Week 5 Progress Report

DATA 450 Capstone

Andrew Nemkov

Week ending: 2/23/23

Time Log

Day	Time	# Hours	Task(s)	GH Commit(s)
Mon.				
Tue.				
Wed.				
Thu.				
Fri.	[10:15-11:15pm]	[1]	[Japanese phrase to word separation]	Link 1
Sat.	[12:00pm-1:00am]	[1]	[watched video of OCR from-scratch development]	Link 1
Sun.	[1:00pm-3:30pm] & [5:00pm-6:00pm]	[3.5]	[completed data testing and all columns]	Link 2

Total # of hours worked this week: 5.5

What you accomplished (or attempted) this week:

This week I completed the difficult data extraction and implementation on my test file. Sepecifically, I began by first taking my Japanese OCR extracted phrases and splitting them into words using the Fugashi tokenization. From this, I was able to add all my necessary columns of data for future analysis, these being: word_US, word_POS, img_series, length, word_freq, hiragana_ratio, katakana_ratio, and kanji_ratio.

Going into detail, the word_US column is an english translation of the word_JAP column of the OCR extracted Japanese word. The img_series or image series is at the moment simply the name of the parent file in which the specific image is located. This will be changed a bit to account for the three Grappler Baki manga series of the normal raw data. The word frequency is the amount of times the specific word appears in the overall word list in a decimal format. The final three column names are the Japanese alphabetical ratios of how many certain types of characters of an alphabet appear in the whole specific word, also in decimal format.

Any setbacks/roadblocks you experienced:

Most of the work this week was somewhat simple, needing only implementation of new libraries or use of already built-in python functions for column calculation. Considering this, the issue for this week was that the translation of the word_JAP Japanese word to English word_US was complicated. Originally I attempted to use the site Jisho.com and it's translation to give me English outputs, but this involved a lot of complexity regarding how many requests you can send to a site without getting blocked from it. In addition, the Jisho dictionary translations were often complex and long, not allowing for simple one-to-one language conversion. Therefore, I changed my strategy and instead implemented the use of Google Translate, which solved both issues significantly, as a result reducing the time needed to run the program on the test images, as well as gettin better simpler translations.

What you plan to work on next:

The next week will involve a very tricky part of the project due to me having to take this code of PaddleOCR.py and running it on the raw data of my three Grappler Baki series. This will take a lot of time, probably several hours of straight program running. I will have to see if I can implement ways for consistent updates to give me a check on the program's progress.

In addition, I might start working on my first question touching on the topic of the relationship between OCR confidence and Japanese word complexity. From this I would also touch on the subject of the limitations of using OCR models to specifically extract Japanese manga text from JPG images.

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

Even though I put in less time on work then compared to last week, I still am on good track for this project. Completing the technical code section is a good pausing point for this week to allow me time to properly finish my data and start working on analysis in week 6.