CS636 Fall 2019
Journey Scraping Project - DNA Research

Team Members
Sam Sadr Azdi
Eric Nersesian
Abhay Bhaga

Team Member Contributions
Overall strategy - Sam and Eric
Architecture design process of code - Sam, Eric and Abhay
Programming the R scripts - Sam, Abhay and Eric
Documentation - Eric

Major Challenges Addressed
Here is a list of the major challenges that we addressed:

- Figuring out the html structure of the website along with the number of years, journals per year, and article per journey, and the html tags that corresponded to the specific deliverables, ie authors, author email, abstract, etc.
- Finding R packages that helped us with the processes, and we ended up using Dplyr, StringR, Rvest, XML R packages.
- Figuring out how to pull the information for one article, we had to figure out how to loop through all journals for a given year, and then loop through all articles per journal.
- Figuring out how to think about user input, and how to handle exceptional situations like the user entering years that do not have journals or input that is not a year.
- Some strings we pulled were in a clean state for use, but some strings had a lot of additional characters and spaces that we needed to use gsub to clean up.
- How to export and convert different file formats of data for example how to export out html files from the web links and how to convert the html files to a text file format or how to export the strings together into a larger text file for final delivery.
- Understanding some of the final deliverables, for example do we need to create html files of all articles from all journals from all years.
- Ran into issues with the website thinking we are a bot as we are downloading too many html files in a short amount of time. The website stopped serving us data because we were sending requests too fast, and we had to initiate a delay with our data requests.
- How to find the current issue number based on the current month this also dealt with finding the last published issue for the year.
- Our solution to dealing with the website timeout is to pull one year at a time, since its easy to pull all the years as needed
- Teammate communication and coordination of work, expectations of regularly meetings are off