Week 4 Progress Report

DATA 450 Capstone

Andrew Nemkov

Week ending: 2/16/23

Time Log

Day	Time	# Hours	Task(s)	GH Commit(s)
Mon.				
Tue.				
Wed.				
Thu.	[4-7:30pm]	[3.5]	[revise proposal and add new questions]	Link 1
Fri.				
Sat.	[8:30-10pm]	[1.5]	[adjusted OCR parameters for improved text recognition]	Link 2
Sun.	[11-1:15pm]	[2.25]	[implemented basic data read-in code for test images]	Link 2

Total # of hours worked this week: 7.25

What you accomplished (or attempted) this week:

I began this week by making further adjustments to my code to specifically improve OCR text-recognition. I fine tuned the values of det_db_box_thresh and det_db_unclip_ratio as well as provide simple image processing code for gray scaling (turning the image to black and white) and applying simple denoising to remove complex background imagery. This slighty improved the amount of recognized text. In addition, I set a limit to the confidence level of extracted text I use so that non below 65% confidence will be implemented into the data.

In addition to the parameter adjustments, I added code to read several Japanese manga JPG images instead of just one, as well as read the extracted information into a CSV file called test_data.csv. This was for testing how to properly read in data into a CSV file, as well as getting a feel for the time and computational effort required for future data preparation process on the real raw data. The test CSV file contains extracted text of 5 images with three pieces of information. These are the specifc phrase extracted from each image, the image name from which the phrase came from, and the confidence level of the extracted phrase.

Any setbacks/roadblocks you experienced:

One specific roadblock that I struggled with was greatly improving the amount of text that the OCR recognizes. From testing, when performing OCR on a manga panel, a image that comes right before or after the main manga text pages and includes aspects such as title and background information, the OCR had a much harder time recognizing text from these due to the lack of speech bubbles. My adjustments made minor improvements, but overall the issue has not been solved.

What you plan to work on next:

For next week I plan to further develop my data preparation, now working with the real raw data images. Since this requires a lot of computational power, I imaging that I would have to make a way for the code to continually update me with its progress or maybe take short breaks between OCR extraction and actual reading into the CSV file. In addition, I plan to start working on refining and calculating all my needed variables for the proper dataset.

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

Currently, even though I have had difficulties and have experiences a roadblock, I am on track with my work, as most of the basic testing parts have been completed and all the minor parameters have been set and refined.