Development Notes

2023-07-18

## Introduction

These are the development notes for scraping programs data from topuniversities.com, and rankings data from YY. Both websites present data as dynamic content. That means that we need a headless browser to simulate the user clicking through the websites.

The best tool for that kind of scraping in the R ecosystem is RSelenium (LINK).

## Installing RSelenium

RSelenium requires the Rjava package. Rjava, in turn, requires access to the Java runtime. I recommend installing [Azul Java](https://www.azul.com/downloads/#zulu). Instructions vary with OS.

We also need a browser that works with RSelenium. The provided scripts work with Chrome.

## Scraping topuniversities.com

To start the scraping, the scripts uses RSelenium’s rsDriver to connect to a browser.

The relevant data from this site is information from the programs. We scraped data for the subjects of Biological Sciences, Computer Science and Information Systems, Data Science, Genetics, Mathematics, Medicine Related Studies, Pharmacology, Pharmacy and Pharmacology, Physics and Astronomy, Statistics and Operational Research in Bahrain, Cyprus, Egypt, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Palestine, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates (UAE) and Yemen.

As a starting point, we navigate to the url that encodes a query for those programs. The website presents the data in pages of 25 elements.

This table features data from the programs, tuition and location data. Since the requirements are not available in this page, we save a link to the program’s page on topuniversities.com to scrap the data from requirements and tuition in a second pass.

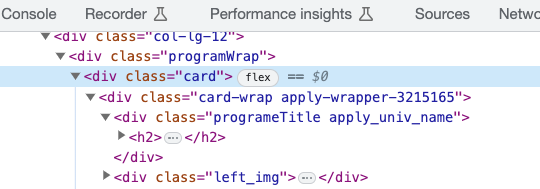
The script for scraping the initial program list is scrap\_programs.R. This script has multiple functions to achieve this objective:

* extract\_row
* scrap\_page
* scrap\_programs. Navigates to:

[https://www.topuniversities.com/programs/bahrain/biological-sciences?country=[OM,PS,AE,BH,IR,IQ,JO,SA,KW,QA,SY,YE,CY,TR,IL,LB,GB,EG]&subjects=[462,468,4049,4055,494,496,554,500,502,508]&pagerlimit=[25]](https://www.topuniversities.com/programs/bahrain/biological-sciences?country=%5BOM,PS,AE,BH,IR,IQ,JO,SA,KW,QA,SY,YE,CY,TR,IL,LB,GB,EG%5D&subjects=%5B462,468,4049,4055,494,496,554,500,502,508%5D&pagerlimit=%5B25%5D)

This is the starting point of the scraping task. From this website, the scrap\_page function scraps pages one by one and saves the 25 results as R objects to the data folder.

To scrap the data, it is necessary to understand the structure of the generated markup in each page. In this case, the data is presented inside a div element with a card class. This code cards <- remDr$findElements(using="css", ".card") finds all such elements. The helper function extract\_row is applied to all found cards to extract all the available data.



After scraping all the pages, we merge all the programs into a single data frame:

files <- list.files('./temp', pattern = ".rds", full.names = TRUE)  
  
# read files and merge  
programs <- files %>%  
 as.list %>%  
 purrr::map\_df(~readRDS(.))  
   
# save  
saveRDS(programs, './data/programs.rds')

The script scrap\_programs\_tution\_fees.R takes the list of urls scraped previously and downloads the tuition fee and requirements data from the program pages.

Since these are static html pages, it is not necessary to use a browser to access this data. The code in this script is straightforward.

## Scraping timeshighereducation.com

The relevant data from this website is in two pages:

The website also renders data dynamically, so we connect to it through a browser. The data is presented in a single page, so it is easier to scrap. The script that does this job is scrap\_rankings.R.

## Geocoding and merging

To geocode the downloaded data, I used the Google Maps API (Link) and the tidygeocoder package. The script that does this is geocode\_programs.R. To access the google maps API it is required to have an API key.

The final data for the Shiny app requires merging the data from different sources. Some of the university names are different, so a little bit of manual processing was necessary. That is achieved in the merge\_the\_topuniversities.R script.

The final data is in the data\_output folder.