Text Analysis and Spatial Data for Economists - First Project

Theresa Gessler

last updated: 2020-02-18

In groups of around 3-5 people, fill out this sheet on current debates about scraping, its scientific use and doing research in R. When you are finished, upload your solutions on GitHub.

Contents

Introduction

GitHub

If you are not registered yet, register for GitHub. I encourage you to set up github on your computer as a version control system but you can also just use your account online and skip the steps for your own computer.

Fork the work sheet and the answer sheet from the usi submit repository. Start filling the answer sheet in a new directory with your name so that you can later commit things without overwriting answers.

Previous experience

Fill in this short survey on your previous experience

Legal regulations and hacktivism

Legality of scraping

Search the web for information whether web scraping is legal.

Describe what the following terms / laws / cases refer to in one sentence each:

- terms of service
- GDPR
- Computer Fraud and Abuse Act
- Sandvig v. Barr

Terms of service

Each of you, choose any social media app or platform that you use frequently (e.g., Snapchat, Instagram, Tik Tok, Facebook). If you do not use social media, you can also look for another platform that you are interested in in your research. Find their Terms of Service (ToS). First, based on their TOS, copy and paste the language about one behavior / activity that is not allowed on that platform. Second, copy and paste parts of the ToS that you find most confusing. If you find something that relates to scraping, also copy that.

robots.txt

Inform yourself about 'robots.txt' files. Check the robots.txt file for a webpage that you are interested in.

Hacktivism

Find information about the following people.

- Aaron Swartz
- Alexandra Elbakyan
- Karrie Karahalios
- Murray Cox
- Paolo Cirio
- another data activist of your choosing

Scientific debate

Using your library article search, identify an academic text / article that relates to company's restrictions on APIs, terms of service etc. Which position does the author take?

Scientific use of scraping

Use of scraping

Search for articles that use scraping on Google Scholar. How many results do you find? Which search terms did you use?

An article that uses webscraped data

Using your library search or google scholar, find an article in your field that uses webscraped data. Tell the other group members about the article and write down how web data is relevant for the article.

Your ideas for scraping

Think about a page that you would be interested in scraping. Paste the address and shortly describe your idea.

Course practicalities

Coding Style

Read a style guide or tutorial on how to write better R code. If you have no idea where to start, try SoftwareCarpentry, R for Reproducible Research, the tidyverse style guide or Code and Data for the Social Sciences: A Practitioner's Guide.

Name one thing that you want to improve about your R code and one thing you disagree with from the style guide.

Coding for others

Show the other group members a piece of code that you have written for a project. It does not have to be the best code you have written, the more that is wrong, the better. Look at the code of the other group members - what do you find counter-intuitive about the way they code?

Write down which decisions were most contested within the group - but of course, don't name names.

RMarkdown

Throughout the course, we will work with files written in RMarkdown. If you have not worked with it before, open RStudio (or your R editor) and create a new RMarkdown file. Read the Introduction to RMarkdown on the RStudio page until (and including) the section on inline code. Identify the elements (text, code chunks, inline code) in your own RMarkdown file. If you do not have RStudio installed, you can use the RSTudio Cloud

Work environment

Come up with a folder structure that you think would make sense for your work in this class. Set it up on your computer!

Submit

Commit your answers and send a pull request to the usi submit repository