

OLD DOMINION UNIVERSITY

# How to Create a Dataset from Twitter or Facebook: Theory and Demonstration

Richard N. Landers  
*Old Dominion University*  
 @rnlanders | rnlanders@odu.edu  
 APA 2017, Washington, DC

---

---

---

---

---

---

---

---

## 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
  - Facebook
  - Twitter
4. Practical Concerns
  - How to learn this skillset
  - Ethical concerns and legal risks

---

---

---

---

---

---

---

---

## Primary References for this Session

- 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
  - Closely tied to my tutorial on Python's *scrapy*
    - <http://rlanders.net/scrapy>
- Website with Slides and Tutorials
  - <http://scraping.tnlab.org>

3

---

---

---

---

---

---

---

---

Foundational Questions

Why scrape social media?  
What are the pros and cons of social media data sources?  
What is machine learning and how do I use it?

---

---

---

---

---

---

---

---

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
- So psychologically, what are social media data?
  - behaviors, the products of person-situation interactions

---

---

---

---

---

---

---

---

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

---

---

---

---

---

---

---

---

## 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 machine learning you want (e.g., OLS regression and Pearson's correlation, naive Bayes classifiers)

7

---

---

---

---

---

---

---

---

---

---

## 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, Dunphy, Smith & Ogilvie, 1966)
  - Topic Analysis
    - Latent Dirichlet allocation (LDA) - Kosinski, Wang, Lakkaraju, & Leskovec (2016)
- Or don't reduce, if you have enough data and don't want to.

8

---

---

---

---

---

---

---

---

---

---

## 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 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?
- Data source theories are the core concept in **theory-driven web scraping**

9

---

---

---

---

---

---

---

---

---

---

## Data Source Theories Imply Testable Predictions

- Make predictions based upon what you think must be true to create a complete data source theory with testable predictions (i.e., hypotheses).

- Example

- Q: How is political engagement represented in tweets?
- H: Twitter posts containing the names of politicians represent political engagement.



- In traditional data collection, we have these same assumptions but they are generally difficult or impossible to test.
  - Content validation is relatively easy.

10

---

---

---

---

---

---

---

---

## 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.'

11

---

---

---

---

---

---

---

---

## 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.

12

---

---

---

---

---

---

---

---



Technical Overview

What steps are involved in scraping social media?

1

---

---

---

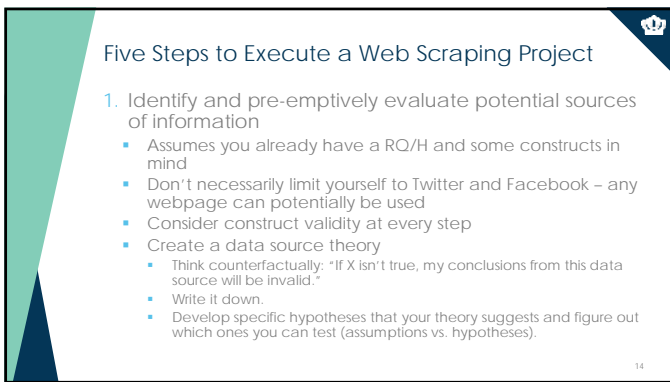
---

---

---

---

---



Five Steps to Execute a Web Scraping Project

1. Identify and pre-emptively evaluate potential sources of information

- Assumes you already have a RQ/H and some constructs in mind
- Don't necessarily limit yourself to Twitter and Facebook – any webpage can potentially be used
- Consider construct validity at every step
- 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).

14

---

---

---

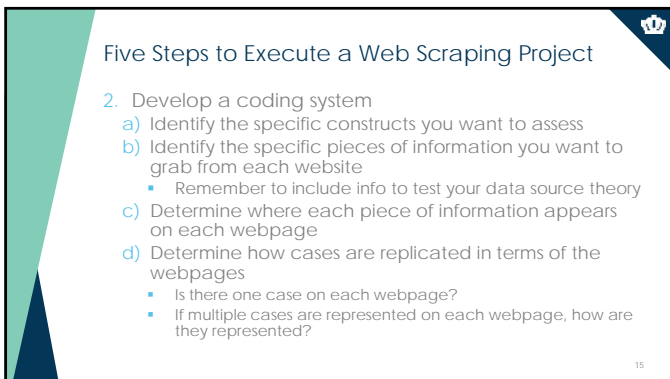
---

---

---

---

---



Five Steps to Execute a Web Scraping Project

2. Develop a coding system

- Identify the specific constructs you want to assess
- Identify the specific pieces of information you want to grab from each website
  - Remember to include info to test your data source theory
- Determine where each piece of information appears on each webpage
- 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?

15

---

---

---

---

---

---

---

---

## 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

16

---

---

---

---

---

---

---

---

## 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.tn1lab.org/add.php>
- It adds two numbers, x and y.
- Try:
  - <http://scraping.tn1lab.org/add.php>
  - <http://scraping.tn1lab.org/add.php?x=1>
  - <http://scraping.tn1lab.org/add.php?x=1&y=muffin>
  - <http://scraping.tn1lab.org/add.php?x=1&y=8>

17

---

---

---

---

---

---

---

---

## 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 ones.
  - For example, only 25 tweets can be returned per simple call, so up to 4500 tweets per 15 minutes

18

---

---

---

---

---

---

---

---

## Experiment with the Facebook API

- Go to <http://developers.facebook.com/tools/explorer> (you'll need to be logged into Facebook)
- Generate a token for yourself ("Get Token")
  - This token will have the permissions that your Facebook account has
- Craft a request using the Explorer, such as:
  - 853552931365745/feed
- Create this same request in your web browser by going to:
  - [https://graph.facebook.com/853552931365745/feed?access\\_token=xxxx](https://graph.facebook.com/853552931365745/feed?access_token=xxxx) (but replace xxxx with the copy/pasted token you generated)

19

---

---

---

---

---

---

---

---

## Facebook Graph Explorer

```
{
  "message": "Hi everyone

Some of you might have seen that I've written a survey about why people leave academia. But it needs respondents, and that's why I'm asking for help: can you please share this survey with people you know who have PhDs/psychology? If you could check it out on social media as well I'd be very grateful.

Survey: https://www.surveymonkey.co.uk/j/2939262

Many thanks

Sorry if you've already seen this survey!"
  "story": "Michael Haselbach shared a link to the group: Psychological Methods Discussion Group.",
  "updated_time": "2017-08-02T21:17:04+0000",
  "id": "853552931365745_14649337931761"
}
```

20

---

---

---

---

---

---

---

---

## JSON Output from Facebook API

```
{
  "data": [
    {
      "message": "Hi, could anyone suggest a comprehensive review of advantages and disadvantages of using 2-point vs. 3-point vs. 5-point scales for measuring attitudes/preferences/opinions, etc.? Thanks!",
      "updated_time": "2017-08-02T21:00:10+0000",
      "id": "853552931365745_14649337931761"
    },
    {
      "message": "Hi everyone! I'd love if you might have seen that I've written a survey about why people leave academia. But it needs respondents, and that's why I'm asking you for help: can you please share this survey with people you know who have PhDs/psychology? If you could check it out on social media as well I'd be very grateful.

Survey: https://www.surveymonkey.co.uk/j/2939262

Many thanks! Sorry if you've already seen this survey!"
      "story": "Michael Haselbach shared a link to the group: Psychological Methods Discussion Group.",
      "updated_time": "2017-08-02T21:17:04+0000",
      "id": "853552931365745_14649337931761"
    },
    {
      "message": "Does anyone have any thoughts on how Nuiser for Nuiser works relative to Nuiser for Nuiser? Is its functionality the same in your experience?",
      "updated_time": "2017-08-02T21:10:10+0000",
      "id": "853552931365745_14649337931761"
    },
    {
      "message": "I apologise if this is old news, and it probably won't be useful to those of you already proficient in R, but I just came across this shiny app for making bar charts, and been quite into it."
      "story": "Michael has shared a link to the group: Psychological Methods Discussion Group.",
      "updated_time": "2017-08-02T21:04:01+0000",
      "id": "853552931365745_14649337931761"
    },
    {
      "message": "There are many ways to estimate publication bias that have been proposed in the last years (e.g., trim, B-curve, trim-curve,...). Are there any R packages (or some other tools) that can calculate all (or many) of these estimates? We're having a summer school on open science and meta analysis and want the participants to be able to try many"
    }
  ]
}
```

21

---

---

---

---

---

---

---

---

## 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: twitterR
  - R package: Rfacebook

22

---

---

---

---

---

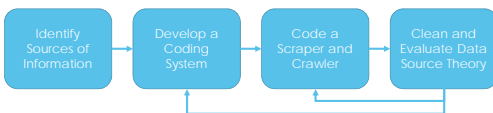
---

---

---

## Five Steps to Execute a Web Scraping Project

- 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



23

---

---

---

---

---

---

---

---

## Regular Expressions

- Regular expressions are enormously powerful and can be very confusing, even if you know what you're doing
  - Can be used to identify or replace text
- Examples of simple regex replacement with "x": I have 9 dogs.
  - `\d` Match any digit I have x dogs.
  - `[ade]` Match letters a, d, or e I hxxx 9 xogs.
  - `\w` Match any alphanumeric x xxxx x xxxx.
  - `\W` Match any non-alphanumeric lxhavex9xdogsx
  - `\s` Match any whitespace lxhavex9xdogs.
- Can get really, really complicated
  - `^\(\{0-9\}(3\)\ | [0-9]{3})[0-9]{3}-[0-9]{4}$`
- Learn with <https://regexone.com/>, test with <http://regex101.com>

24

---

---

---

---

---

---

---

---



Five Steps to Execute a Web Scraping Project

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

25

---

---

---


---

---

---

---

---



Pre-Demonstration Break

---

---

---

---

---

---

---

---

Demonstrations

Facebook and Twitter

---

---

---

---

---

---

---

---

# Practical Concerns

How to learn this skillset  
Ethical concerns and legal risks

---

---

---

---

---

---

---

---

# Why Do This Yourself?

- The old way
  - URAs 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
- You should really learn R anyway
  - <http://www.slop.org/tip/july16/crash.aspx>

---

---

---

---

---

---

---

---

# How to Learn This Skillset

- There are three major skillsets involved:
  - HTTP, to know how URLs are created and how they are used
  - 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
    - Machine learning libraries in R, Python, SPSS, etc. to run analytic algorithms
- To learn HTTP, <http://www.slop.org/tip/july17/crash.aspx>
- 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>

---

---

---

---

---

---

---

---

## 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

31

---

---

---

---

---

---

---

---

## 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**
  - A case related to LinkedIn is *currently in the court system*

32

---

---

---


---

---

---

---

---



OLD DOMINION UNIVERSITY

## Questions?

For easily digestible descriptions of new technologies, see my column in the *Industrial-Organizational Psychologist!*

For example, natural language processing:  
<http://www.slop.org/tip/april17/crash.aspx>

Richard N. Landers  
 Old Dominion University  
 @rnlayers | rnlayers@odu.edu  
 APA 2017, Washington, DC

---

---

---

---

---

---

---

---