IRisk Lab Data Discovery and Consolidation

Week 15 Report

Rohit Valmeekam

What was accomplished
- ❖ To address issues related to subpages on Loopnet, such as listings and apartments, I refined my code to handle URL encoding more effectively. This involved ensuring that the scraper recognizes and navigates through various types of subpages on the Loopnet website.
- ❖ Extensive Input Testing:
    - ➢ Rigorous testing was conducted using randomly generated addresses to assess the adaptability and accuracy of the code. This comprehensive input testing helped identify and rectify potential errors in handling diverse input scenarios.
- ❖ Conducted extensive input testing
    - ➢ Randomly generated addresses and tested them with program:
        - ■ Identified errors with Google search detecting the web-scraping bot and blocking requests
            - ● Attempted to fix errors through a series of search request delays (randomly generating delay times between API calls to avoid bot detection)
                - ◆ Did not seem to get past Google's bot detection algorithm
                - ◆ Another approach involved altering request headers to mimic legitimate browser requests. This included modifying user-agents and other identifying information to appear more like a genuine user.
                    - ➢ Did not end up working

Next Steps
- ❖ Next semester I have a list of things I want to accomplish
    - ➢ Work on a CSS stylesheet for the database website
    - ➢ Finish Integrating the Loopnet Scraper and Google Map Business finder and send packaged product to Eli
    - ➢ Integrate my already-existing pipeline to support file type conversion within the database and to give a more extensive interface for next semester's researchers to integrate datasets into the database