Ben Questions to Jin:

* What scraping methods have you tried?
* What scraping methods do you want to know more about?

Background:

So right now, I wrote scraping code using Beautiful Soup, which works. However, it is hard to maintain, and not the entire process is automated (I still have to do some clicking).

My plan is the switch to Scrapy, a Python module (?) that really makes things a lot easier.

(Ask the professor if he knows about Scrapy. From looking at his webpage, it seems he mostly uses R, so he might not even know what Scrapy is. If not, that’s completely fine.)

Current strategy:

1. Get a list of starting points for scraping (say <https://www.yelp.com/search?find_desc=cafe&find_loc=60637>. And replace 60637 by all possible zip codes). (DONE)
2. Scrape for website links first. So for each starting point, gather all links
3. Filter those links based on some criteria (and remove duplicates).
4. Scrape each link.

Questions to ask about Scraping:

How were you able to scale your scraping?

What cloud platform did you use (if any)?

How did you avoid being blocked while maintaining speed?

How were you able to keep track of when there’s a divergence between what you expect from scraping a webpage and the results of the scrape?

How were you able to keep track of when you “left off”? For example, if your scrape suddenly stopped, is there an efficient way for your scraper to resume where it left off, as opposed to going back to the starting point?

Questions about Data Analysis:

Do you know of any way to do this large-scale data analysis? I expect around 15-20 businesses for each zip code, and there are 40000 zip codes. I also expect around 30 reviews for each. Then that’s 18 million reviews. I know how to access the Midway supercomputer (I’ve done it before for computational biology to sequence genes). However, I know that a supercomputer does not necessarily improve performance at all unless the code is written in a certain way. I know how to write R code to run in parallel, but I have no clue how to do that in Python.

What are your experiences with NLP?