Dictionary-Based Classification of Text:
Demonstration and Possible Applications

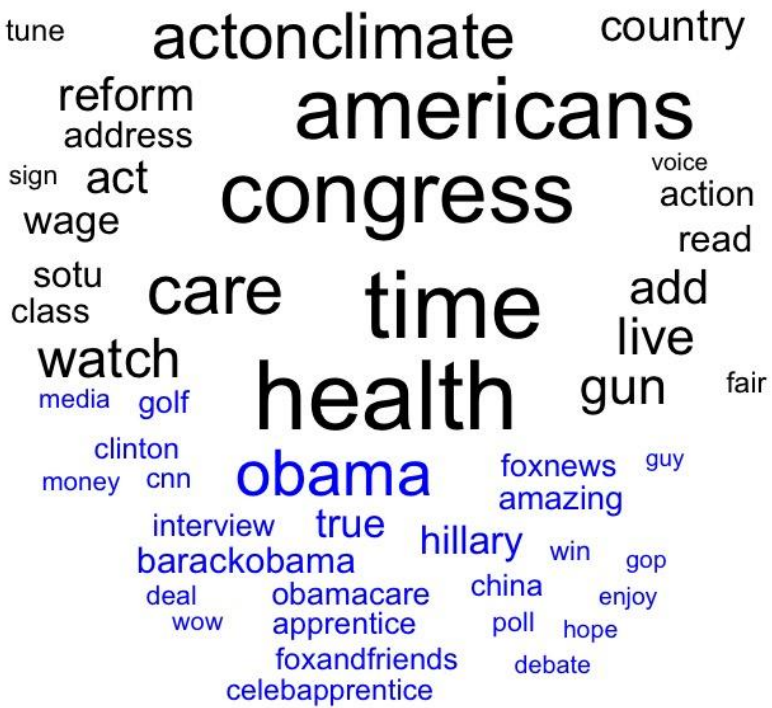
Oja Pathak, Jordan Dewar,
Henry Overos
April 19, 2021

iLCSS

Roadmap

1. What is Text Analysis?
    a. Why is it helpful for Political Science?
    b. General types of Text Analysis methods
    c. Four important principles of Text Analysis
2. What does Text Analysis look like in R?
3. Example: Coding data on ethnic minorities through news articles
4. Application: Twitter data and posting around mass shootings
    a. How to get Twitter data through an Academic Developer account
    b. Using Text Analysis to determine changes in Twitter conversations over time

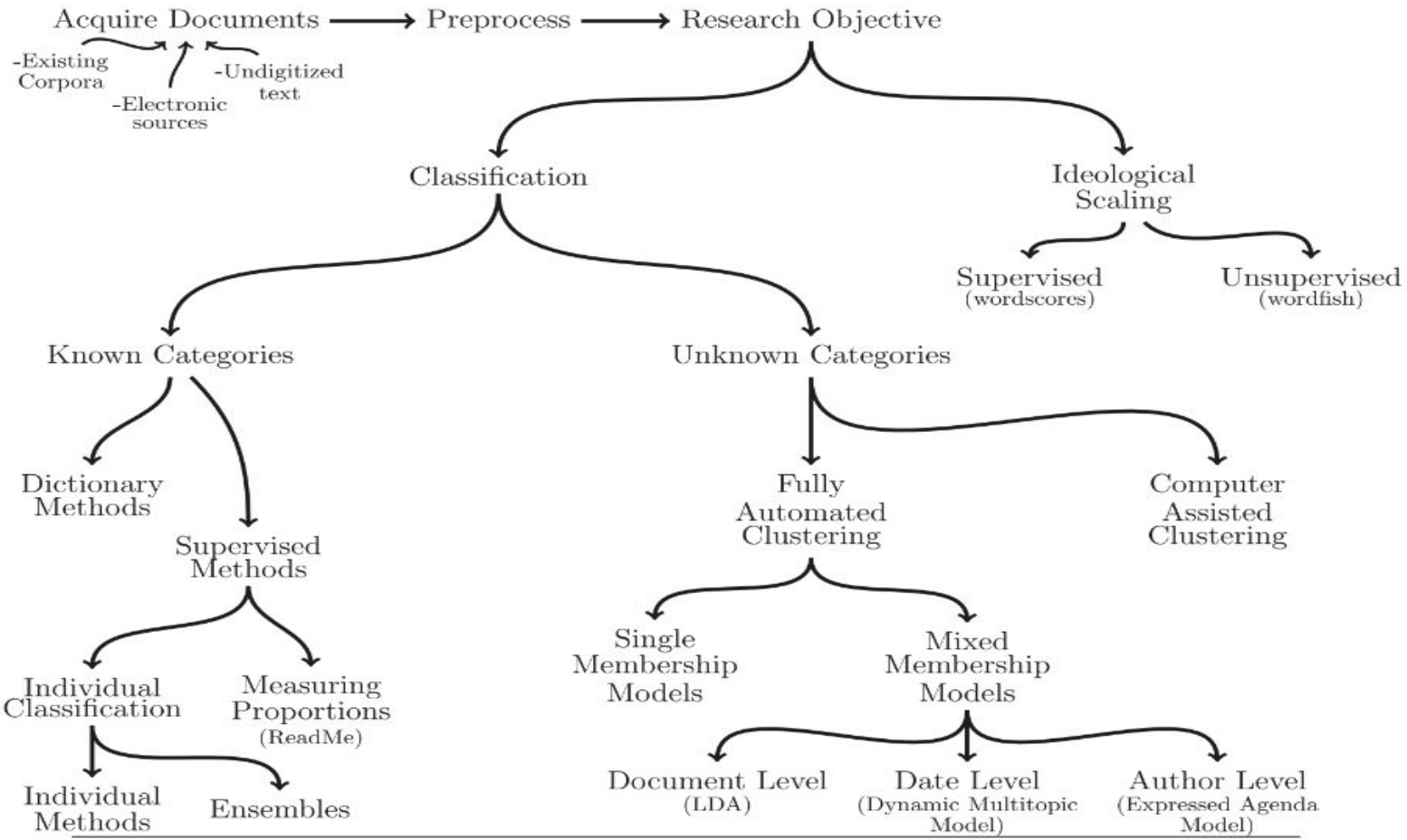
Obama


tune
actonclimate
country
reform
americans
address
sign
act
congress
voice
action
wage
read
sotu
care
time
add
class
live
watch
health
gun
fair
media
golf
clinton
money
cnn
obama
foxnews
guy
amazing
interview
true
hillary
win
barackobama
gop
deal
obamacare
china
enjoy
wow
apprentice
poll
hope
foxandfriends
debate
celebapprentice

Trump

Introduction to Text Analysis

- Text analysis is
  - a way to find patterns in documents
  - theory driven
  - computational but requires human knowledge and interpretation
- Reduces the cost of analyzing a large collection of texts
- Let’s say you acquire text documents. Now what?



Grimmer and Stewart (2013): Overview of Text Analysis

Usage in Political Science

- Useful in a variety of political contexts
  - Analysing the reuse of language in bills and international treaties (Allee and Elsig 2019)
  - Identifying how movements strategize to draw in new audiences (Nielsen 2020)
  - Using dictionary-based classifiers for understanding the link between propaganda and protests (Carter and Carter 2020)
  - Using text to predict armed conflict (Mueller and Rauh 2017)

1: Text Analysis Models are Wrong but Useful

- The goal of TA is *not causation* but prediction
- Ex: *He eats, shoots, and leaves.* vs. *He eats shoots and leaves.*
- They do help overcome practical social science problems (categorization, topic modeling, sentiment analysis, etc.)

2. They Cannot Replace Humans

- Still need careful thought by researchers
- Processing and decision making still necessary
- Could amplify human abilities

3. No one model fits all

- Dataset-dependent
- Goal of the research
- Research design
- Research Question

4. VALIDATE

- Validation, above all
  - Can you replicate through human coding?
  - What other evidence do you have?

Running Text Analysis in R

Well-known, supported packages:

- Quanteda (https://quanteda.io/)
- Tidytext (https://cran.r-project.org/web/packages/tidytext/index.html)

Example: From raw text to usable data

```
> texts

[1] "Four score and seven years ago"

[2] "Hello, world!"

[3] "I coulda had class. I coulda been a contender. I coulda been somebody, instead of a bum,
which is what I am, let's face it."'
```

1. Text to Corpus

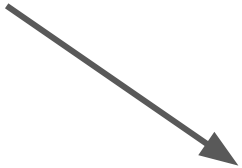
```
> corpus <- corpus(texts)
> corpus
Corpus consisting of 3 documents.

text1 :
"Four score and seven years ago"

text2 :
"Hello, world!"

text3 :
"I coulda had class. I coulda been a contender. I coulda been..."
```

2. Corpus to Tokens

```
> toks <- tokens(corpus, remove_punct =T)
> toks <- tokens_tolower(toks)
> toks
Tokens consisting of 3 documents.
text1 :
[1] "four"  "score" "and"   "seven" "years" "ago"
text2 :
[1] "hello" "world"
text3 :
 [1] "i"         "coulda"    "had"       "class"     "i"         "coulda"    "been"      "a"         "contender"
[10] "i"         "coulda"    "been"
[ ... and 13 more ]
```

3. Document-Feature Matrices

Document	Text
1	four score and seven years ago
2	hello world



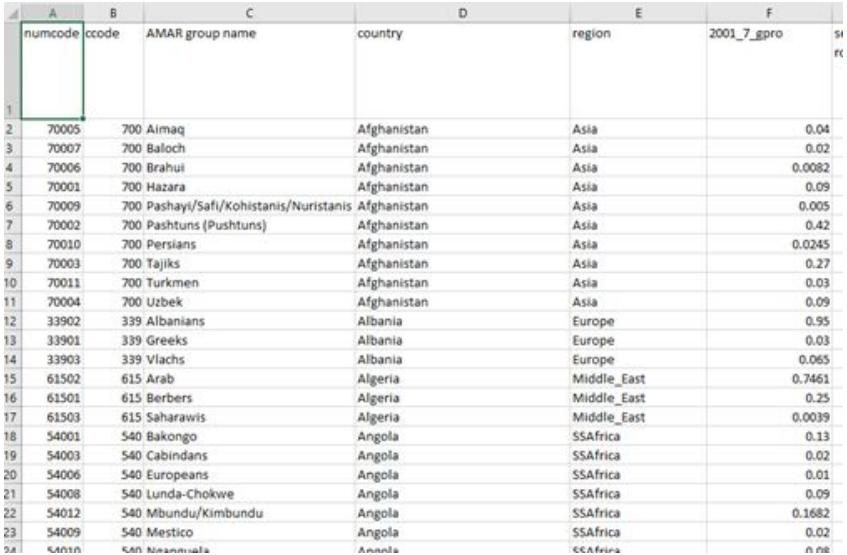
Doc	four	score	and	seven	years	ago	hello	world
1	1	1	1	1	1	1	0	0
2	0	0	1	0	0	0	1	1

3. Document-Feature Matrices

```
> dfm <- dfm(toks)
> dim(dfm)
[1]  3 26
> dfm
Document-feature matrix of: 3 documents, 26 features (66.7% sparse).
       features
 docs    four score and seven years ago hello world i coulda
  text1    1     1   1     1     1   1     0     0 0      0
  text2    0     0   0     0     0   0     1     1 0      0
  text3    0     0   0     0     0   0     0     0 4      3
[ reached max_nfeat ... 16 more features ]
```

Research Example: Updating the All Minorities at Risk Data Frame

- *All Minorities at Risk* (AMAR)
  - Large data frame coding for global ethnic minorities
- To update AMAR
  - Sift through openly available news media
  - For data on over 1200+ socially relevant ethnic groups
  - Across time



	A	B	C	D	E	F
1	numcode	ccode	AMAR group name	country	region	2001_7_gpro
2	70005	700	Aimaq	Afghanistan	Asia	0.04
3	70007	700	Baloch	Afghanistan	Asia	0.02
4	70006	700	Brahui	Afghanistan	Asia	0.0082
5	70001	700	Hazara	Afghanistan	Asia	0.09
6	70009	700	Pashayi/Safi/Kohistanis/Nuristanis	Afghanistan	Asia	0.005
7	70002	700	Pashtuns (Pushtuns)	Afghanistan	Asia	0.42
8	70010	700	Persians	Afghanistan	Asia	0.0245
9	70003	700	Tajiks	Afghanistan	Asia	0.27
10	70011	700	Turkmen	Afghanistan	Asia	0.03
11	70004	700	Uzbek	Afghanistan	Asia	0.09
12	33902	339	Albanians	Albania	Europe	0.95
13	33901	339	Greeks	Albania	Europe	0.03
14	33903	339	Vlachs	Albania	Europe	0.065
15	61502	615	Arab	Algeria	Middle_East	0.7461
16	61501	615	Berbers	Algeria	Middle_East	0.25
17	61503	615	Saharawis	Algeria	Middle_East	0.0039
18	54001	540	Bakongo	Angola	SSAfrica	0.13
19	54003	540	Cabindans	Angola	SSAfrica	0.02
20	54006	540	Europeans	Angola	SSAfrica	0.01
21	54008	540	Lunda-Chokwe	Angola	SSAfrica	0.09
22	54012	540	Mbundu/Kimbundu	Angola	SSAfrica	0.1682
23	54009	540	Mestico	Angola	SSAfrica	0.02
24	54010	540	Nzanguela	Angola	SSAfrica	0.08

The Proposed Solution: Machine-Assisted Text Categorization

Overview

1. Create a dictionary of seed words related to classes/topics
2. Label documents that contain seed words from dictionary
3. Estimate a naïve Bayes classifier on the labeled documents from step 2

Table 2: Dictionary of U.S. AMAR Group Search Terms

AMAR Class	Seed-Words
Black	african_american*, black_american*, black_lives_matter, black*
Asian	asian_american*, asia*, east_asia*, southeast_asia*, asian_immigrant*, china, chinese, korea*, japan*, india*, phillipin*
Latin American:	hispanic*, latin_american*, latin*, illegal_immigrant*, mexic*, guatamal*, hondur*, peru*, cuba*, puerto
Muslim:	muslim*, islam*, hijab, hajj, ramadan, halal, imam*, head-scar*, mosque*, sharia, koran, cleric*
Jewish:	jew, jews, jewish, synagogue*, rabbi*, hasidic, orthodox, israel, torah
Native American	native_american*, american_indian*, indigenous, tribe*, tribal, sioux, cherokee, reservation

Test Run: Identifying articles in newspapers about protest

- Protest is a documented variable in the AMAR Data Frame
- It should be an easy test, we expect variation in protest by ethnic group and across time

Data: *All the News* Dataset

- Used an open-source dataset of 150,000 US newspaper articles collected from 2016-2017
- Sampled 45,000 articles grouped by publication and time

Table 3: Distribution of Articles by Publication in US News Dataset

	Original Data		Sample Data	
Publication	N	Percent	N	Percent
Atlantic	7179	0.05	3000	0.07
Breitbart	23781	0.17	3000	0.07
Business Insider	6757	0.05	3000	0.07
Buzzfeed News	4854	0.03	3000	0.07
CNN	11488	0.08	3000	0.07
Fox News	4354	0.03	3000	0.07
Guardian	8681	0.06	3000	0.07
National Review	6203	0.04	3000	0.07
New York Post	17493	0.12	3000	0.07
New York Times	7803	0.05	3000	0.07
NPR	11992	0.08	3000	0.07
Reuters	10710	0.08	3000	0.07
Talking Points Memo	5214	0.04	3000	0.07
Vox	4947	0.03	3000	0.07
Washington Post	11114	0.08	3000	0.07

Note: Table generated in R 4.0.2 using the xtable package.

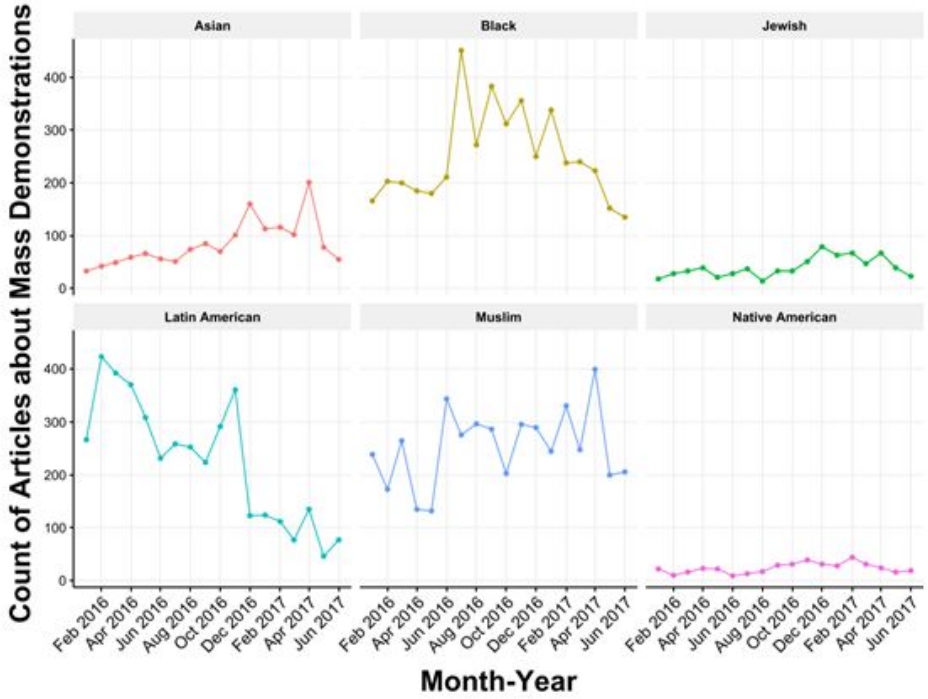


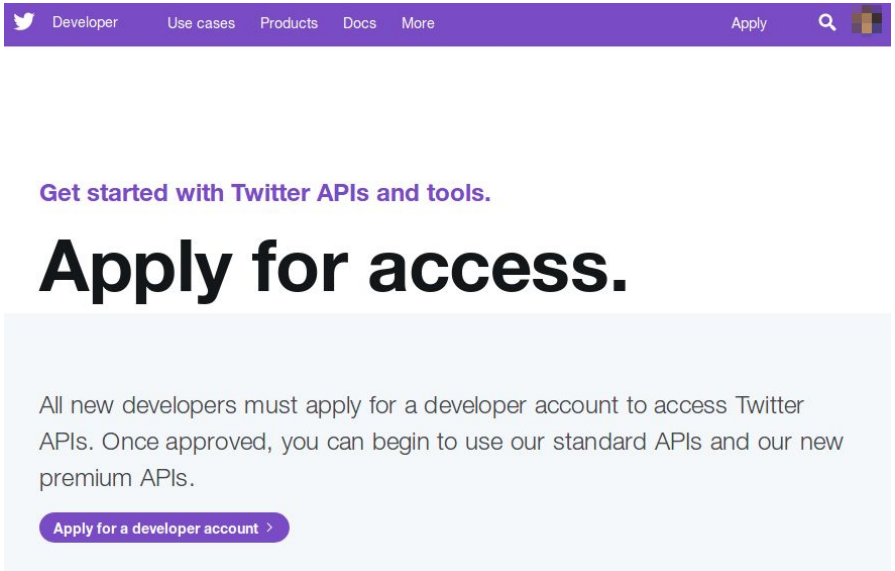
Figure 6: Count of articles about mass demonstrations in US Newspapers from January 2016 to June 2017. Subplots represent counts separated by tagged *AMAR* ethnic group.

How else can we gather text data?

1. Twitter is becoming an increasingly popular venue for discussion about politics for both citizens and political leaders
2. It’s difficult to gather twitter data unless you have a twitter developer account
3. So how do we get this account and gather data?

How do you apply for an academic twitter developer account?

1. Start by applying for a standard developer account
2. Afterwards, you can apply for an upgrade to an academic account
    a. The academic account will get you the ability to pull up to 10 million tweets a month



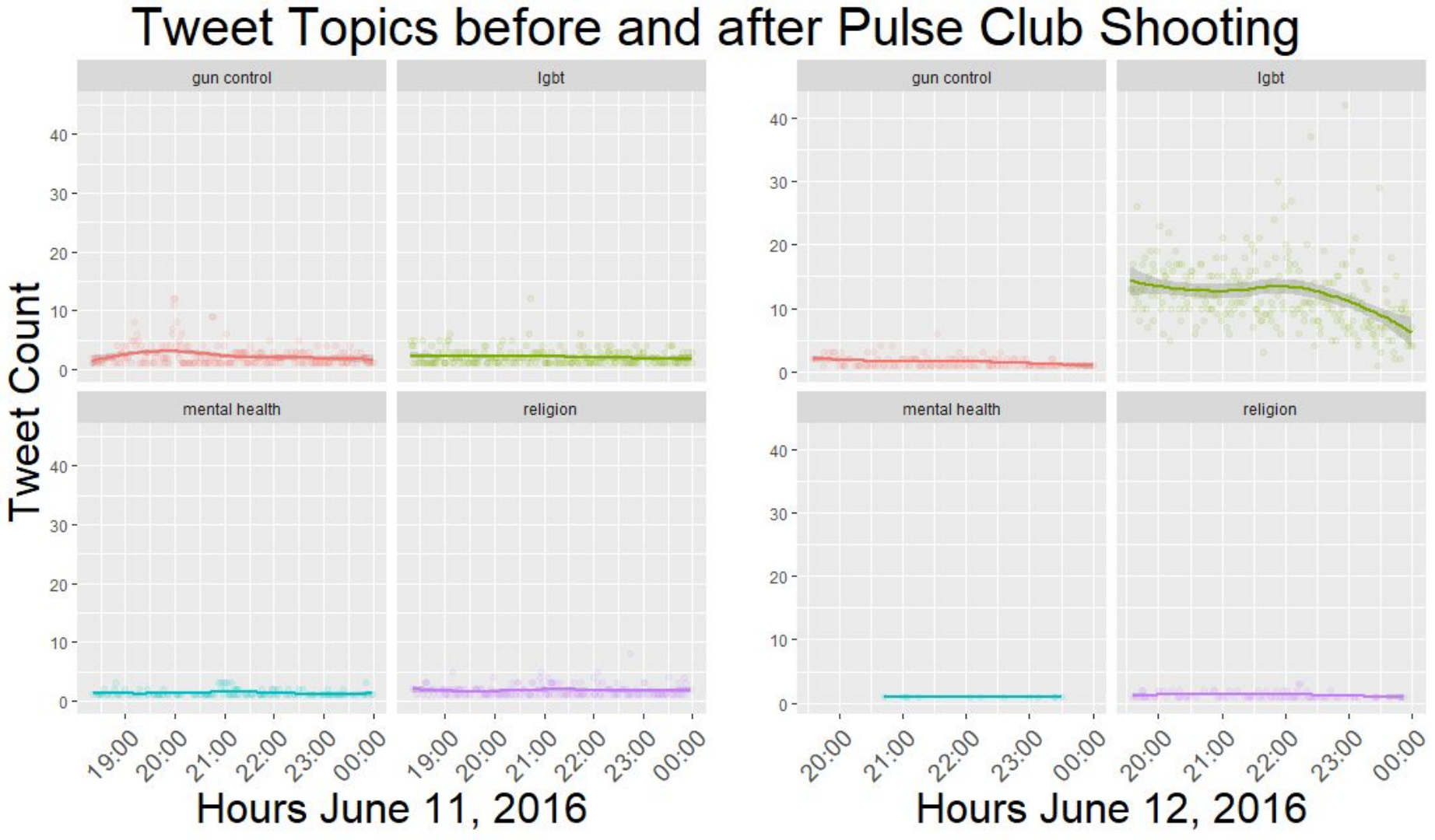
An example of tweet scraping in R

Our application - Pulse Nightclub Shooting and Identity on Twitter

1. We collected about 200,000 tweets from June 11th, 2016 and June 12th, 2016
    a. The model focused on four main topic areas:
        i. Gun control
        ii. Mental health
        iii. LGBT issues
        iv. Religion

Test Dictionary

```
issues:
     lgbt:[lgbt*, gay_community, lesbian*, gay*, bisexual*, transgender*, trans, queer*, womxn]
     religion:[religion, god, allah, tradition, culture war, abomination, radical gender norms,
bible, jesus, quran, psalm, hadith]
     gun control:[second amendment, 2nd amendment, nra, ar-15, wal_mart, bullet, gun registry]
     mental health:[mental_health, mental_illness, suicide_prevention, mentally_ill,
anxiety_and_depression, loneliness_epidemic, dsm, icd]
```


Tweet Topics before and after Pulse Club Shooting
gun control
lgbt
mental health
religion
Tweet Count
0
10
20
30
40
19:00
20:00
21:00
22:00
23:00
00:00
Hours June 11, 2016
Hours June 12, 2016

Thank you!
Questions?