Journal Report 2
9/9/19-9/16/19
Bryan Lu
Computer Systems Research Lab
Period 2, White

Daily Log

Monday, September 9

I researched how to scrape an infinite-scrolling website. Later, I implemented a `BeautifulSoup` script combined with the `selenium` package in Python to scrape the first few problems loaded on the High School Olympiads forum page.

Tuesday, September 10

I attempted to scale up the performance of the script by running it for longer and longer periods of time. I researched how to prompt the script to scroll just the box in which the posts were loaded, instead of the entire page.

Thursday, September 12

I learned how to use the `requests` package in Python to make AJAX calls to the servers, and learned about how HTML requests work. I tried to replicate the AJAX call that the forum used to dynamically update the page with more posts.

Timeline

Date	Goal	Met
8/26	N/A	N/A
9/2	Scrape at least 400 problems from the AoPS website, through brute force and operating on the Contest Collections.	No, as earlier years have a Shortlist with only 3-4 relevant geometry problems apiece. I got in the neighborhood of 200-270 problems through manual scraping.
9/9	Scrape at least 1000 problems from the AoPS website – in particular, the High School Olympiads (HSO) forum.	No, but the scraper that I have is nearly functional. Up to scaling, my script can scrape an arbitrary amount of the problems on the page. I did not get many new problems this week.
9/16	Finish writing the webscraper to scrape arbitrarily many problems off of the forums, properly formatted. Begin the process of filtering posts from the dataset.	N/A
9/23	Filter posts that are not standard olympiad geometry problems, and construct a standard lexicon of keywords to look for in a problem.	N/A

Reflection

This week, I tried to create a legitimate webscraper that could scrape forum posts from the High School Olympiads forum on AoPS. The main challenge this week was to figure out how to get more posts loaded onto the page, as I had already written a piece of code that I could just modify slightly to scrape text from a `div` of a certain class in order to get the text I wanted.

I tried multiple approaches to get this auto-scrolling to work. In the first part of the week, I thought I could do this with a naïve approach that made the entire body of the page scroll down, but the way the forum is coded doesn't allow this to work, because the scrolling part of the page is embedded within the page. Thus, I tried to use the `requests` package in order to get more posts, which I figured out were being loaded with AJAX calls. I haven't yet gotten code using this package to successfully scroll the page, so I've reached out to the TJ Sysadmins who have much more experience than I do with this kind of scraping.

I plan on working with the Sysadmins during lunch on Monday and possibly Tuesday to figure out how to scroll the page successfully. In the meantime, I'll have to begin the process of cleaning my data set during the week before I have a lot of problems saved. I hope to get the problems that I have already saved trimmed down to a set of standard olympiad geometry problems that would lend themselves easily to a fixed diagram, and performing a frequency analysis on these remaining problems to get a set of the most common words/terms in these problems. This should prepare me well for the algorithms I'll need to code in the future.