Text Analysis: Day 1

Rachel Porter
Odum Institute for Research in Social Science
rachsur@live.unc.edu

February 18, 2020

Outline

1. Collection
   - Web-scraping
   - Application Programming Interface
2. Processing
3. Analysis
   - Dictonary-based Approaches
   - Topic-based Approaches

Why Text Analysis?

- We recognized that text can convey powerful information
  - Content
  - Sentiment, Tone, or Emotion
  - Patterns and trends
- However, using text as data in research can be expensive...
  - Time consuming to read and code
  - Expensive to hire assistants
  - Methodology can be inconsistent / hard to generalize

For Example. . .

Topo offers a beautiful view of Franklin Street that's perfect for people watching in the afternoon. During the day, it's a casual spot to grab dinner with family or friends and the food is pretty decent tasting but can be pricey for a predominately college aged crowd. The drinks are always pretty good and they have a good selection of beers on tap with a selection of their own vodka that they distill.

During the night, this place turns into your typical college free for all with a DJ, dancefloor, loud music and college kids packed in like sardines. The outdoor patio offers a nice relief from it all when you need a water break from dancing. It's a pretty decent spot, a fun nighttime bar and a staple in Chapel Hill you have to visit at least once just to say that you did.

Corey C. voted for this review

Useful 1 | Funny | Cool

I came here today during their afternoon menu hours, which are 3:00-5:00. The limited menu is definitely smaller than the lunch and dinner menus, but overall had a fair number of choices. Unfortunately, none of them particularly stood out to me, but I know I can be picky. I ended up ordering the Bavarian pretzels with beer cheese, which were okay. The pretzels tasted fine, but were stale and possibly reheated. The cheese had already started to congeal by the time it was brought to the table. It tasted fine, but nothing special. They were quite busy, so the service was alright given the number of people who were there. A great bonus is the location and the patio seating. It's on the top floor of a building right on the corner of downtown Chapel Hill, and the view was beautiful today. They also have covered outdoor seating for those rainy days. I'll have to come back during dinner some time because I've heard that the short ribs are quite good, and I've been wanting to try them. Overall, it's a great spot to hang out. They're known for their bar and beer selection, since they have their own brewery. It does, however, get very crowded with college students on weekends and during events.

Nicole L. voted for this review

Useful 1 | Funny 1 | Cool 1

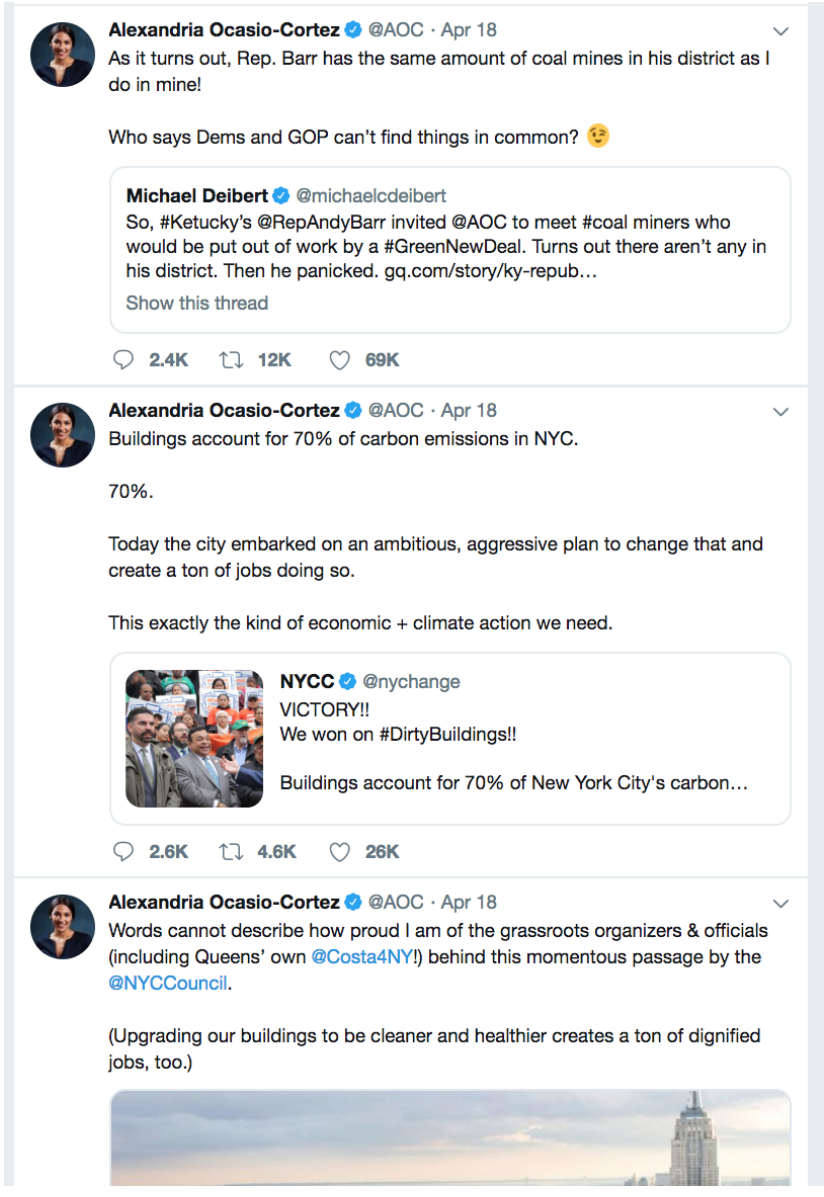
Releated Fields

- **Machine Learning**: A set of statistical and programming tools for extracting patterns for (often) large, complex data sources
- **Statistics**: developing methods for the interpretation of data and experimental outcomes in reaching conclusions with a certain degree of confidence
- **Natural language processing** (NLP): The study of language and how it is used, often within and/or across languages

Some Initial Text Analysis Jargon

- A **corpus** is a collection of documents, think of it as a dataset of text
- **Documents** are the individual collections of text, like a news article or interview transcript
  - Documents can be broken down into smaller pieces, this is called "**parsing**"
  - For instance, in a transcript, we may want to analyze each speaker's text not the document as a whole
- To learn about text, we often employ **metadata**
  - For example: date, author information, auxillary information, etc.

For Example. . .



Alexandria Ocasio-Cortez @AOC · Apr 18

As it turns out, Rep. Barr has the same amount of coal mines in his district as I do in mine!

Who says Dems and GOP can't find things in common? 😉

> **Michael Deibert** @michaelcdeibert
> So, #Ketucky's @RepAndyBarr invited @AOC to meet #coal miners who would be put out of work by a #GreenNewDeal. Turns out there aren't any in his district. Then he panicked. gq.com/story/ky-repub…
> Show this thread

2.4K 12K 69K

Alexandria Ocasio-Cortez @AOC · Apr 18

Buildings account for 70% of carbon emissions in NYC.

70%.

Today the city embarked on an ambitious, aggressive plan to change that and create a ton of jobs doing so.

This exactly the kind of economic + climate action we need.

>
> **NYCC** @nychange
> VICTORY!!
> We won on #DirtyBuildings!!
>
> Buildings account for 70% of New York City's carbon…

2.6K 4.6K 26K

Alexandria Ocasio-Cortez @AOC · Apr 18

Words cannot describe how proud I am of the grassroots organizers & officials (including Queens' own @Costa4NY!) behind this momentous passage by the @NYCCouncil.

(Upgrading our buildings to be cleaner and healthier creates a ton of dignified jobs, too.)

Collecting Text Online: Webscraping

- Web scraping or screenscraping refers to the process of automatically extracting data from web pages. There are three broad steps involved in scraping:
  1. Identify the page or pages with data or text of interest
  2. Download the source code (HTML or XML)
  3. Parse source code and create data set

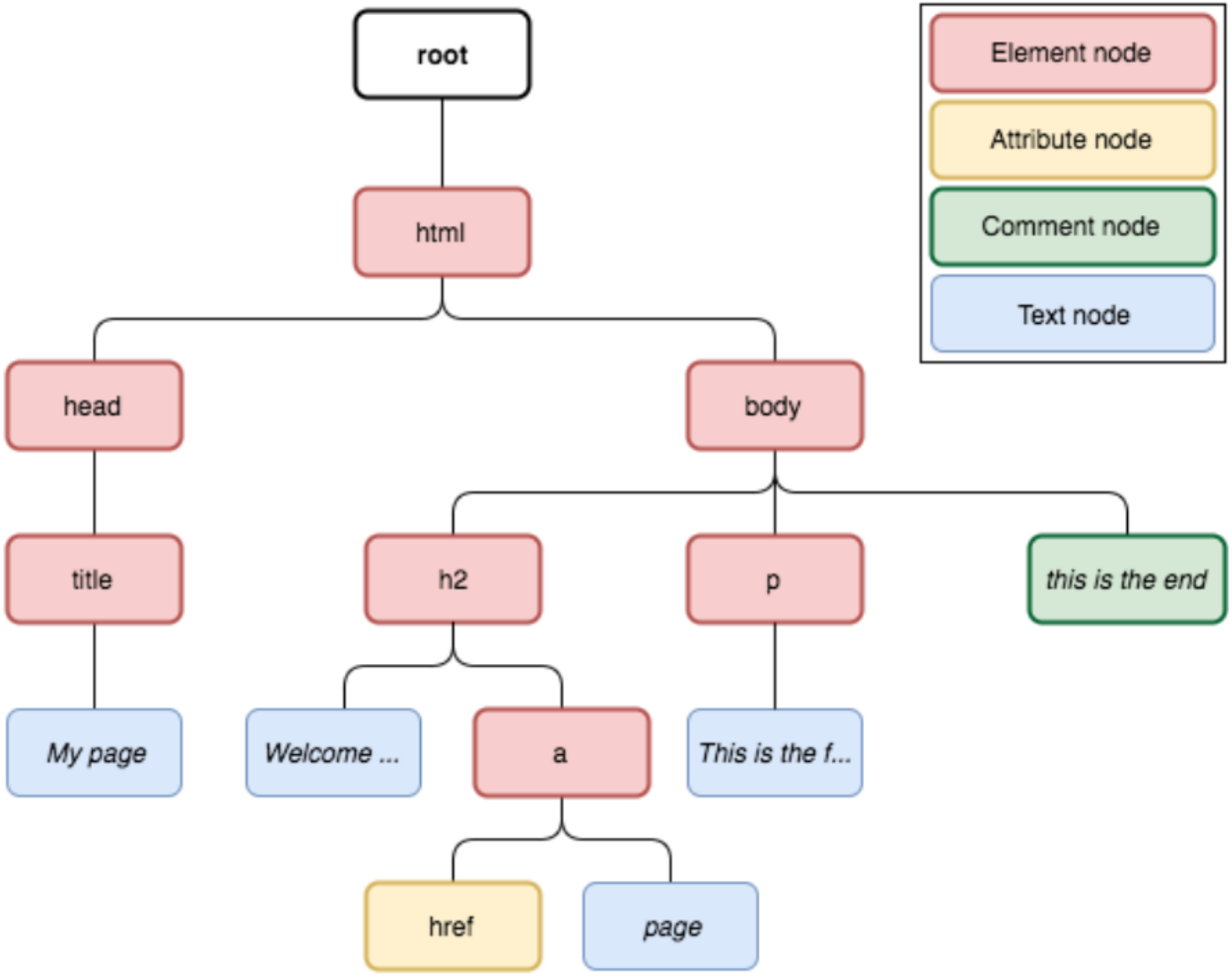
Step 1: Identify the Webpage

- https://politicalscience.unc.edu/people/faculty/

Step 2: Download the source code

```
library(rvest)
faculty_page <- read_html("https://politicalscience.unc.edu
```

Step 3: Parse source code



Step 3: Parse source code

```
bio_nodes <- html_nodes(faculty_page, css = ".entry-title
                                .entry-content")
bio_text <- html_text(bio_nodes)
```

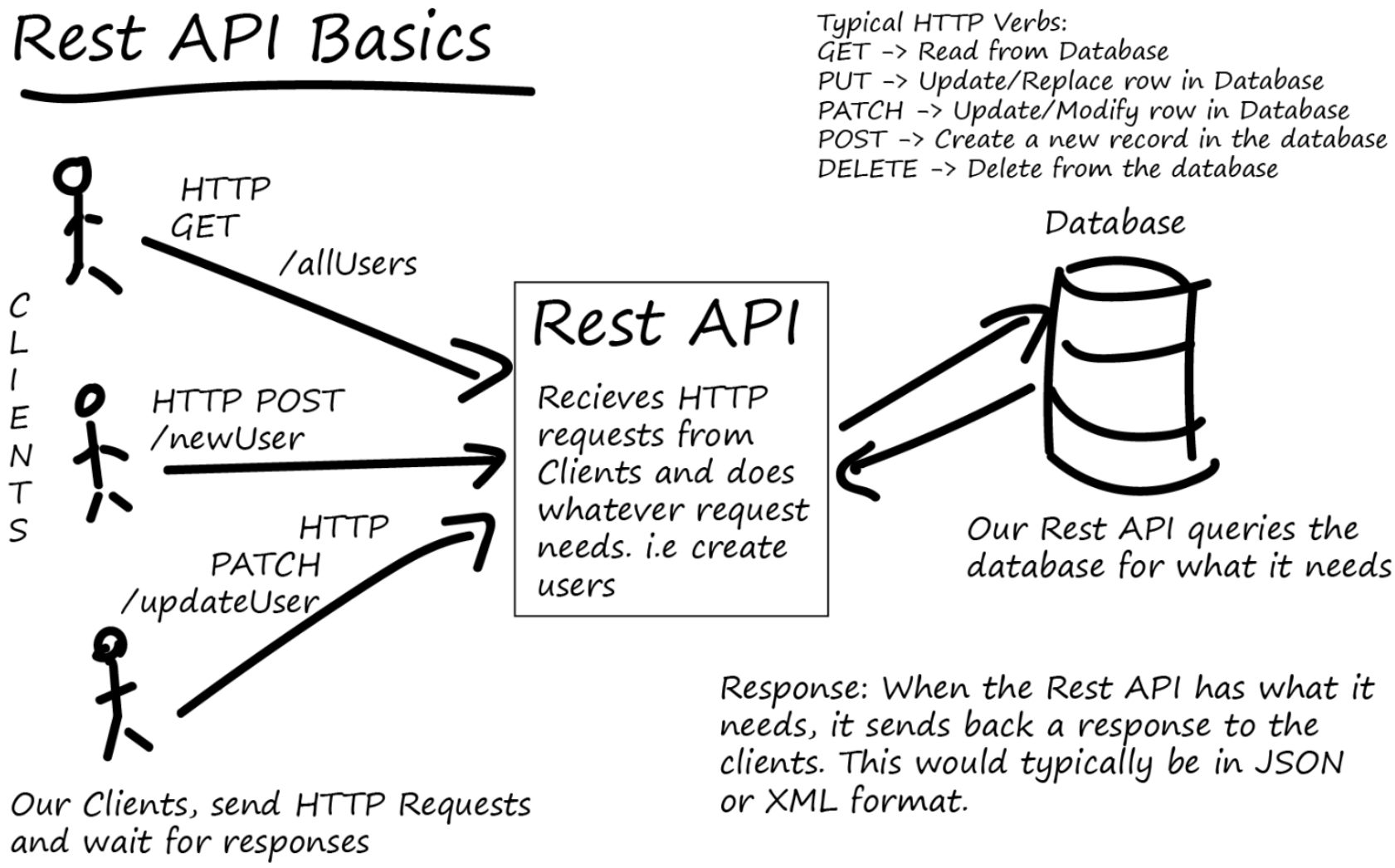
```
"Assistant Profressor317 Hamilton HallOffice Hours:
Fall 2018:  Wed. from 2:00-5:00919.962.0434cambr@email.unc
WebsiteCurriculum VitaeCameron Ballard-Rosa is an Assistant
Professor of Political Science at the University of North C
Chapel Hill.  He received his Ph.D. in Political Science fr
University, and also holds an M.A. in Economics from Yale.
interests include political economy, international relation
comparative politics, and formal theory.Cameron is currentl
a book project on the political logic of international sove
with particular emphasis on the ways that urban-rural confl
sensitive food subsidies, may vary across different regime
He uses formal theory, large-n statistical analysis, and cl
study reading of several countries to present substantive a
evidence for his primary hypotheses explaining sovereign de
```

A note on Webscraping

- Many websites do not allow webscraping, be sure to read the terms of use before scraping text
- For example, Twitter, Facebook, the New York Times, Federal government websites all prohibit scraping
- Alternatively, these types of sites often allow for text to be collected via Automated Programming Interfaces

Twitter (and only pursuant to the applicable terms and conditions), unless you have been specifically allowed to do so in a separate agreement with Twitter (NOTE: crawling the Services is permissible if done in accordance with the provisions of the robots.txt file, however, scraping the Services without the prior consent of Twitter is expressly prohibited); (iv) forge any TCP/IP packet header or any part of the header information in any email or posting, or in any way use the Services to send altered, deceptive or false source-identifying information; or (v) interfere with, or disrupt, (or attempt to do so), the access of any user, host or network, including, without limitation, sending a virus, overloading, flooding, spamming, mail-bombing the Services, or by scripting the creation of Content in such a

Automated Programming Interfaces



Preprocessing Unstructured Text

- To prepare documents for statistical analysis, we apply a series of rules to the documents in order to simplify their representations
- Plain text may have elements incompatible with text analysis
  - Numbers, punctuation, html code, emojis, etc.

For Example. . .

The new monuments are the Birmingham Civil Rights National Monument, Freedom Riders National Monument and Reconstruction Era National Monument.

- Birmingham Civil Rights National Monument: The Birmingham Civil Rights National Monument will protect the historic A.G. Gaston Motel in Birmingham, Alabama, which served at one point as the headquarters for the civil rights campaign led by Dr. Martin Luther King Jr. that helped lead to the passage of the Civil Rights Act of 1964. The monument will also tell the stories associated with other nearby Birmingham historic sites, including the Sixteenth Street Baptist Church– which was the site of a bombing in 1963; and Kelly Ingram Park, where Birmingham Public Safety Commissioner Bull Connor turned hoses and dogs on young civil rights protesters.
- Freedom Riders National Monument: The Freedom Riders National Monument is located in Anniston, Alabama and contains two sites that help underscore. . .

Preprocessing Jargon

- **Types** are unique sequences of letters or characters, i.e., words, phrases, numbers, and/or punctuation
- **Tokens** are 'particular instances of a type'
- **Terms** are types that occur in a dictionary
- **Vocabularies** are the set of types or tokens that occur in a corpus
- **Vocabulary size** is the number of unique types/tokens in a corpus

Preprocessing Unstructured Text

1. Stop words
   - Common words that aren't meaningful: for instance, 'and', 'or', 'it', 'be'
2. Numbers, punctuation, captialization
   - These can be considered tokens but increase the dimensionality of the text
3. Stemming
   - Bringing words to their root, once again reduces dimensionality of text
   - Combines common words

What do we mean by text dimensionality?

features concept

instances

w_1	w_2	w_3	...	w_n	label
1	1	0	...	0	health
0	0	0	...	0	other
0	0	0	...	0	other
0	1	0	...	1	other
⋮	⋮	⋮	...	0	⋮
1	0	0	...	1	health

Preprocessing Unstructured Text

1. Stop words
   - Common words that aren't meaningful: for instance, 'and', 'or', 'it', 'be'
2. Numbers, punctuation, captialization
   - These can be considered tokens but increase the dimensionality of the text
3. Stemming
   - Bringing words to their root, once again reduces dimensionality of text
   - Combines common words

Preprocessing Unstructured Text

4. Tokenizing
   - Split text into meaningful pieces
     - "Hello Mr. Smith, how are you doing today? The weather is great, and city is awesome. The sky is pinkish-blue. You shouldn't eat cardboard"
     - ['Hello Mr. Smith, how are you doing today?', 'The weather is great, and city is awesome.', 'The sky is pinkish-blue.', "You shouldn't eat cardboard"]
5. Reducing vocabulary size
   - Removing sparsely occuring tokens, reducing dimensionality of text

For example. . .

"I will be the greatest jobs president God ever created."

↓

```
'i' 'will' 'be' 'the' 'great-' 'job-' 'presid-' 'god' 'ever'
                                'creat-'
```

Another example. . .

"Our children can achieve great things."

↓

```
'our' 'child-' 'can' 'achiev-' 'great-' 'things'
```

In dataset form...

	great-	job-	presid-	god	ever	creat-	child-	achiev-
Trump	1	1	1	1	1	0	0	0
Bush	1	0	0	0	0	0	1	1

Lab Exercise

Thanks to Brice Acree, Jaime Arguello, and Chris Bail with the Summer Institute in Computational Social Science for making their materials available. Portions of this presentation draw on their resources.