Social Media Mining: An Introduction using R & Twitter

Social Science Data Analytics Workshop

Caleb Lucas (@calebjlucas)

September 25, 2019

Michigan State University

Agenda

1. Web APIs
2. Twitter API
3. `rtweet`
4. R sesh!

Introduction

- A bit about me (and Shane!)

Introduction

- A bit about me (and Shane!)
- Code of Conduct: Respect & Inclusivity

Introduction

- A bit about me (and Shane!)
- Code of Conduct: Respect & Inclusivity
- Learning goals:

Introduction

- A bit about me (and Shane!)
- Code of Conduct: Respect & Inclusivity
- Learning goals:
  - Understand how social media APIs work

Introduction

- A bit about me (and Shane!)
- Code of Conduct: Respect & Inclusivity
- Learning goals:
  - Understand how social media APIs work
  - Collect, visualize, and store tweets using `rtweet`

Introduction

- A bit about me (and Shane!)
- Code of Conduct: Respect & Inclusivity
- Learning goals:
  - Understand how social media APIs work
  - Collect, visualize, and store tweets using `rtweet`
  - Understand how to prepare tweets for analysis

Introduction

- A bit about me (and Shane!)
- Code of Conduct: Respect & Inclusivity
- Learning goals:
  - Understand how social media APIs work
  - Collect, visualize, and store tweets using `rtweet`
  - Understand how to prepare tweets for analysis
  - Analyze tweet corpus with text analysis

Web APIs

What's an API anyway?

- Application Programming Interface (API)

What's an API anyway?

- Application Programming Interface (API)
- Broadly: APIs receive principled requests for data and send responses from a remote server

What's an API anyway?

- Application Programming Interface (API)
- Broadly: APIs receive principled requests for data and send responses from a remote server
- They comprise a huge portion of the internet

What's an API anyway?

- Application Programming Interface (API)
- Broadly: APIs receive principled requests for data and send responses from a remote server
- They comprise a huge portion of the internet
  - Most connections between apps, websites, servers!

What's an API anyway?

- Restaurant metaphor:

What's an API anyway?

- Restaurant metaphor:
  - YOU are a diner at a restaurant

What's an API anyway?

- Restaurant metaphor:
  - `YOU` are a diner at a restaurant
  - The `WAITER` is an API

What's an API anyway?

- Restaurant metaphor:
  - `YOU` are a diner at a restaurant
  - The `WAITER` is an API
  - The `MENU` is the data the restaurant can serve you (with rules/restrictions)

What's an API anyway?

- Restaurant metaphor:
  - `YOU` are a diner at a restaurant
  - The `WAITER` is an API
  - The `MENU` is the data the restaurant can serve you (with rules/restrictions)
  - The `KITCHEN` is the server that hosts the data

What's an API anyway?

- Restaurant metaphor:
  - `YOU` are a diner at a restaurant
  - The `WAITER` is an API
  - The `MENU` is the data the restaurant can serve you (with rules/restrictions)
  - The `KITCHEN` is the server that hosts the data
  - The `WAITER` delivers from the `FOOD` from the `KITCHEN` to `YOU`

What's an API anyway?

- Businesses use APIs for many reasons

What's an API anyway?

- Businesses use APIs for many reasons
  - Banks to provide secure access to account info

What's an API anyway?

- Businesses use APIs for many reasons
  - Banks to provide secure access to account info
  - Third party apps (Twitter, Reddit, etc.)

What's an API anyway?

- Businesses use APIs for many reasons
  - Banks to provide secure access to account info
  - Third party apps (Twitter, Reddit, etc.)
  - Google Maps API to apps that want to provide directions

What's an API anyway?

- Businesses use APIs for many reasons
  - Banks to provide secure access to account info
  - Third party apps (Twitter, Reddit, etc.)
  - Google Maps API to apps that want to provide directions
- We use APIs to get data/run experiments

What's an API anyway?

- Businesses use APIs for many reasons
  - Banks to provide secure access to account info
  - Third party apps (Twitter, Reddit, etc.)
  - Google Maps API to apps that want to provide directions
- We use APIs to get data/run experiments
  - Facebook, Instagram, Telegram, YouTube, Google, Reddit, Gab, MTurk, Gmail, etc.

What's an API anyway?

- Businesses use APIs for many reasons
  - Banks to provide secure access to account info
  - Third party apps (Twitter, Reddit, etc.)
  - Google Maps API to apps that want to provide directions
- We use APIs to get data/run experiments
  - Facebook, Instagram, Telegram, YouTube, Google, Reddit, Gab, MTurk, Gmail, etc.
  - Twitter!

DESIGNED BY
Mehdi Medjaoui

The API Landscape

Last Update: March, 2017

POWERED BY
spoke Intelligence

Business Processes as an API/API-as a Product/Transactional APIs (151)

API Lifecycle platform (78)

Backend Building Tools/MBaaS (14)

API Abstraction / Integration Platform (32)

APIs vs Scraping

- Huge benefit of APIs is they serve data in principled, consistent manner

APIs vs Scraping

- Huge benefit of APIs is they serve data in principled, consistent manner
- Obscures much of the complicated behind-the-scenes activity

APIs vs Scraping

- Huge benefit of APIs is they serve data in principled, consistent manner
- Obscures much of the complicated behind-the-scenes activity
- Host server knows what to expect and how to provide data

APIs vs Scraping

- Huge benefit of APIs is they serve data in principled, consistent manner
- Obscures much of the complicated behind-the-scenes activity
- Host server knows what to expect and how to provide data
- We know exactly what data we can get + how

APIs vs Scraping

- Huge benefit of APIs is they serve data in principled, consistent manner
- Obscures much of the complicated behind-the-scenes activity
- Host server knows what to expect and how to provide data
- We know exactly what data we can get + how
- We typically scrape data that was assembled for a different purpose

APIs vs Scraping

- Huge benefit of APIs is they serve data in principled, consistent manner
- Obscures much of the complicated behind-the-scenes activity
- Host server knows what to expect and how to provide data
- We know exactly what data we can get + how
- We typically scrape data that was assembled for a different purpose
    - Presentational code meant to be rendered in a browser

APIs vs Scraping

- Huge benefit of APIs is they serve data in principled, consistent manner
- Obscures much of the complicated behind-the-scenes activity
- Host server knows what to expect and how to provide data
- We know exactly what data we can get + how
- We typically scrape data that was assembled for a different purpose
    - Presentational code meant to be rendered in a browser
    - Mucked up with HTML, etc.

Why APIs are Awesome

- Collecting data at scale with EASE

Why APIs are Awesome

- Collecting data at scale with EASE
- Automating tasks - DMing, following, etc.

Why APIs are Awesome

- Collecting data at scale with EASE
- Automating tasks - DMing, following, etc.
    - Customer service accounts on Twitter

Why APIs are Awesome

- Collecting data at scale with EASE
- Automating tasks - DMing, following, etc.
  - Customer service accounts on Twitter
  - [Bot examples]

Why APIs are Awesome

- Collecting data at scale with EASE
- Automating tasks - DMing, following, etc.
  - Customer service accounts on Twitter
  - [Bot examples]
- Programming packages make this even easier

Why APIs are Awesome

- Collecting data at scale with EASE
- Automating tasks - DMing, following, etc.
  - Customer service accounts on Twitter
  - [Bot examples]
- Programming packages make this even easier
- Yes, they can also be used for evil (Cambridge Analytica, etc)... but lets ignore that

General Social Media API Tips

1. If you want something, get it ASAP

General Social Media API Tips

1. If you want something, get it ASAP
   - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)

General Social Media API Tips

1. If you want something, get it ASAP
    - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)
2. Think about storage ahead of time

General Social Media API Tips

1. If you want something, get it ASAP
   - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)
2. Think about storage ahead of time
   - Social media sites produce massive amounts of data

General Social Media API Tips

1. If you want something, get it ASAP
   - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)
2. Think about storage ahead of time
   - Social media sites produce massive amounts of data
   - Where should the data be stored? How?

General Social Media API Tips

1. If you want something, get it ASAP
   - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)
2. Think about storage ahead of time
   - Social media sites produce massive amounts of data
   - Where should the data be stored? How?
3. Build scale into research design

General Social Media API Tips

1. If you want something, get it ASAP
   - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)
2. Think about storage ahead of time
   - Social media sites produce massive amounts of data
   - Where should the data be stored? How?
3. Build scale into research design
   - Can you reasonably analyze what you collect? Can you sample from it?

General Social Media API Tips

1. If you want something, get it ASAP
   - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)
2. Think about storage ahead of time
   - Social media sites produce massive amounts of data
   - Where should the data be stored? How?
3. Build scale into research design
   - Can you reasonably analyze what you collect? Can you sample from it?
4. Beware limits (remember the `menu`)

General Social Media API Tips

1. If you want something, get it ASAP
   - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)
2. Think about storage ahead of time
   - Social media sites produce massive amounts of data
   - Where should the data be stored? How?
3. Build scale into research design
   - Can you reasonably analyze what you collect? Can you sample from it?
4. Beware limits (remember the `menu`)
   - Easier with programming packages [now]

General Social Media API Tips

1. If you want something, get it ASAP
   - APIs constantly change, oftentimes for the worse (Twitter), but sometimes for the better (Telegram)
2. Think about storage ahead of time
   - Social media sites produce massive amounts of data
   - Where should the data be stored? How?
3. Build scale into research design
   - Can you reasonably analyze what you collect? Can you sample from it?
4. Beware limits (remember the `menu`)
   - Easier with programming packages [now]
   - Can you collect what you want to?

Twitter API

What can you do?

- Search, stream, follow, unfollow, DM, etc.

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do
  - Most data YOU can see, the API can get

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do
  - Most data YOU can see, the API can get
  - Regarding functionality: think bots...

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do
  - Most data YOU can see, the API can get
  - Regarding functionality: think bots...
  - However, this is getting harder post-2018

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do
  - Most data YOU can see, the API can get
  - Regarding functionality: think bots...
  - However, this is getting harder post-2018
- Only public accounts, priv. accounts return user info

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do
  - Most data YOU can see, the API can get
  - Regarding functionality: think bots...
  - However, this is getting harder post-2018
- Only public accounts, priv. accounts return user info
- [Rate limits] change occasionally

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do
  - Most data YOU can see, the API can get
  - Regarding functionality: think bots...
  - However, this is getting harder post-2018
- Only public accounts, priv. accounts return user info
- [Rate limits] change occasionally
- [Example] of request/response

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do
  - Most data YOU can see, the API can get
  - Regarding functionality: think bots...
  - However, this is getting harder post-2018
- Only public accounts, priv. accounts return user info
- [Rate limits] change occasionally
- [Example] of request/response
- Search the past 7 days, otherwise $$$

What can you do?

- Search, stream, follow, unfollow, DM, etc.
  - Most tasks YOU can do, the API can do
  - Most data YOU can see, the API can get
  - Regarding functionality: think bots...
  - However, this is getting harder post-2018
- Only public accounts, priv. accounts return user info
- [Rate limits] change occasionally
- [Example] of request/response
- Search the past 7 days, otherwise $$$
- The API returns JSON, but `rtweet` flattens it

Edit profile
Caleb Lucas
@calebjlucas
PhD candidate in political science @michiganstateu | civil war, terrorism, mena | @universityofky alum
East Lansing, MI
caleblucas.com
Joined December 2012
227 Following
210 Followers

Tweets 72 | Following 227 | Followers 210 | Likes 439

Caleb Lucas

@calebjlucas

PhD candidate in political science @michiganstateu | civil war, terrorism, mena | @universityofky alum

East Lansing, MI

caleblucas.com

Joined December 2012

This account's Tweets are protected.

Only confirmed followers have access to @calebjlucas's Tweets and complete profile. Click the "Follow" button to send a follow request.

Caleb Lucas
72 Tweets

Tweets | Tweets & replies | Media | Likes

You Retweeted

Jessica Schoenherr @jessica_ann87 · Sep 13
Replying to @ProfHansNoel
I am on the market! I am a Ph.D. Candidate at @MSU_poli_sci studying American political institutions and the U.S. Supreme Court. My research focuses on how attorneys use written legal arguments to influence the justices and the law.

Welcome!
Ph.D. Candidate, Michigan State University
jaschoenherr.com

11 19

You Retweeted

Aycan Katitas @Aycan_ka · Sep 13
Replying to @ProfHansNoel
ABD @PoliticsUVA, I advance a top-down theory of trade preferences, exploring subnational variation of opposition to understand how trade resurged as a salient issue recently - Candidates tap into ethnocentric districts with anti-trade rhetoric to win elections in the U.S.

1 12 19

Trends

Trending in United States
#millennialretirementplans
5,193 Tweets

#HowToBeABoss
Anna Wintour's MasterClass on creativity, leadership, and owning it
Promoted by MasterClass

Trending in United States
Bork
5,082 Tweets

Trending in United States
Hispanics
37.3K Tweets

Politics
'Who do you like more, the country or the Hispanics?' Trump asks advisor Steve Cortes

Trending in United States
The Man Who Sold America
8,874 Tweets

Trending in United States
#InsteadOfFlyingCarsWeHave
3,231 Tweets

Trending in United States
Saved by the Bell
Trending with: Battlestar Galactica, Punky Brewster

Television
Saved by the Bell, Punky Brewster and Battlestar Galactica reboots coming to NBC's Peacock

Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

rtweet

The `rtweet` Package

- Constantly updated, good documentation

The `rtweet` Package

- Constantly updated, good documentation
  - Some bugs, but lets help improve it!

The `rtweet` Package

- Constantly updated, good documentation
  - Some bugs, but lets help improve it!
- `tweepy` is excellent Python package

The `rtweet` Package

- Constantly updated, good documentation
  - Some bugs, but lets help improve it!
- `tweepy` is excellent Python package
- Twitter is constantly updating/maintaining/changing the API

The `rtweet` Package

- Constantly updated, good documentation
  - Some bugs, but lets help improve it!
- `tweepy` is excellent Python package
- Twitter is constantly updating/maintaining/changing the API
  - Difficult to maintain an error-free package that interacts with the API

The `rtweet` Package

- Constantly updated, good documentation
  - Some bugs, but lets help improve it!
- `tweepy` is excellent Python package
- Twitter is constantly updating/maintaining/changing the API
  - Difficult to maintain an error-free package that interacts with the API
  - Be mindful of this if you encounter an error

The `rtweet` Package

- Constantly updated, good documentation
  - Some bugs, but lets help improve it!
- `tweepy` is excellent Python package
- Twitter is constantly updating/maintaining/changing the API
  - Difficult to maintain an error-free package that interacts with the API
  - Be mindful of this if you encounter an error
  - Old code can be problematic, API can change too

R sesh!

R sesh!

- This PDF:
  `caleblucas.com/files/rtweets_ssda.pdf`
- Jupyter NB:
  `caleblucas.com/files/rtweets_ssda.ipynb`
- Jupyter NB HTML:
  `caleblucas.com/files/rtweets_ssda.html`
- R Script:
  `caleblucas.com/files/ssda.R`