Towards Data Science, webscrapping for fun*

How to dynamically read and optmize skimming processes

Celio Oliveira

Contents

Web harvesting :: Towards Data Science

This document is a starter of a web harvesting project that aims to optimize the search for articles on the website Towards Data Science. Towards DS is a platform using Medium that exchanges ideas and expands the understanding of data science. It has a mixed audience, consisting of readers entirely new to the subject and expert professionals who want to share their inventions and discoveries.

This document was inspired in (Radecic 2019) and (Oliveira 2020). It was analyzed using R (R Core Team 2020), the "tidyverse" package written by (Wickham et al. 2019), "dplyr" package written by (Wickham et al. 2021), and "rvest" written by (Wickham 2020).

*Code and data are available at: https://github.com/CROliveira/R_WebHarvesting

Introduction

This sample brings the last publications on the landing page of Towards DS website using a rvast package that involves creating an object that we can use to parse the HTML from a webpage. Furthermore, rvest can connect to a webpage and scrape / parse its HTML in a single package. We use syntax similar to dplyr and other tidyverse packages by using %>%.

In further phases, I aim to filter the title of publications, author name, date of publishing, last updated on, how many claps the article received, etc. It is not as easy as I though working with scraped data in R and I need to spend some more time on the documentation and understanding of the parameters to get the data I intend to.

The table below shows the last publications uploaded on the website:

```
##                                                                      titles
## 1                                                  A Better Way To Vote
## 2                                    9 Distance Measures in Data Science
## 3                                 Data Scientists Should Be More End-to-End
## 4                                  A Bayesian Take On Model Regularization
## 5 Answering 10 Most Commonly Asked Questions About Artificial Intelligence
## 6                   Best Python IDEs and Code Editors You Should Know - Part 2
## 7                  Change Data Capture(CDC) tools accelerate data lake adoption
## 8            NLP Profiler: Profiling datasets with one or more text columns
## 9                                                Cloth Filters for Espresso
```

References

Oliveira, Celio. 2020. *VaexWebHarvesting*. https://github.com/CROliveira/WebHarvesting-.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Radecic, Dario. 2019. *A Step-by-Step Guide to Web Scraping with r*. https://towardsdatascience.com/web-scraping-with-r-easier-than-python-c06024f6bf52.

Wickham, Hadley. 2020. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.