

Screen-scraping Facebook: technical and ethical aspects

Moreno Mancosu

Lucerne, October 28, 2020
Lucerne RUG Workshop

Motivation

- ▶ Human life is happening more and more on the internet
- ▶ Many events occurring on the internet are accessible to everyone, so in principle they can be observed and studied
- ▶ Researchers are more and more interested in people's behavior on **Social Network Sites**
- ▶ **Facebook** is arguably the most widely used SNS in the world

The problem

- ▶ *Until early 2018*: FB public information (e.g. posts, reactions, comments) could be collected legally using their Application Programming Interface (API)
- ▶ *February 2018–August 2019*: Facebook tightens the API access to third parties – Software like **RFacebook** or **Netvizz** is now unusable
- ▶ *May 2018*: GDPR – Researchers' freedom to get and publish user information is further restricted
- ▶ Yet, the same information is publicly available and can be accessed by everyone from FB's website

Screen scraping: a technique to automatically extract data from web pages

Premises

- ▶ FB information is still public – and, we bet, will remain public
- ▶ If you can **see** the information, you already **have it** on your computer
- ▶ So you can download it and organize it into a dataset
- ▶ You can instruct a web browser (e.g. Chrome) to do so automatically

The logic of screen scraping

By means of automatic browsing, researchers can instruct a browser to:

- ▶ **Head** to a certain (public) page
- ▶ **Show** all the information that they might need (e.g. posts + number of reactions, comments)
- ▶ **Extract** the information by parsing the page
- ▶ **Organize** the information into a structured dataset

An example

Is it **ethical**?

- ▶ Yes, because you do not scrape any type of personal information, and when you do you pseudonymize information


Is it **legal**?

- ▶ Yes, because you do not scrape any type of personal information, and when you do you pseudonymize information

Does it comply with FB **Terms of service**?



What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data

Social Media + Society
July-September 2020: 1–11
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2056305120940703
journals.sagepub.com/home/sms


Moreno Mancosu¹  and Federico Vegetti^{1,2}

Abstract

In reaction to the Cambridge Analytica scandal, Facebook has restricted the access to its Application Programming Interface (API). This new policy has damaged the possibility for independent researchers to study relevant topics in political and social behavior. Yet, much of the public information that the researchers may be interested in is still available on Facebook, and can be still systematically collected through web scraping techniques. The goal of this article is twofold. First, we discuss some ethical and legal issues that researchers should consider as they plan their collection and possible publication of Facebook data. In particular, we discuss what kind of information can be ethically gathered about the users (public information), how published data should look like to comply with privacy regulations (like the GDPR), and what consequences violating Facebook's terms of service may entail for the researcher. Second, we present a scraping routine for public Facebook posts, and discuss some technical adjustments that can be performed for the data to be ethically and legally acceptable. The code employs screen scraping to collect the list of reactions to a Facebook public post, and performs a one-way cryptographic hash function on the users' identifiers to pseudonymize their personal information, while still keeping them traceable within the data. This article contributes to the debate around freedom of internet research and the ethical concerns that might arise by scraping data from the social web.

Keywords

web scraping, social networks, research ethics, Facebook

Introducing fbSim

There's also an R package to simulate a Facebook user's behavior and automatically collect information from public Facebook pages via Google Chrome automatic sessions

3 functions:

1. **fbSetAccount**: produces a new Chrome profile with your FB credentials stored in it
2. **fbSimPosts**: navigates and gets information about posts from a target public FB page
3. **fbSimLikes**: gets information about the pages likes by a target FB page

You can find the package here

<https://github.com/morenomancosu/fbSim>

We are looking for beta testers! Please contact us if you would like to be one

Thank you for your
attention!

`moreno.mancosu@unito.it`