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

ODU STOB Dean’s Research Seminar
September 2017

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

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?

- Why do social media exist?
  - 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

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 have enough data and 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?

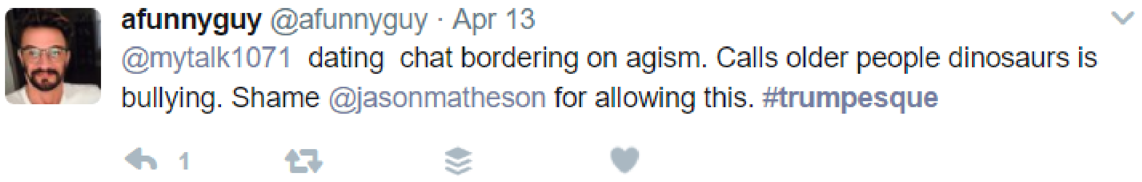
- Data source theories are the core concept in **theory-driven web scraping**
- **Data structure**
  - How are target constructs represented both visually and in code?
  - Is there inconsistency in how target constructs are represented?
  - Do data appear on only one type of webpage?
  - How is user content created and captured?
  - How much content available on each page?
  - Is the content consistently available?

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.

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
  - RQ: 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.

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-ISH:** This is probably illegal, but this may change soon.
- **We’ll come back to this in the last section:** A lot of web scrapers are criminals.

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.

Technical Overview

What steps are involved in scraping social media?

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).

Five Steps to Execute a Web Scraping Project

2. Develop a coding system
   a) Identify the specific constructs you want to assess
   b) Determine how those constructs are represented from a technical standpoint
      a) Are they recoded from text?
      b) Are they structured pieces of information?
      c) Where are they? 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
        - Will probably need multiple calls to grab large datasets
        - 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

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
    - TSV: 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

JSON Output from Facebook API

Secure | https://graph.facebook.com/853552931365745/feed?access_token=EAACEdEose0cBANJClsJ9dadoE


```
{
   "data": [
      {
         "message": "#New_Significance #P_Less_Than_005\n#Type_I_Error\n#Type_II_Error\n#Error_Balance \nI dic
average effect size in social psychology) and computed sample sizes for different type-I and type-II error prc
alpha = .05,  beta = .75  Ratio 1/15\nN = 100, alpha = .05, beta = .50, Ratio 1/10\nN = 200, alpha = .05, beta
338, alpha = .005, beta = .20, Ratio 1/40\nN = 500, alpha = .005, beta = .05, Ratio 1/10\nN = 600, alpha = .00
power, which implies 20\u0025 Type-II errors, we fail to provide evidence for a true hypothesis with effect si
far, social psychologists have been using sample sizes of n = 20 per cell (N = 40 total) to chase these effect
75\u0025 and a type-I / type-II error ratio of 1/15.  \nIf social psychologists would do a priori power analys
times as many participants). \nUsing the same N = 200 and the new significance criterion of p \u003C .005,  pc
suggesting that type-II errors are much less important than type-I errors.  \nTo get back to a 1/4 ratio, samp
applies to d = .4, which is an average effect size, meaning power is lower for half of the studies.  \nAre we
         "story": "Uli Schimmack created a poll in Psychological Methods Discussion Group.",
         "updated_time": "2017-07-27T19:56:44+0000",
         "id": "853552931365745_1457448990976133"
      },
      {
         "message": "More comments on #new_significance \n\nIs it better to have no significance (threshold)?
         "story": "Uli Schimmack shared a link to the group: Psychological Methods Discussion Group.",
         "updated_time": "2017-07-27T19:41:59+0000",
         "id": "853552931365745_1458680730852959"
      },
      {
         "message": "Hi everybody,\n\nI\u2019m considering using p-curve and/or p uniform as supplementary pub
dependencies in the data, so to investigate other publication bias indices (trim-and-fill, PET-PEESE, selectic
package). \n\nDoes p-curve and p uniform in meta-analysis also assume \nthat all effect size estimates are inc
sizes, b) impute a p-value from the aggregated dependent effect size, and c) perform p-curve and/or p uniform
I\u2019ve read several papers on these methods, but so far have not seen any discussion on this issue.",
         "updated_time": "2017-07-27T19:14:04+0000",
         "id": "853552931365745_1458154720905560"
```

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=xxxxx (but replace xxxxx with the copy/pasted token you generated)

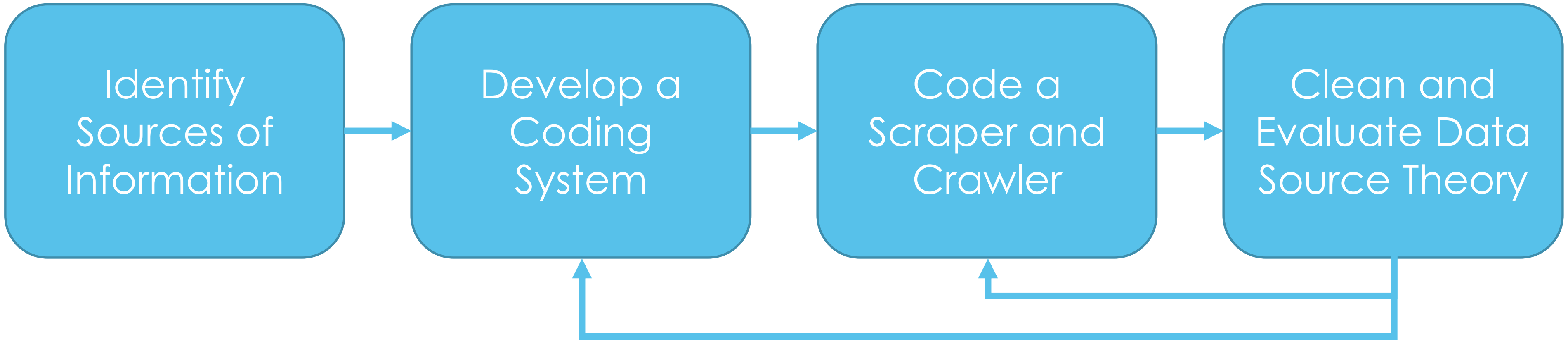
Getting What You Want

- 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

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

Demonstration

Facebook

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

How to Learn This Skillset

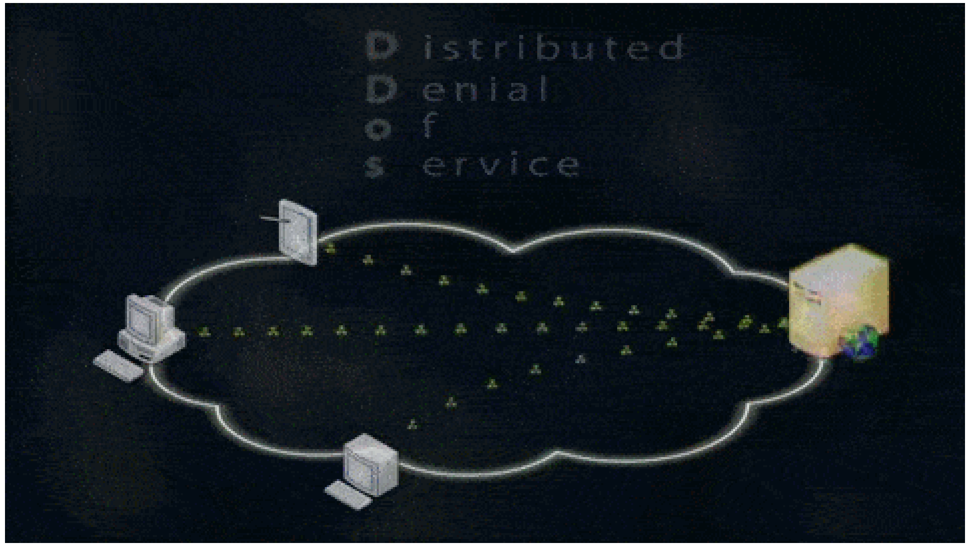
- There are two major skillsets involved:
  - HTML, to know how web pages are structured
  - Statistical programming (e.g., in R or Python) in general, to be able to run algorithms
    - Web scraping libraries in R or Python, to run specific extraction algorithms
    - Machine learning libraries in R,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

- 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

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*

OLD DOMINION UNIVERSITY

http://scraping.tntlab.org

Questions?

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

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

Richard N. Landers
Old Dominion University

@rnlanders | rnlanders@odu.edu

ODU STOB Dean's Research Seminar
September 2017