Journal Report 1

9/2/19-9/8/19
Bryan Lu
Computer Systems Research Lab
Period 2, White

Daily Log

Tuesday, September 3

I used a primitive HTML parser I coded in Python to copy IMO Shortlist geometry problems from the Art of Problem Solving (AoPS) website to my computer. I saved the webpages that stored full Shortlist from the years 1985-2018 and ran my script on them to extract the relevant problems.

Thursday, September 5

I used my primitive parser to scrape all the problems from the IMO Shortlist (years 1979-1985) from the AoPS website. I also looked for alternative ways to accumulate a large quantity of problems quickly, including asking people who interned for AoPS if they could pull the relevant posts from the High School Olympiads forum from the AoPS databases.

Timeline

Date	Goal	Met
8/19	N/A	N/A
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.	
9/16	Clean up the problem statements, adjusting the dataset for problems that don't explicitly mention points or are posed in a three-dimensional context, and removing LaTeX formatting from problems.	

Reflection

I was able to use a small snippet of Python code using BeautifulSoup that took HTML code and gave me the problems I wanted from the page fairly well, and I thought I could get much closer to

400 problems this week, assuming the Shortlist from each year had about 7-8 geometry problems, with about 60 years of problems. It became clear that was definitely not the case when I was spending 2-3 minutes on each of the years from 1990 backwards only getting 3-4 problems, on average, from each year. This method definitely was not a successful one.

Instead, I tried to find another centralized source of geometry problems, one of which was the forum posts tagged with the "geometry" tag on the HSO forum in the AoPS Community. I had not yet built an actual web spider that could crawl through the results page and retrieve all the posts, so I saw if I could do it quickly by asking a former TJ alum that (temporarily) had a personal server at AoPS for testing. I found out this didn't work this weekend, because those servers were shut down while processing the request to the databases. Although I haven't thoroughly smashed my goal by getting all of these posts, I think I still have a decent shot at meeting my goal of 1000 problems this week if I can extract problems from this more reliable source. Thus, my goals remain the same for this week, if I can code and run a webspider to get all of these posts, which is equivalent to retrieving the desired problems.