Getting & Using Data From the Internet

Alyssa Smith (smith.alyss@northeastern.edu)

INTROS

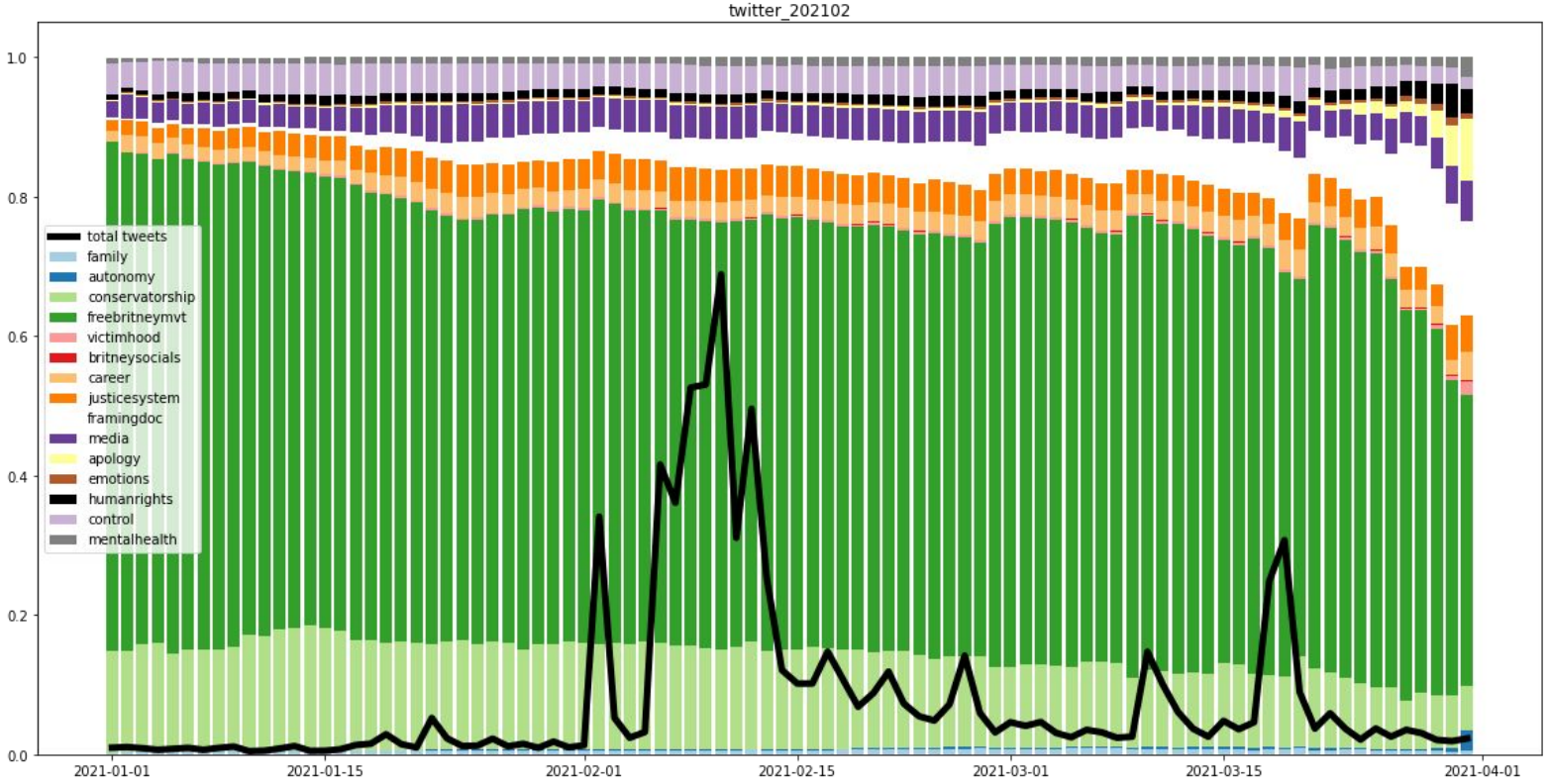
Introductions

- Name / pronouns
- Favorite real/fictional animal
- What are you thinking of studying for your capstone project?
- What kind of internet data would you like to look at?

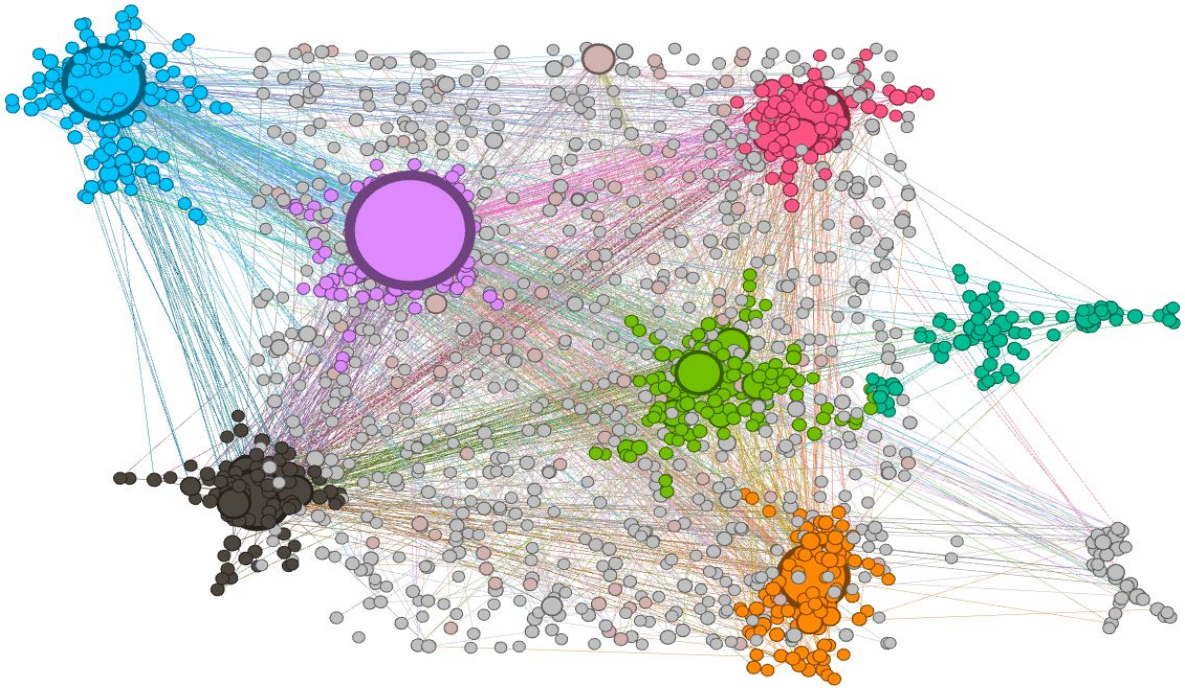
Why Internet Data?

- Internet data can give us a way to (very imperfectly) quantify people's social lives online.
  - What are people talking about?
  - Who do people interact with?
  - How do communities form?
- It is especially useful at large scales - getting this kind of information on how people associate without social media data would be very difficult, if not impossible!
- Internet data is very rich in terms of context, content, and usability

Some Cool Projects I've Done With Internet Data



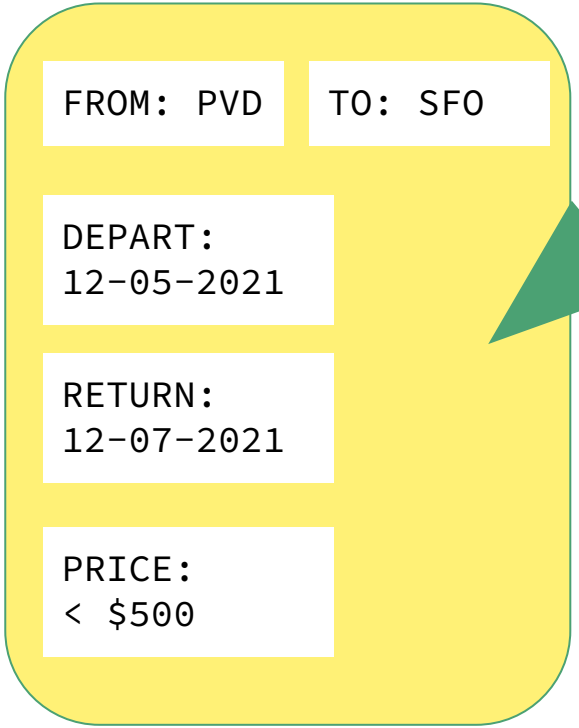
Some Cool Projects I've Done With Internet Data



Some Cool Projects I've Done With Internet Data

How do websites & apps work?

How does it work?



Definitely an app window in a real phone

The user interacts with the user interface. The app knows what she's trying to get, but it doesn't have the information she wants. So it calls up a friend.

Calling the API

Dearest server, can I have all the flights from PVD to SFO departing on 12/05/2021 and ….

`This is still definitely the app`

Sure thing! We have these flights:
1) United, leaving at 5:00AM….
2) JetBlue, leaving at 7:30 AM….

Calling the API

But how do the app and the server know how to talk to each other?

Dearest server, can I have all the flights from PVD to SFO departing on 12/05/2021 and ….

This is still definitely the app

Sure thing! We have these flights:
1) United, leaving at 5:00AM….
2) JetBlue, leaving at 7:30 AM….

What an API does

Speaking very generally, an API is how two or more computer programs interact with each other.

It’s a set of rules that govern how one program asks the other for things, what things it can get, and how the things will look when they’re returned.

Calling the API

Here's what that looks like in practice:

This is still definitely the app

www.flyingisfun.com/?origin=PVD&dest=SFO&depart=20211205&return=20211207&price_lt=500

{Flights: {'airline': 'United', 'depart_time': '20211205T050000', …}, {'airline': 'JetBlue', …}}

What’s going on?

The app knows how to phrase its request so that the server understands it, and the server knows how to construct its response so that the app can make sense of it.

An API is a way for computer programs to talk to each other.

Why do I care?

APIs are great for researchers and developers!

If you are trying to get a lot of information repeatedly from somebody else's computer program, an API is the way to do it!

This might look like:

- An app that schedules your flights for you
- An analysis of all reddit posts mentioning "potato farming"
- A program that emails you every time your advisor tweets something with negative sentiment

How APIs Work

- Most APIs are wrappers on a database or a bunch of computer functions.
- That means that when we talk to the API, we have to give it information in a way that will make sense to the database/computer program that is doing all the work.
- We'll see some examples of this in a second.

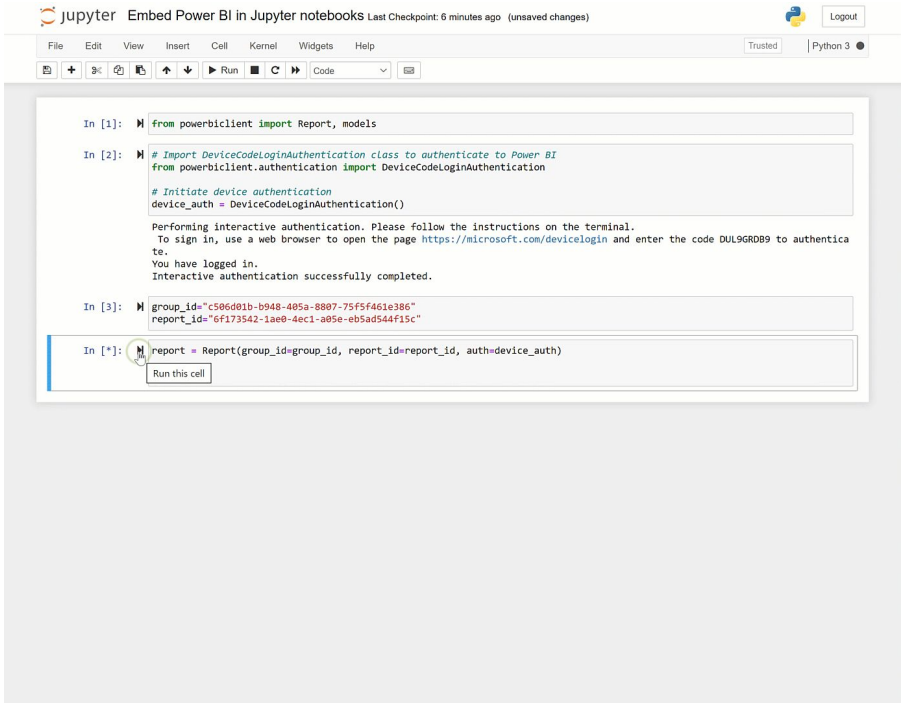
How the heck do I do this?

Programming & Computers

- Programming is a way to talk to a computer and get it to do what you want.
- There are a lot of **programming languages** out there that have different pros & cons; today we'll be using **Python**.
- If you have used formulas in Excel, congratulations. You have done programming. You are totally equipped to do this.
- Computer programs are a great way to deal with a lot of data and learn from it!

Jupyter notebooks

- In the interactive portion of the class, we'll be using a Jupyter notebook. This "notebook" allow you to run small chunks of code at a time.
- You can modify these chunks of code and run them multiple times in order to try different inputs or techniques.
-



Software Packages that talk to APIs

There are a few ways to use APIs.

Some APIs have pre-built packages that let you access them with less friction and parsing involved. Praw (Python Reddit API Wrapper) is one that I've used a bit. There will be an example of this in the code!!

Reddit has a Python package - this is easy to use!

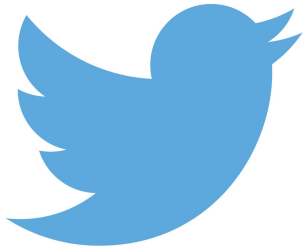
```
29 APP_NAME = creds['app_name']
30
31 MY_SUBREDDIT = 'wallstreetbets'
32 SEARCH_TERM = 'gamestop&(potato|facebook)'
33 reddit = praw.Reddit(client_id=REDDIT_ID, client_secret=REDDIT_SECRET, user_agent=APP_NAME)
34 subreddit = reddit.subreddit(MY_SUBREDDIT)
35
```

Using the API itself

Not all APIs have nice Python or R packages, though.

Sometimes you have to use the API itself; generally we use the requests module in Python to do so, along with json to parse the responses. I've put an example of this in the code as well!

Twitter has an API that is well-documented and easy to use.



Scraping

Sometimes websites don’t even have an API; in that case, you’ll have to scrape the website.

When you scrape a website, you pull the whole webpage, parse it, and extract the data you want.

This works better on structured websites that don’t block bots (if you are scraping a website, you are a bot). If you run into CAPTCHAs, I am sorry. Don’t try to beat them.

Archive of Our Own (the fanfiction website) is an example of a website that is easy to scrape.

Giving up

Some websites or apps have things built in that make it hard for researchers to scrape them. Others don’t have an API you can use and are hard to scrape by nature.

Facebook, Instagram, and TikTok are examples of these.

Unfortunately, if you are hoping to use data from one of these websites, you are very unlikely to be successful.

Ethical Considerations

Contextual Privacy

- If you're dealing with social media data that is scrapable or visible using a public API, chances are that it is publicly visible.
- When we think about privacy online, though, we want to think of it as *contextual*. What someone might be comfortable saying in one context, with a known set of expectations, might not be something they're okay saying to a researcher or that researcher's audience.
- Search, archiving, and tracking online are very powerful now. This potentially allows people to deduce a lot of information from a single social media post, especially if that post is shared to an audience that is vastly different from its intended audience.

Keeping People Safe

- Generally speaking, it is risky to publicize the username, profile picture, or exact text of a social media post or profile.
- If you have more information on the participants, such as survey data linked to their social media profile, it’s even more important to keep that information safe.
- You can anonymize usernames by redacting or hashing them; profile pictures can also be redacted or hashed.
- To show example posts etc, you can make up your own or heavily redact them.

Consider Vulnerability

- Consider the populations you're studying. Some people may be at risk if their social media profiles were somehow made available to others in their lives.
- Some examples:
  - LGBTQ+ people who aren't out (or are only out in some parts of their lives) can be outed through social media profiles that are surfaced to the wrong audience.
  - Marginalized populations talking about controversial topics often use techniques to avoid algorithmic detection and search (e.g. deliberate but human-readable misspellings of names) - making the specifics of these techniques explicit and public can open up these communities to online harassment they are trying to avoid.
- Think about the power imbalance between you (the researcher) and the people you are studying.
- People have the right to **not be counted** in your research.

Questions so far?

Coding Time!!