Webscrabing Doctors' Information from the United States Immigration website as a first step to track down doctors and facilities that have been harassing immigrants

Rutendo Madziwo , Maggie Szlosek , Rachel LaFlamme

Abstract

Ru's working on it.

Text based on plos sample manuscript, see *http://journals.plos.org/ploscompbiol/s/latex*

Introduction

Maggie's

The research question(s) Background/significance of the research

Method

We collected the doctors' data using the package rvest for web-scraping, RSelenium for web navigation and an external platform Docker for virtual interaction with the web browser.

Docker is a platform to develop, deploy, and run applications inside containers[**???**]. We were able to virtually interact with the USCIS website by connecting to a port in Docker and opening Chrome and this enabled us to control and see what was happening on the website at a given time. Initially, Docker was installed before installing and loading the needed R Packages. The command `docker run -d -p 4445:4444 selenium/standalone-chrome` was then run in the R Terminal after we had installed our packages. This command sets up the virtual Chrome container to enable interaction with the Chrome web browser. In order to check if Docker is running, one can type in `docker ps`. We eventually open the browser using RSelenium commands before scraping our data.

RSelenium is a package in R which helps one connect to a Selenium server. This server in turn connects to the Chrome web browser and hence allowed us to automate our webscraping experience. RSelenium is responsible not only for opening and closing the browser, but it allowed us to virtually navigate the web page and automatically control the scraping. This was especially useful as our website had no endpoint urls and hence could not rely on more traditional web scraping methods. In addition, it made the process of scraping the data faster as one can simply allow the code to run and scrape multiple pages without needing to manually click the specific website.

While RSelenium was responsible for most of the web manouvering, the package we used for scraping the data from each of the pages was `rvest`. This package makes harvesting data from a website easy as it can find specific html nodes, and their children. It also allows one to use both XPaths and CSS selectors so though we eventually stuck to using basic elements, we were not limited to one option. As a side note, we chose to use CSS selectors for web navigation with the RSelenium package.

We created a function to scrape this data and took advantage of the purrr package in R to map all our scraped elements together. To clean up our data, `dplyr` and `tidyverse` were used for text-processing the such that zipcodes were in a separate column from the rest of the address in the resulting doctors' dataset. At the moment, this function is running in a for loop but will be converted to a while loop in order to allow for different state scenarios.

The final code written to collect the doctors' information allows a user to input one zipcode at a time in order to scrape data. Once that zipcode is entered, the doctors and facilities on that web page are harvested using `rvest` before moving on to the next page. Clicking to the next page has been automated using `RSelenium` and a for loop was implemented in our code such that for a certain number of times, the website's `Next` button is clicked, moves on to the next page, scrapes that page and so on. The website itself has been written in such a way that an actual user can keep clicking to find the nearest doctors within a 500 miles radius. As such, we have also manually entered different zipcodes in different parts of the USA so as to capture all the doctors in the country and create different datasets.

Maggie's

A discussion of the research, the limitations of the current research, reasonableness of any assumptions made, possibilities of future work/studies that should be conducted, etc.

Ru's

The title of the project and a one-paragraph abstract of the entire project with recommended length of no more than 150 words.

Here are two sample references: [1,2].

References

1. Feynman R, Vernon Jr. F. The theory of a general quantum system interacting with a linear dissipative system. Annals of Physics. 1963;24: 118–173. doi:10.1016/0003-4916(63)90068-X

2. Dirac P. The lorentz transformation and absolute time. Physica. 1953;19: 888–896. doi:10.1016/S0031-8914(53)80099-6