Development Notes

2023-07-18

## Introduction

These are the development notes for scraping data from [Top Universities](www.topuniversities.com) and [Times Higher Education](www.timeshighereducation.com). Since both websites present data as dynamic content they require the use of a headless browser to simulate the user clicking through the sites to generate the content. The best tool for that kind of scraping in the R ecosystem is [RSelenium](https://docs.ropensci.org/RSelenium/).

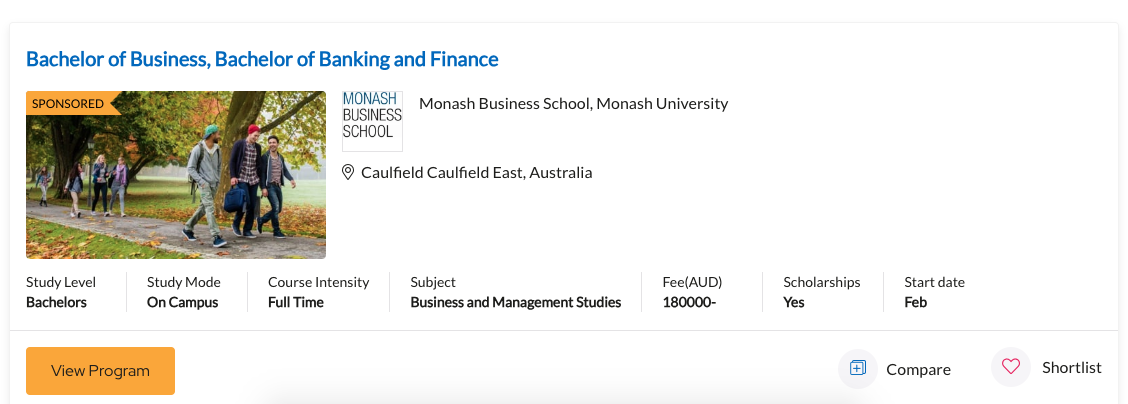
## Installing RSelenium

RSelenium requires the Rjava package. Rjava, in turn, requires access to the Java runtime. For Windows machines, Detailed installation instructions can be found in [this tutorial](https://www.youtube.com/watch?v=GnpJujF9dBw).

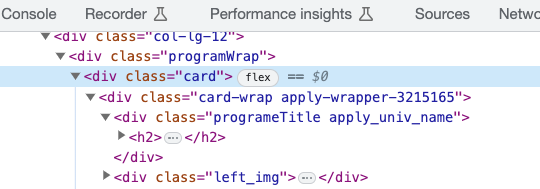
We also need a browser that works with RSelenium. The provided scripts work with Chrome.

## Scraping Top Universities

The topuniversities.com websites provides a [programs finder](https://www.topuniversities.com/programs?qs_qp=topnav). We used this tool to extract information for programs related to Bioinformatics, Machine Learning and Data Science. This is how the website presents the data:



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:



The findElements method finds all elements with a class of card. The helper function extract\_row is applied to all found cards to extract the required data using css selectors.

To start scraping the website, we use RSelenium’s rsDriver to connect to a browser:

Once the connection with the browser is established, we navigate to the starting url.

This url encodes a starting query of the programs we are interested in, filtering programs with 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. The website presents the data in pages of 25 elements:



The scripts queries one page at a time and downloads the data for each. The function scrap\_page(i) scraps page i from this query. The 25 results from each page are saved to the temp folder.

Since the requirements information for each program are not available in this page, we save a link to the program’s page to scrap the data from requirements and tuition in a second pass.

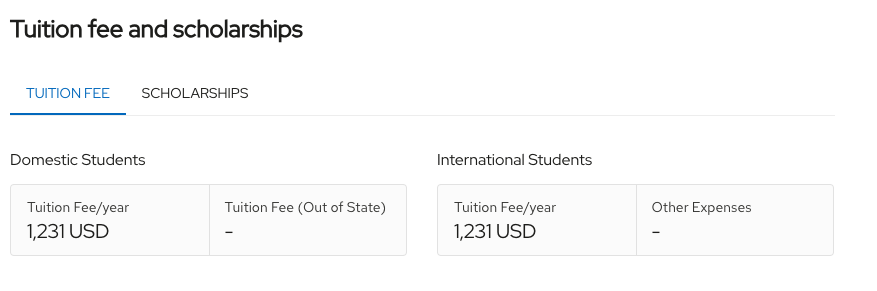
After scraping all the pages, we merge all the programs into a single data frame and save it:

### Scraping the requirements and tuitions data

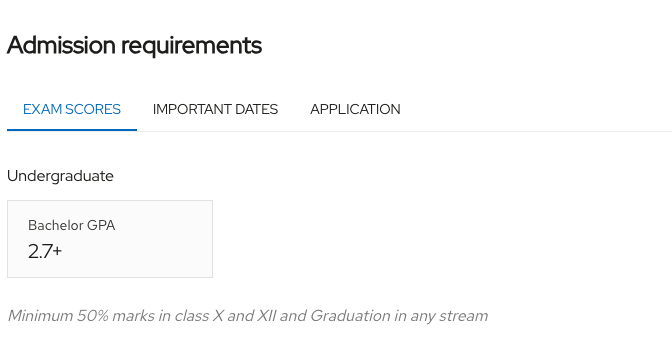
The script top\_univeristies\_scrap\_programs\_tuition\_fee.R takes the list of urls scraped previously and downloads the tuition fee and requirements data from the program pages.

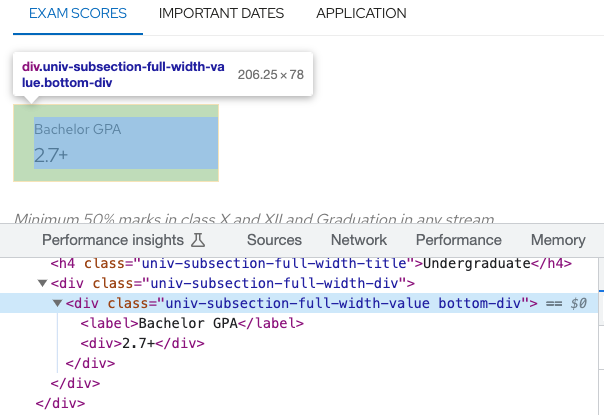
Since these are static html pages, it is not necessary to use a browser to access this data. The html pages are downloaded with rvest::read\_html.

This is how the webpages presents the tuition fee data:



And the requirements:

 The associated markup:



To extract the necessary information, we download the html page using read\_html, and then we specify css selectors that have the information we need to the html\_elements.

### Cleaning up the data

After scraping the data from all the programs, it is necessary to clean it by removing duplicates, parsing the requirements information and converting tuition fees to the same currency (British Pounds), removing programs that are not related to Machine Learning and Data Science.

Parsing each programs requirements involves taking the string from the html webpage and converting it to columns in our dataset. Each column corresponds to a requirement (TOEFL, IELTS, etc.). The value for each program in the column represents the minimum requirement of the program.

Currencies reported for the tuition fees are reportes in Euro, US Dollars, British Pounds and Singapore Dollars. I used the following exchange rates to convert currencies to British Pounds:

|  |  |
| --- | --- |
| Currency | Exchange |
| USD | 1.30 |
| EUR | 1.16 |
| SGD | 1.72 |

The final dataset is located in '../data\_output/programs\_geolocated.rds' and has 5356 programs. Here are the basic stats from the dataset:

Data summary

|  |  |
| --- | --- |
| Name | df |
| Number of rows | 2405 |
| Number of columns | 35 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 11 |
| numeric | 24 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | empty | n\_unique |
| university | 0 | 1.00 | 0 | 141 |
| requirements | 0 | 1.00 | 0 | 921 |
| study\_level | 0 | 1.00 | 0 | 3 |
| location | 0 | 1.00 | 611 | 146 |
| program\_title | 0 | 1.00 | 0 | 2098 |
| subject | 0 | 1.00 | 0 | 36 |
| study\_mode | 1730 | 0.28 | 0 | 3 |
| course\_intensity | 1248 | 0.48 | 0 | 2 |
| duration | 383 | 0.84 | 0 | 23 |
| currency | 1307 | 0.46 | 0 | 3 |
| duration\_units | 383 | 0.84 | 0 | 1 |

**Variable type: numeric**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd |
| fee | 1307 | 0.46 | 21074.80 | 14653.51 |
| duration\_length | 383 | 0.84 | 35.63 | 17.07 |
| fee\_gbp | 1307 | 0.46 | 20623.15 | 13438.83 |
| toefl | 0 | 1.00 | 50.29 | 41.49 |
| ielts | 0 | 1.00 | 6.03 | 0.47 |
| bachelor\_gpa | 1778 | 0.26 | 2.48 | 0.54 |
| cambridge\_cae\_advanced | 1479 | 0.39 | 170.86 | 20.22 |
| pte\_academic | 1212 | 0.50 | 58.41 | 8.19 |
| a\_levels | 1581 | 0.34 | 13.18 | 33.59 |
| international\_baccalaureate | 1439 | 0.40 | 27.55 | 9.41 |
| ucas\_tariff | 1758 | 0.27 | 92.81 | 34.26 |
| atar | 2159 | 0.10 | 85.08 | 7.47 |
| sat | 1863 | 0.23 | 1059.39 | 334.52 |
| act | 2354 | 0.02 | 25.22 | 1.35 |
| gre | 2394 | 0.00 | 250.00 | 75.00 |
| gpa | 2384 | 0.01 | 2.69 | 0.45 |
| btec\_qualifications | 1901 | 0.21 | 4.60 | 20.36 |
| op | 2378 | 0.01 | 19.52 | 5.34 |
| sqa\_highers | 2292 | 0.05 | 19.19 | 40.67 |
| sqa\_advanced | 2371 | 0.01 | 0.00 | 0.00 |
| as\_levels | 2373 | 0.01 | 35.12 | 44.06 |
| year\_12\_scores | 2402 | 0.00 | 60.00 | 0.00 |
| latitude | 792 | 0.67 | 51.97 | 5.00 |
| longitude | 792 | 0.67 | 0.04 | 9.53 |

### Filtering relevant programs in interdisciplinary subjects

topuniversities.com classifies the programs by subject. While, some subjects are directly related to Data Science and Machine Learning, others, like Biology include traditional programs and others related to Machine Learning (Bioinformatics, etc.).

To classify this programs, I extracted keywords from the program titles and kept titles with some relevant keywords (data, intelligence, computational, etc.). The keywords are extracting using R’s tm package.

## Scraping timeshighereducation.com

The website also renders data dynamically, so we connect to it through a browser. The initial url is <https://www.timeshighereducation.com/world-university-rankings/2023/world-ranking#!/length/-1/sort_by/rank/sort_order/asc/cols/stats>

The data is presented in a single page, so it is easier to scrap. The script that achieves this is the\_scrap\_rankings.R. It uses the same functions from Rselenium and another custom function to extract the data from the webpage.

The scraped data is located in data\_output/the\_ranking\_data.rds, and it includes 267 universities with its ranking and scores on the following categories: overall score, teaching, research, citations, industry income and international outlook.

Data summary

|  |  |
| --- | --- |
| Name | the\_ranking\_data |
| Number of rows | 267 |
| Number of columns | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 9 |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | empty | n\_unique |
| university | 0 | 1.0 | 0 | 163 |
| location | 0 | 1.0 | 49 | 216 |
| overall\_score | 28 | 0.9 | 0 | 34 |
| teaching | 28 | 0.9 | 0 | 86 |
| research | 28 | 0.9 | 0 | 97 |
| citations | 28 | 0.9 | 0 | 101 |
| industry\_income. | 28 | 0.9 | 0 | 101 |
| international\_outlook | 28 | 0.9 | 0 | 59 |
| rank | 28 | 0.9 | 0 | 34 |

**Variable type: numeric**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd |
| longitude | 49 | 0.82 | 7.94 | 19.51 |
| latitude | 49 | 0.82 | 47.66 | 10.47 |
| n | 0 | 1.00 | 39.24 | 50.60 |

## Geocoding

To geocode the downloaded data, we used the [Google Maps API](https://developers.google.com/maps/) and the tidygeocoder [package](https://jessecambon.github.io/tidygeocoder/). The script that does this is geocode\_programs.R. The Google maps API requires an API key.

## Scripts

#### the\_scrap\_rankings.R

Downloads data from [THE](https://www.timeshighereducation.com/) and outputs ./data/ranks.rds and ./data/scores.rds.

#### top\_univeristies\_scrap\_programs\_tuition\_fee.R

Takes ./data/programs.rds as input, downloads tuition and requirements data and outputs ./data/programs\_requirements\_fee.rds.

#### top\_universities\_clean.R

Takes ./data/programs\_requirements\_fee.rds as input, wrangles requirements data, converts tuition fees, filters programs not related to Bioinformatics, etc, recodes university names to match the names from the Times Higher Education website and merges the program data with the /data/geolocated\_locations.rds dataset.

#### top\_universities\_geocode.R

Takes './data/programs\_clean.rds' as input adds coordinates and outputs ./data\_output/programs\_geolocated.rds.

### the\_clean.R

Takes ./data/geolocated\_locations.rds, ranks.rds and scores.rds as input and outputs a single data frame with all the locations with coordinates and the rankings and scores of the corresponding universities.