Week 5 Progress Report

DATA 450 Capstone

Andrew Nemkov

Week ending: 3/2/23

Time Log

Day	Time	# Hours	Task(s)	GH Commit(s)
Mon.				
Tue.				
Wed.				
Thu.				
Fri.				
Sat.	[10:30pm-1:00am]	[2.5]	[read in first of three series (Hanma Baki)]	Link 1
Sun.	[11:30am-12:30am] & [6:30pm-7:30pm]	[2]	[failed to finish data creation]	Link 2

Total # of hours worked this week: 4.5

What you accomplished (or attempted) this week:

This week I completed processing all of the jpg images from Baki1Hanma raw data folder. These images are all from the Hanma Baki manga series, the first of three series to be analyzed for this project. To correctly run in this data into the new realData.csv file, I adjusted the value of img_series. After this successfully processed, I continued to try to process the other two manga series images.

Any setbacks/roadblocks you experienced:

I hit quite a significant roadblock today regarding finishing processing all JPG images. It seems that the Paddle OCR model can't read or understand one of the images in the Baki2Dou2 folder, causing a run time error to occur after around 30 minutes of processing. I believed this originally to be a possible memory issue on my computer, but this is not the case as after implementing some basic code, the memory percentage is never exceeded. I will have to do further testing and fidgeting with the code, but most likely this is an issue with one of the JPG images in the folder.

What you plan to work on next:

The next week will require a significant amount of work regarding the finalization of data creation for all three manga series as well as starting to work on question 1 of the project for analysis of the relationship between OCR confidence and Japanese word complexity.

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

I am not on track this week due to the above significant setback. I will have to spend significanly more time next week working on this project to get back on to normal work pace. In addition, I believe it will be better for me to try to focus my attention onto this project for weekdays that I have time, most likely Monday, Wednesday, and Friday.