URMP: Report Week 1

January 20, 2020

Panayot Vassilevski

Austen Nelson

1 Introduction

This is still the early stages of my research so most of my progress is reading related materials and coming up with a more constructed direction and plan. I read a few related papers and am in the progress of creating an annotated bibliography for self reference. In this process I have come across quality inspiration.

2 Progress

The first step in my actual research will be doing data collection of recipes. A few of the papers I read referenced a recipe database called ”open recipes” but when I visited the site the database seems to be removed. I have emailed the administrators to see if I can get access to the database but haven’t heard back yet. This isn’t a huge problem as I can scrape this data myself. I have started with allrecipes.com. This week I wrote a basic scraping tool and have accumulated about 20,000 recipes. For this project I would like to implement clustering algorithms designed for very large graphs so I would like to have 200,000+ recipes. To accumulate this data will take some time but it should be possible.

3 Plans

One of the first graphs I plan on constructing is an ingredient — recipe graph where each ingredient and recipe is a node and an edges exist between a recipe and an ingredient if the ingredient is present in the recipe. This edge will also be weighted with the quantity. This is a nice bipartite graph and should have some nice properties to investigate. Another graph I plan on constructing will be the ingredient co-occurrence graph with ingredients as nodes and edges between ingredients that show up in recipes together. The weight of those edges will be the number of recipes with the co-occurrence. When representing this graph as an adjacency matrix the result is a nice symmetric matrix that can be clustered using modularity functionals.

4 Potential Issues

Now that I have some starting data I can move on to the next portion of my project of starting to create these graphs. The first challenge will be processing ingredient lists of recipes into sensible abstractions. Originally I was considering using regular expressions to try and pull out the quantity, unit, ingredient, comments, and excess information from each ingredient but quickly realized this cannot be done. A more flexible and advanced option would be to create a formal grammar and write a parser to do this processing. This could potentially work but there are many edge cases to be considered and ambiguous or inconsistent situations would still be challenging to overcome. Currently I am considering a more probabilistic approach using machine learning. A natural language processing technique for this task I am considering is Conditional Random Fields. There are numerous libraries for this kind of task with training sets such as CRF++ that I could use but I might benefit from implementing a specific version for recipe parsing. NYT did this for their cooking database but it doesn’t look maintained. For the sake of learning I would usually like to do what I can on my own if possible, though.
My plans for next week include:

- writing more parsers to expand my dataset
- researching NLP techniques for segmenting and labeling my ingredient list
- researching more background research related to creative computation for culinary application

My advisor hoped that I would have the recipe-ingredient occurrence graph built by our next meeting on February 6th but this might not be possible if I have to use statistical modeling to wrangle and clean my data.