? If not, how will you get back on track?

I have gotten back on track this week due to putting in a lot more effort on the project on Monday and Thursday. This allowed me to get over the hardest hurdle of this project to continue to actually answering the questions.