Sucheen Sundaram

Metadata Extraction Process

I am a student researcher for the William & Linda H. Frost Program at California Polytechnic State University at San Luis Obispo. My project this summer was to find metadata and organize it into a data frame. The goal of extracting the metadata from various repositories was to find more meaningful “search terms” for users to explore datasets looking by more complex or particular terms, such as classification task, author affiliation, size, delimiter, and more.

The process of extracting this metadata began with identifying what columns (search filters) I needed to enhance the searching process. My project lead, Dr. Hunter Glanz, began the work with scraped data from the CORGIS Project and the UCI Machine Learning Repository. He extracted the numerical variables of the datasets, character variables, and delimiter. I used the same framework to scrape my metadata as well. I decided on file size, author, and most importantly, tags. In addition to the basic fields, I also thought I should include some popularity metrics, including the number of downloads and followers.

After deciding on the desired columns, the next step of this process is to find metadata repositories that provide these fields. I searched around and I found Harvard Dataverse, re3data, Edinburgh Datashare, California Open Data, data.gov, and more repositories. I learned the R package *rvest* prior to starting the project so I could use HTML parsing techniques with CSS selectors to scrape the text from metadata. To make this process easier, I wrote a function that accepts a url and CSS selector as parameters and returns the scraped HTML-parsed text.

I then wrote the code to do the invidiual-scraping process for the metadata; these are functions that take in a URL as a parameter (URL to the dataset) and return the scraped metadata attributes that I could find. An example of this process is the following: Graphical user interface, text, application

Description automatically generated

This statement includes most of the desired columns (search terms) I want to extract from this repository.

Throughout this code, I called my *scrape\_rvest* function that returns the scraped content from a URL and CSS selector. After scraping each individual item, I then combined them into a data frame with one row, with one column for each search filter that I found in the repository. To address scraping the file information (i.e. name, size, downloads), I stored the columns and data in separate variables with prefix “file\_info.” After scraping the file information, my initial plan was to store it in a separate data frame and nest that data frame as another column in the metadata columns. However, it proved not to be feasible when reading it in from a .csv file in which the completed metadata data sets are stored. To solve this issue, I stored the data “laterally”. That is, I stored the file name and its accompanying metrics in separate R vectors and one would need to match the indicies to obtain an individual observation. Some example code is as follows:

