OLD DOMINION UNIVERSITY

Anyone Can Learn Web Scraping

Richard N. Landers, Ph.D.
Old Dominion University
@rnlanders | rlanders@tntlab.org

SUNY-Binghamton 2018, Binghamton, NY

Agenda/Learning Objectives

1. Foundational Questions
   - Why scrape social media?
   - What are the pros and cons of social media data sources?
2. Technical Overview
   - What steps are involved in scraping social media?
   - How are Facebook and Twitter accessed?
3. Demonstrations
   - Twitter
   - Facebook
4. Practical Concerns
   - How to learn this skillset
   - Ethical concerns and legal risks

Primary Reference for this Workshop

- **Landers, R. N.**, Brusso, R. C., Cavanaugh, K. J. & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods, 21,* 475-492.
  - Steps you through the creation of data source theories and an example in much greater detail than what I'll talk about here
  - Illustrates some technical concepts in greater detail
- Additional resources I've provided
  - A quick summary of scraping and APIs at https://tinyurl.com/ydhy5j93
  - All workshop materials at http://scraping.tntlab.org
  - My tutorial on Python's *scrapy is* available at http://rlanders.net/scrapy

3

Foundational Questions

Why scrape social media?
What are the pros and cons of social media data sources?

Why scrape social media?

- What is social media?
  - A consequence of the Web 2.0 movement toward interactivity on the internet
    - "user generated content"
- What does user-generated content entail?
  - purposive data
    - user profiles
    - content
  - incidental metadata (see Ghostery on http://abcnews.com)
    - trail of breadcrumbs

Examples of social media data

- Facebook
  - **Data:** profile content, job history, education history, places of residences, pictures, picture captions, family relationships, feed posts, tags, photos, group memberships, likes, comments
  - **Metadata:** photo meta-data (e.g., locations), posting locations, post times, like meta-data (down the rabbit hole)
- Twitter
  - **Data:** posts, photos, tags, retweets
  - **Metadata:** posting locations, retweet and tag networks
- Any discussion board (or other social websites)
  - **Data:** posts, profiles (whatever they might contain)
  - **Metadata:** varies widely

So what can I do with scraped data?

- The first step of "big data science," data wrangling/munging

https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

- Can be followed up with any sort of descriptive or inferential analysis you want

So what can I do with scraped data?

- Text data is commonly subjected to follow-up data complexity reduction techniques
  - Linguistic Inquiry and Word Count (LIWC)
    - Outputs an enormous variety of summary statistics about text, including linguistic (types of words), psychological (traits), high-level (e.g., authenticity, emotional tone)
    - See Tausczik & Pennebaker (2010)
  - Sentiment
    - Uses existing lexica to classify words as positive or negative (such as LIWC)
    - The Harvard General Inquirer (from Stone, Dunphry, Smith & Ogilvie, 1966)
  - Topic Analysis
    - Latent Dirichlet allocation (LDA) - Kosinski, Wang, Lakkaraju, & Leskovec (2016)
- Or don't reduce, if you don't want to.

Data Source Theories (and example RQs)

- Develop a list of your assumptions about the data sources you are considering related to:
  - **Data origin/population characteristics**
    - Why does this website exist?
    - Who owns the data available on this website?
    - Why would someone want to visit this website?
    - Why would a content creator want to contribute?
    - What type of data do content creators provide?
    - Do users pay to participate?
    - Are creators restricted in the kind of content they can contribute?
- Data source theories are the core concept in **theory-driven web scraping**
  - **Data structure**
    - How are target constructs represented both visually and in code?
    - Is there inconsistency in how target constructs are represented?
    - Do data appear on only one type of webpage?
    - How is user content created and captured?
    - How much content available on each page?
    - Is the content consistently available?

Common Assumptions About Social Media

- A huge variety of Facebook data and metadata are available about basically everyone in the United States.
  - **PARTLY TRUE:** Only if their privacy settings allow it.
- Unlimited information about everyone that has ever posted on Twitter is available.
  - **PARTLY TRUE:** Most people get access to Twitter data via the 'firehose.'
- I can get full job histories about anyone on LinkedIn.
- I can get full job histories about anyone whose privacy settings allow it.
  - **FALSE UNLESS YOU'RE A CRIMINAL:** This is almost certainly illegal.
- **We'll come back to this in the last section:** A lot of web scrapers are criminals.

10

More Specific Data Source Theories

- **Facebook**
  - The data you can scrape vary based upon who you are and what access you have obtained for yourself.
  - In practice, there are two ways to do this:
    - Scrape content from public groups/pages
    - Create an app that people sign up for and scrape profile content
  - There are **time limitations**.
- **Twitter**
  - Almost all profiles are public, so that's much easier.
  - Birthdays may be available.
  - Geographic data is available, sort of.
  - Search tools don't allow unrestricted access; there are per-query access limits.
- **Other websites**
  - Varies by website features and terms of service.

11

Technical Overview

What steps are involved in scraping social media?

12

Five Steps to Execute a Web Scraping Project

1. Identify and pre-emptively evaluate potential sources of information
   - Assumes you already have a purpose in mind
   - Don't necessarily limit yourself to Twitter and Facebook – any webpage can potentially be used
   - Create a data source theory
     - Think counterfactually: "If X isn't true, my conclusions from this data source will be invalid."
     - Write it down.
     - Develop specific hypotheses that your theory suggests and figure out which ones you can test (assumptions vs. hypotheses).

Five Steps to Execute a Web Scraping Project

2. Develop a coding system
   a) Identify the specific pieces of information you want to grab from each website
      - Remember to include info to test your data source theory
   b) Determine where each piece of information appears on each webpage
   c) Determine how cases are replicated in terms of the webpages
      - Is there one case on each webpage?
      - If multiple cases are represented on each webpage, how are they represented?

Steps to Execute a Web Scraping Project

3. Code a scraper and potentially a crawler
   - When scraping, data will come from one of two sources depending upon which website's data you're trying to access
   - If an API is available, you want to use the API
     - Returns **structured** data with variables pre-defined
     - Legally unambiguous
   - If an API is not available, you'll need to scrape manually
     - Returns **unstructured** data
     - Requires a lot more work
     - Legally ambiguous in some cases

So what's an API?

- **API: Application Programming Interface**
  - A data gateway into someone else's system
  - Created by the provider of the service
  - Almost universally intended and designed for real-time access by other websites, but you can use them too
  - Requires learning API documentation – they're all different
- Let's start easy. I've created an API at http://scraping.tntlab.org/add.php
- It adds two numbers, x and y.
- Try:
  - http://scraping.tntlab.org/add.php
  - http://scraping.tntlab.org/add.php?x=1
  - http://scraping.tntlab.org/add.php?x=1&y=muffin
  - http://scraping.tntlab.org/add.php?x=1&y=8

16

What format of data do APIs provide?

- The output of an API can be in essentially *any* format, but some are more common.
  - If you're lucky
    - CSV: comma-separated values file
    - DAT: tab-delimited data file
  - More than likely
    - JSON: JavaScript object notation
- Both Facebook and Twitter return JSON files
- These APIs also have **rate limits** in terms of the number of requests you are allowed to send and how quickly; Twitter for example limits to 180 calls every 15 minutes for simple requests and 15 calls every 15 minutes for complex one.
  - For example, only 25 tweets can be returned per simple call, so up to 4500 tweets per 15 minutes

17

JSON Output from Facebook API

18

Getting What You Want

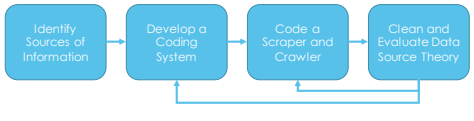
- Learn the documentation to understand what you can and can't actually scrape
  - Twitter: https://dev.twitter.com/docs
  - Facebook: https://developers.facebook.com/docs/
- The next challenge is to convert the JSON file into a format you want. You can do this in any program you want, but I find R is easiest
  - R package: twitteR
  - R package: Rfacebook # might not be useful in a post-CA world

If There's No API, You Need to Scrape

```
<html>
<head>
<title>My Webpage's Title</title>
<link href="mycss.css" type="text/css" rel="stylesheet">
</head>
<body>
<h1>My Wepage</h1>
<p>This is my webpage.</p>
<p>Here's data I want.</p>
</body>
</html>
```


Five Steps to Execute a Web Scraping Project

4. Clean the data and revise the data source theory
   - Once you have your data in hand, run all hypothesis tests possible from your data source theory
   - You will almost certainly identify problems with your coding system at this stage; time to revise



Five Steps to Execute a Web Scraping Project

5. Analyze!
   - Natural language processing
   - Data simplification
   - Simple profile reporting

Demonstrations

Twitter API
Scraping Facebook

Practical Concerns

How to learn this skillset
Ethical concerns and legal risks

Why Do This Yourself?

- The old way
  - Hand-coding text (~2 minutes per subject; with 2 coders, at 60 per hour, coding 500 entries would take 8.3 hours of coding time)
- The new way
  - In ~8 hours, we captured >100,000 text entries

- If you don't want to code, you can't use APIs
- If you already know R, you'll find API calls fairly easy
  - Does require learning a bit about how the internet works

How to Learn This Skillset

- There are two major skillsets involved:
  - HTML, to know how web pages are structured
  - Statistical programming (e.g., in R or Python) in general, to be able to run algorithms
    - Web scraping libraries in R or Python, to run specific extraction algorithms

- To learn HTML, https://www.codecademy.com/learn/learn-html-css

- To learn R, Python, and their libraries:
  https://www.datacamp.com/tracks/data-scientist-with-r
  https://www.datacamp.com/tracks/data-scientist-with-python
  http://datascience.tntlab.org

26

Ethics and Legal Risks - Hacking

- Don't look like a hacker and you won't be treated like one (honeypots)

- Remember to read API documentation (and to authenticate)
- Look for tutorials/examples of those that have done this before
- Don't go hunting for statistical significance with the standard psych toolkit

27

Ethics and Legal Risks – Fair and Commercial Use

- **Fair use:** Often unclear what is usable
  - Harvesting data when a policy is in place explicitly forbidding it is definitely unethical and probably illegal (see eBay v Bidder's Edge, 2000 and Ticketmaster Corp vs Tickets.com, 2000)
  - Harvesting data behind a login wall without a policy is probably unethical and probably illegal (APIs protect you from this)
  - Harvesting public data that is not explicitly linked anywhere is probably unethical and probably illegal (see the story of Andrew Auernheimer, aka *weev*)
  - Harvesting public social media data that is plainly visible through simple web browsing might be ethical but is **probably legal**
  - Recent case of LinkedIn v. Hi-Q

28

OLD DOMINION UNIVERSITY

Thank you! Questions?

Richard N. Landers, Ph.D.
Old Dominion University
@rnlanders | rlanders@tntlab.org

SUNY-Binghamton 2018, Binghamton, NY