Journal Report 4
9/23/19-9/30/19
Bryan Lu
Computer Systems Research Lab
Period 2, White

Daily Log

Monday, September 23

I looked through my data set and started to remove the problems that were not actually geometry problems or that were too vague and didn't contain references to specific points/objects.

Tuesday, September 24

I finished removing unusable problems from my data set by explicitly searching for capital strings in the line, and started to format the problems I had into a usable form by removing random blank spaces, expanding problems with two parts into two problems, etc.

Thursday, September 26

I finished formatting problems by inserting spaces between mathematical operators and symbols and wrote a quick code snippet to find the most common words present in the problems.

Timeline

Date	Goal	Met
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.	I was not able to get my webscraper to work, but I've successfully started to format the approximately 300 problems I actually have.
9/23	Filter posts that are not standard olympiad geometry problems, and construct a standard lexicon of keywords to look for in a problem both as objects and as relations.	Yes, I created two separate files for various geometrical objects and relations that I found that appeared in my data set.
9/30	Write code that creates a graph structure corresponding to the problem statement, with objects as nodes and relations as edges.	N/A
10/7	Research how to write code to create a log-linear classifier and figure out what features of the problem statement should be passed to it.	N/A

Reflection

This week I made decent progress by finishing up my lexicon of words that I would be looking for as well as polishing the problems I did have. I'm ready to start executing the method that I will use to parse the problems and figure out what relations are valid in the context of the problem, using a probabilistic learning method. There are possibly a couple of errant parentheses that will likely make interpretation a bit more difficult, but I'll figure out how to handle this next week. I'm also not sure yet how to handle different forms of a word, but this will something I will also handle next week when I refine my lexicon and identify the objects and relations I want to detect explicitly.

I didn't filter out the most common words in English, but below is the output file of the most common words in these problems. As expected, the vast majority of these problems revolve around a triangle usually called ABC, and a lot of them are A-centric, i.e. the problem is framed around lines through A and its opposite side, BC.

```
the 1227
and 842
of 628
that 409
be 345
a 305
Let 294
triangle 279
ABC 256
at 255
angle 241
= 234
is 232
point 205
to 200
on 195
BC 184
Prove 180
are 172
AB 158
points 153
line 148
respectively 145
A 144
The 128
circle 124
with 117
P 113
AC 113
lines 110
such 108
C 108
B 100
```