URMP: Report Week 2

January 27, 2020

Panayot Vassilevski

Austen Nelson

1 Progress

This week I made decent progress on cleaning and organizing the data I collected. I spent some time annotating my ingredient phrases with tags so I could put my data through a conditional random field to parse out the extra information and separate a phase into quantity, measurement, and name. The tool I used is CRF++. This tool doesn't have any API and only deals with external data files but I integrated the installed program into my rust data-flow. The accuracy isn't prefect but it is working well enough to begin analysis. My dataset has over 40,000 recipes with around 15,000 unique ingredients. With this information it was pretty trivial to create the bipartite graph between recipes and ingredients. I did some very basic exploration like looking at the most commonly used ingredients.

2 Plans

The next step will be to create the other data structures that I have planned and begin doing analysis on the data. I'm still doing research on the analysis that I am going to do and am meeting with my advisor on Thursday the 30th. I also plan on writing some other scrapers and spiders for recipes other than allrecipes. Even though the allrecipes database is large, it seems to be lower quality and geared towards writing recipes around products that companies pay the site to promote. It would be interesting to see if analysis differs between platforms.