My idea for the final project is to create a data visualization/network graph of SNL cast members and their non-SNL-related projects to analyze the frequency by which cast members include other SNL cast members and alumi in their creative projects. Although they were only minor parts, this visualization would show, for example, how in Andy Samberg’s Brooklyn 99, SNL alums like Fred Armisten played Mlep(clay)nos (the clay is silent) or Maya Rudolph played U.S. Marshal Karen Haas.

The data that I plan to use was scraped from snlarchives.net, an online archive of all things SNL created by Joel Navaroli (@snlmedia). The data was scraped and published on Kaggle and GitHub by Hendrik Hilleckes (@hhllcks, hllcks@gmail.com, blog.hhllcks.de) and Colin Morris (http://colinmorris.github.io/). The current database on GitHub is missing a few seasons of data so I will likely be using the scraping technique that the Hilleckes details in his GitHub repository for the project.

To build a foundation of understanding regarding snlarchives.net, I explored the website quite a bit, paying keen attention to organization and choices made. The home page is broken down into 10 categories: Cast; Characters; Commercials; Crew; Episodes; Guests; Impressions; Seasons; Sketches; and Updates. Each category is then broken down alphabetically, organized by first letter of first name (including mononymous names) except for the categories of Cast and Guest which use last name (unless the person is mononymous).

A picture containing text, posing, person, screenshot

Description automatically generated The SNL archive includes descriptions of every skit from “Live from New York” to the goodbyes and includes screen caps of each. A user could, for insance, search for a cast member like Kate McKinnon and get a list of every Impression and character that the cast member has done in addition to the seasons that they’ve been on SNL.

Another user might search for every portrayl of Hillary Clinton and get results which includes every actor who portrayed Hillary Clinton in addition to the individual appearances of each portrayl. When the archive was scraped by Hendrick Hilleckes, it was parsed into the following csv and json files: actors, appearances, casts, characters, episodes, hosts, impressions, seasons, sketches, tenure, titles. As far as I can tell, all of the data was kept intact.

Graphical user interface, website

Description automatically generated

Application, table

Description automatically generatedMy vision for this project is to take the file of cast members and use their names to scrape each of their IMDb pages. Then I’m going to attempt to visualize the overlaps for each cast member’s other projects. Since SNL has been on since 1975, with different casts every few years and new cast members every season, I am expecting the final dataset to be large. There have been a total of 159 SNL cast members. I plan on working with a subset of the data while I figure out how to do everything.

While exploring the cast.csv file, I noticed some odd columns, like “season\_fraction,” and strange formatting decisions, such as making the “first\_epid” dates into float values.

I’m sure that I’m going to have to do a good amount of data cleaning and organizing but I’m grateful that there is a clear place to begin this project.

A screenshot of a computer

Description automatically generated with low confidence