Class 8: Scraping Social Media Data

MAST5953: Web Scraping and Text Mining

Dr. Miriam Sorace

www.miriamsorace.net

04 February 2021

Outline of Today's Class

Social Media Data

Twitter API Access

Scraping Tweets with R

Social Media Data

Ideal Point Estimation Using Twitter Data

Barbera, P. 2015 *Political Analysis*

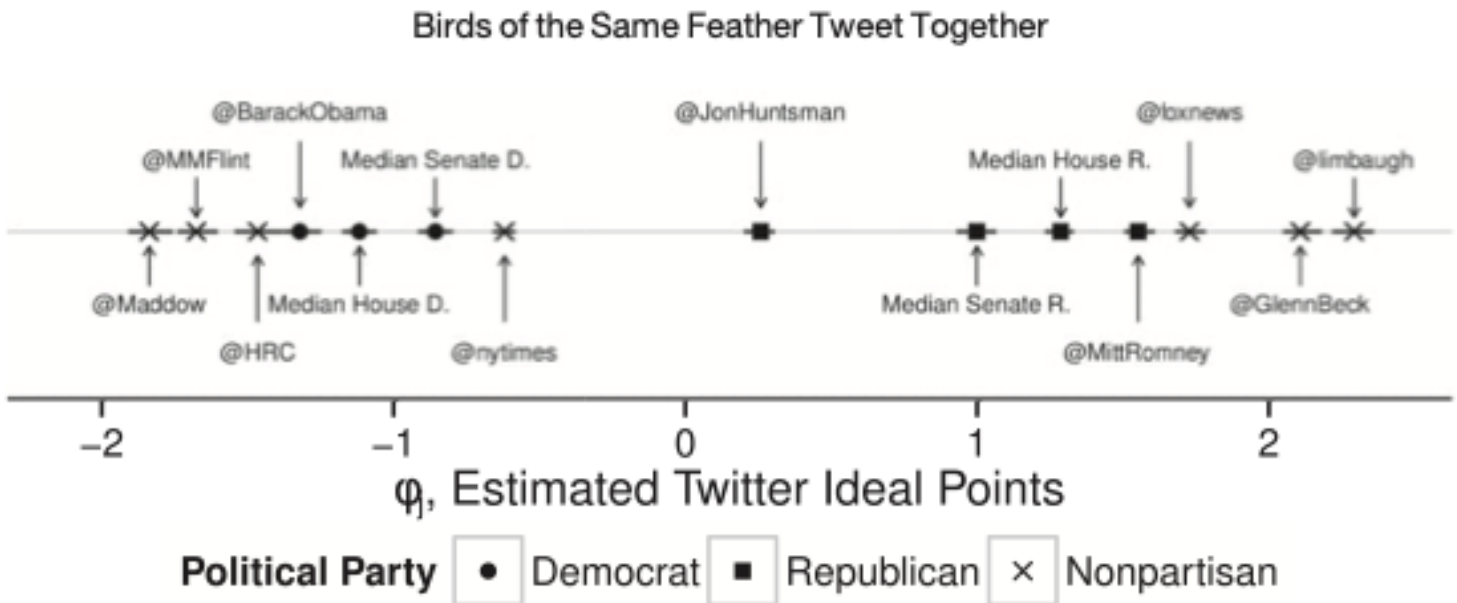


Fig. 2 Ideal point estimates for key political actors in the United States.

Predicting the Brexit Vote Using Twitter Data

Lopez et al. 2018 *Statistics, Politics and Policy*

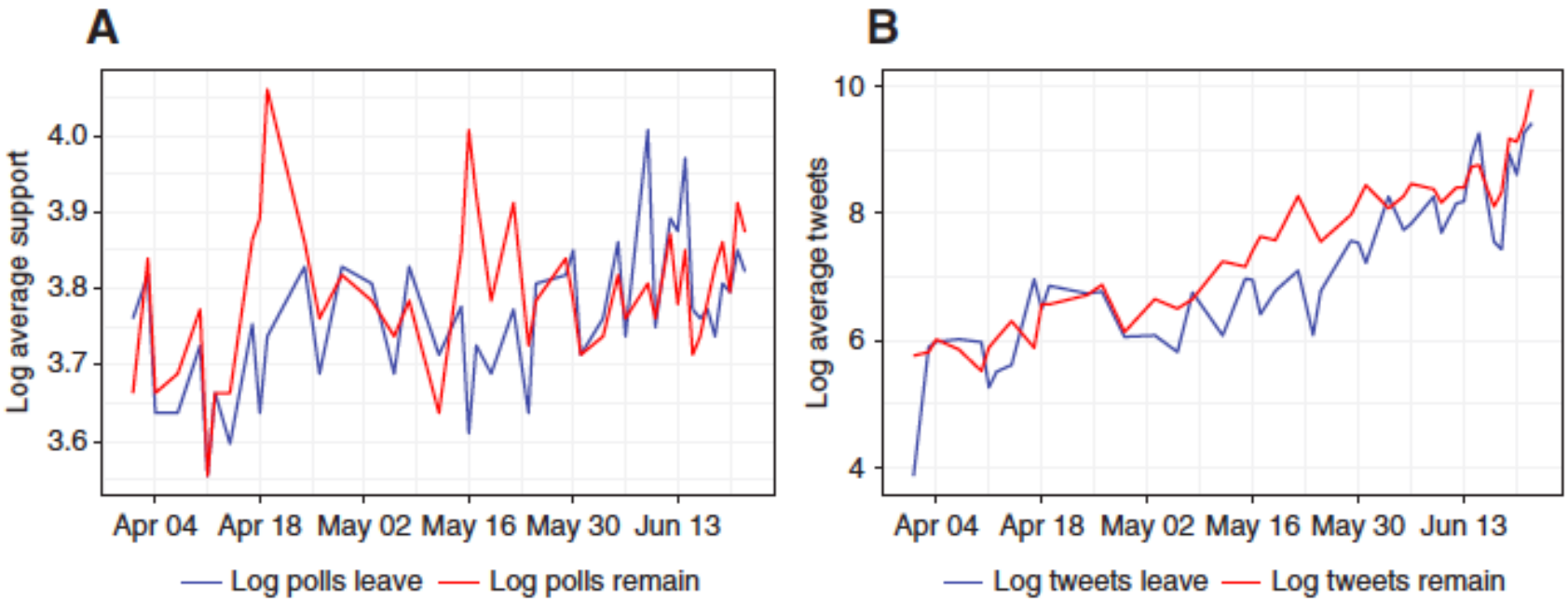
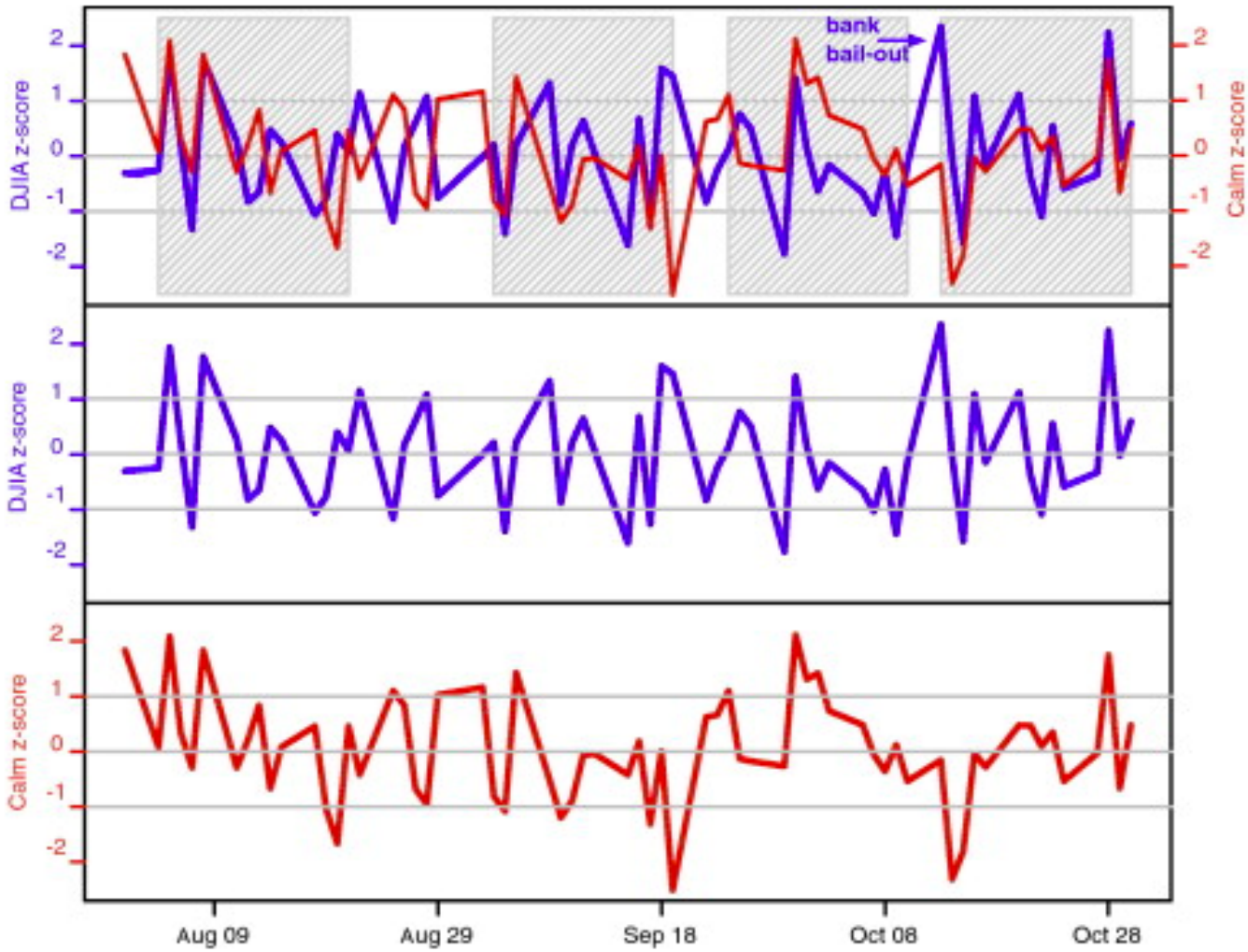


Figure 1: Time-series of Opinion Polls and Tweets. (A) Opinion Polls (B) Twitter.

Twitter Mood Predicts the Stock Market

Bollen et al. 2011 *Journal of Computational Science*



Influenza Surveillance through Twitter

Broniatowski et al. 2013 *PLOS One*

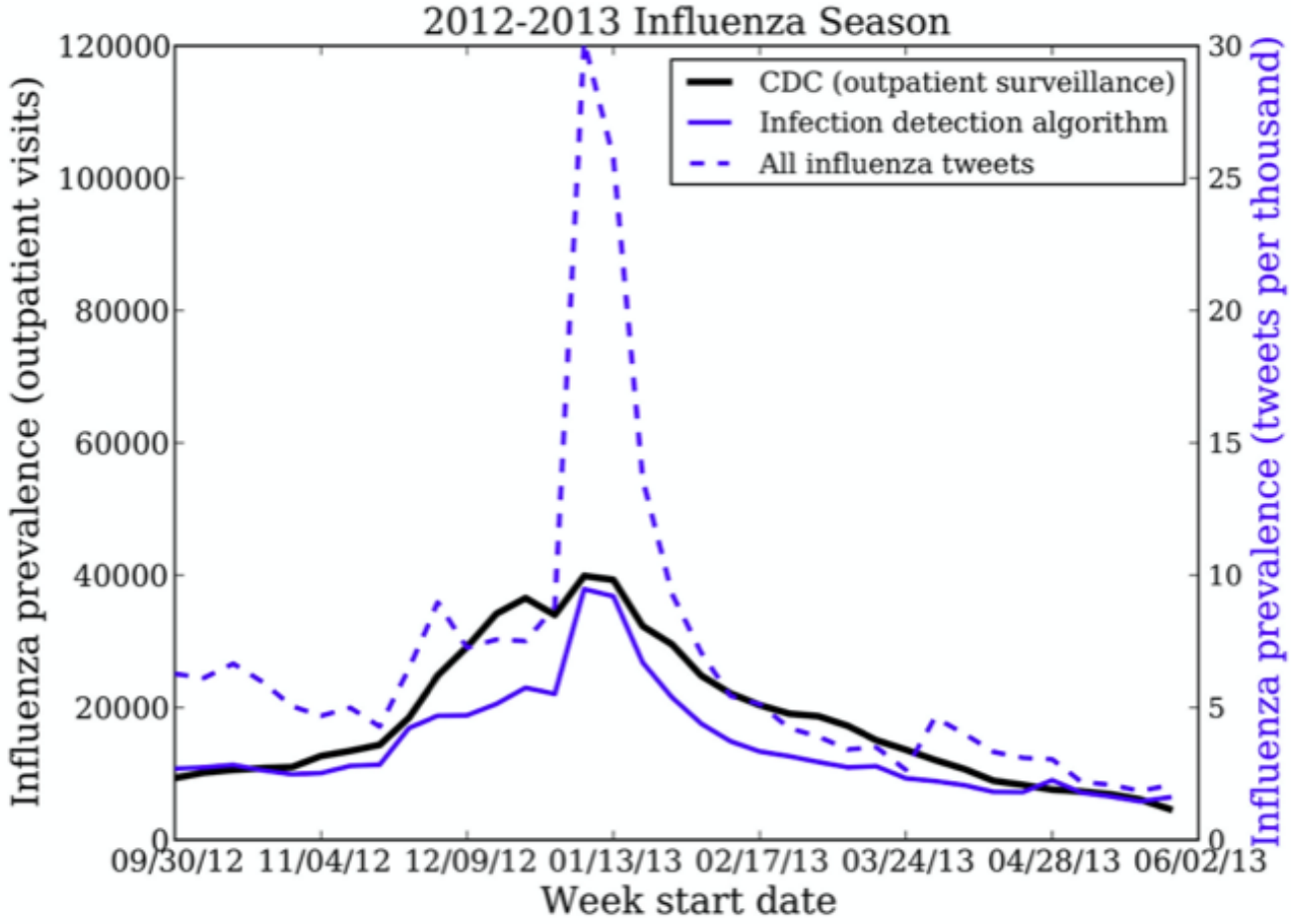


Figure 2. 2012-2013 national influenza rates from Twitter and CDC surveillance. This figure shows national influenza rates of the United States as predicted by two Twitter-based algorithms alongside the influenza-like illness surveillance network data from the US Centers for Disease Control and Prevention (CDC). The dashed blue line is the measure estimated by a simple model of keyword matching, while the solid blue line is the measure estimated by our new infection detection model. Our new algorithm more closely matches the CDC data (solid black line), while the simpler keyword model infers spurious spikes due to other Twitter chatter, e.g. in early December and early April.

doi: 10.1371/journal.pone.0083672.g002

Social Media Data: Potential

- Most of contemporary political behaviour takes place online through social media networks: affordable field experiments possible
- It can be used to produce valid measures of users' ideology.
- Real-time data about people's experiences & beliefs: useful for prediction and prevention (*wisdom of the crowds* effect)

Social Media Data: Pitfalls

- Representativeness:
  - 22% of US citizens are on Twitter
  - Tweeters are younger, more educated and more liberal than general US public
  - 80% of US tweets are created by 10% of users
  - see `https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/`
  - Spam and bots: not all users are real
- More useful in studying elites (political, media, interest groups)
- Ethics: difficult to claim informed consent of users. Protect privacy!

Twitter API Access

Revision: Scraping Methods

1. Regular Expressions
2. Node queries
3. APIs (*Application Programming Interfaces*)

Revision: APIs

- A 'communication' method (interface) that allows a user to access somebody else's code and implementation
- APIs trigger prescribed actions when invoked
- Check this video out!

Twitter API Access

- To access Twitter APIs you need to:
  1. Set up a Twitter Account if you don't have one
  2. Apply for a Twitter Developer Account
  3. Create a Twitter application
  4. To scrape: set up a connection to either REST or Streaming API via `R` wrapper functions (we'll see demo of this later)

First Things First: Twitter API Terms of Use

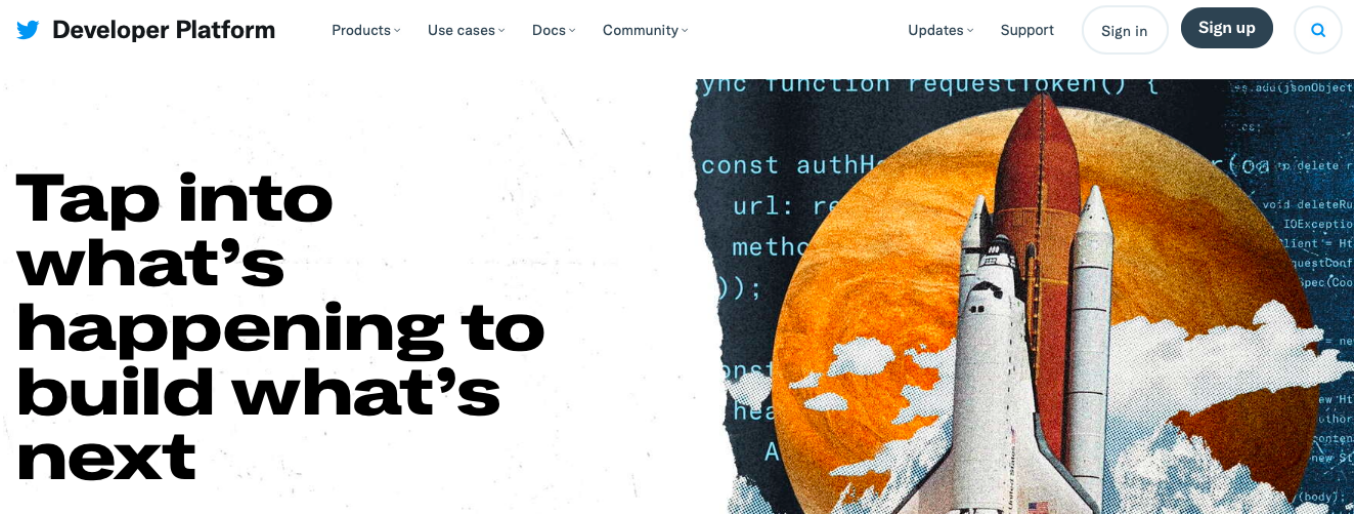
Steps

- Go to:
  - `https://developer.twitter.com/en/developer-terms/agreement-and-policy`
  - `https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases`
- **Lab work (in pairs of groups): read these terms, summarize key ones relevant to our intended usage**
  - Save them and keep them in mind going forwards!!

Apply for a Developer Account

Steps

- Go to `https://developer.twitter.com/en` 'Sign Up' on the top right & follow instructions
- If you do not have a Twitter account, click 'Sign up' again (on the bottom), if you do sign in with your Twitter handle & password



Apply for a Developer Account

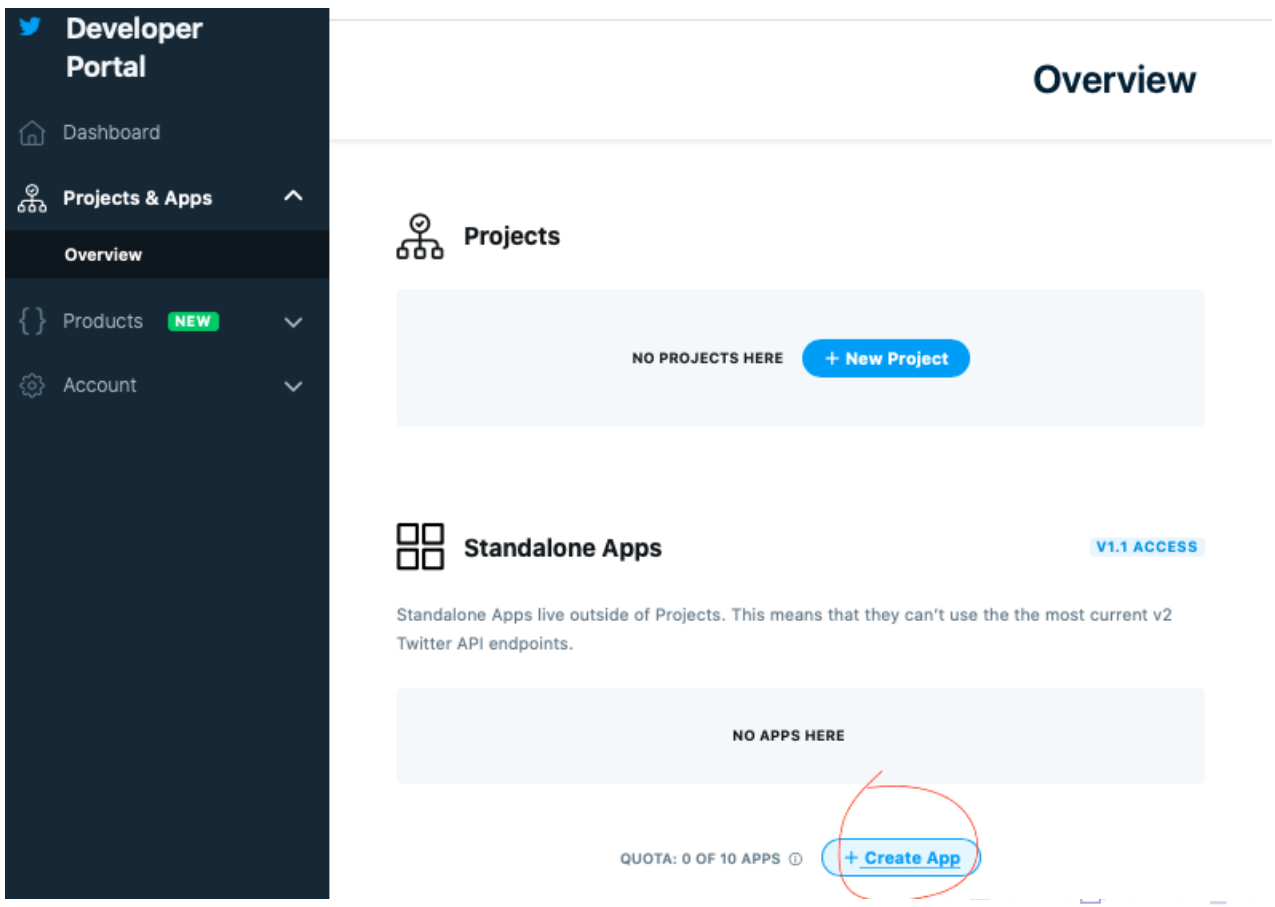
Steps

- You will be re-directed to some sort of application form, where you need to:
  - input your details (user, email, country)
  - write-up your 'Reasons of Use' and 'Planned Use'.
  - **Lab work: complete your application forms, ask fellow classmate(s) to proofread/advise**:
    - Mention something like: 'I am learning text analysis via machine learning as part of a University project and we're using politicians' Tweets as texts' ... 'I will use sentiment analysis and topic modelling algorithms on the Tweets, but not intended for publication'
    - Individual tweets won't be submitted anywhere just aggregate results from text analysis will be submitted as part of the coursework
    - Reassure Twitter your developer access has educational scopes (non-commercial), and that you won't breach the terms of use/ethical standards

Apply for a Developer Account

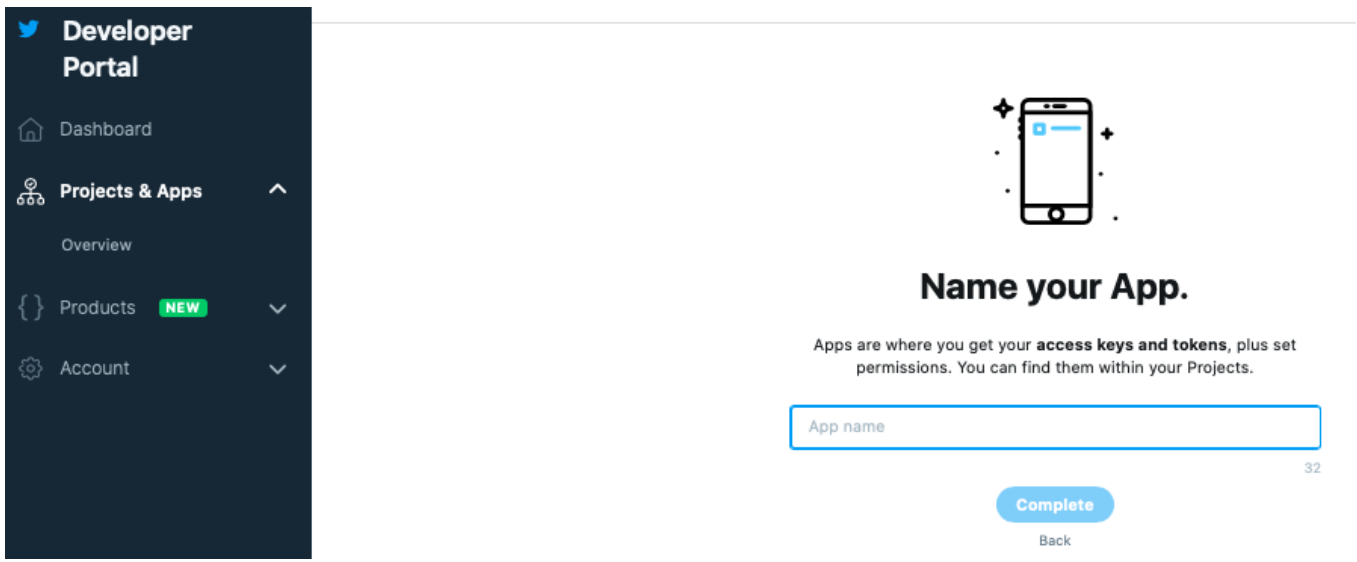
Steps

- You will be directed to the Twitter developer dashboard: `https://developer.twitter.com/en/apps/`
- In 'Project & Apps' click 'Overview', then 'Create App'



Apply for a Developer Account

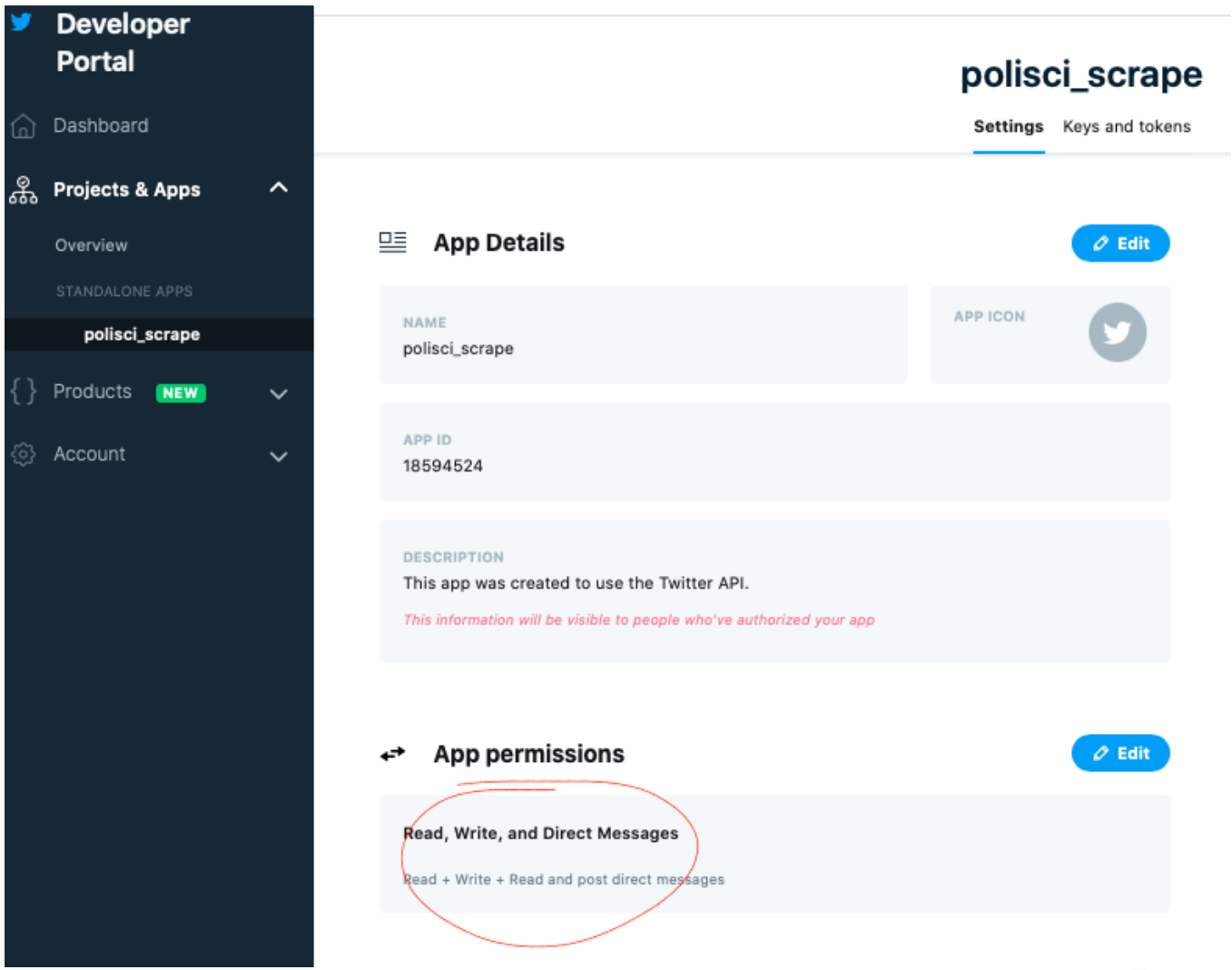
Steps



- Use something like 'MAST5953_' + your Kent userID ... something that is unlikely to exist already

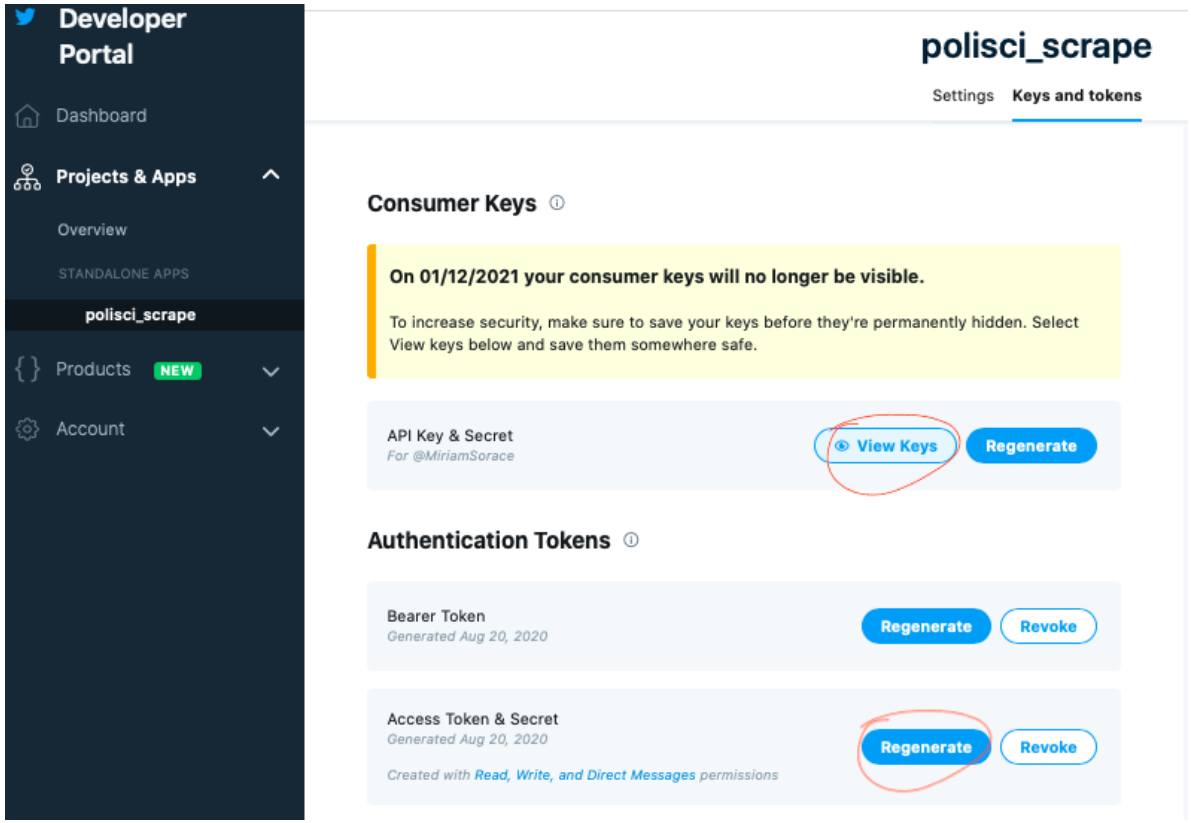
Apply for a Developer Account

Steps



Apply for a Developer Account

Steps



- The 'View Keys' option is not disabled, save your API credentials as they are generated, or regenerate them.
- You need to store both **Keys** and **Authentication** credentials (usually called secret & tokens)

Scraping Tweets with R

Authenticate via OAuth

```
library(ROAuth)
my_oauth <- list(consumer_key = "CONSUMER_KEY",
    consumer_secret = "CONSUMER_SECRET",
    access_token="ACCESS_TOKEN",
    access_token_secret = "ACCESS_TOKEN_SECRET")
save(my_oauth, file="~/my_oauth")
```

Authenticate via OAuth

- If R returns this error:
  - 'Error: Not a valid access token.'
- it means that your Keys and Authentication codes were copied/pasted incorrectly: go back to the Developer Dashboard, regenerate them and copy/paste the correct ones in the correct fields!

Twitter Rest API

- Used to retrieve *static* information
  - individual user's data, friends and followers
  - searches by keyword

Twitter Streaming API

- Used to retrieve *real-time* information
  - tweets as they happen
  - access to global data stream

Core Function

Scrape User Tweets

- `library("rtweet")`
- `get_timelines(c("BarackObama"), n = 3200, parse=T, token=my_oauth)`

R code demonstration

- Code for Twitter Scraping assignment
- Advanced code demo

What we have learnt today ...

- APIs wrapper function logic
- Twitter APIs
  - REST
  - Streaming
- The authentication steps to gain access to Twitter APIs
- Scraping user tweets or keyword-based tweets
- Parsing and converting Twitter data into .csv