ABOUT SCRAPY

DATA PROCESSING

DIGEST

NEXT STEPS

EXPLORING CONGRESS

Lora Johns & Eric Ma

Digest

Why Examine Government Data?

- Mr. Speaker!
  - Newt Gingrich when he was first elected in the 90's, would present in front of the always on camera, after session was over in front of an empty chamber, uninterrupted But he would always point the finger at the chamber itself. "Congress was broken. Your political system is broken."
  - Newt is largely attributed for starting the partisan polarized era that persists to this day.
  - Newt was exceptionally media savvy for his era. Now we have more than C-Span. We have a news network for every kind of viewer, twitter, and facebook.

Digest

Why Examine Government Data?

- The Floor Recognizes
  - Is congress working for you? Is what they say to the public the same as what they're doing in their chambers?

Tools for staying on top of Congress

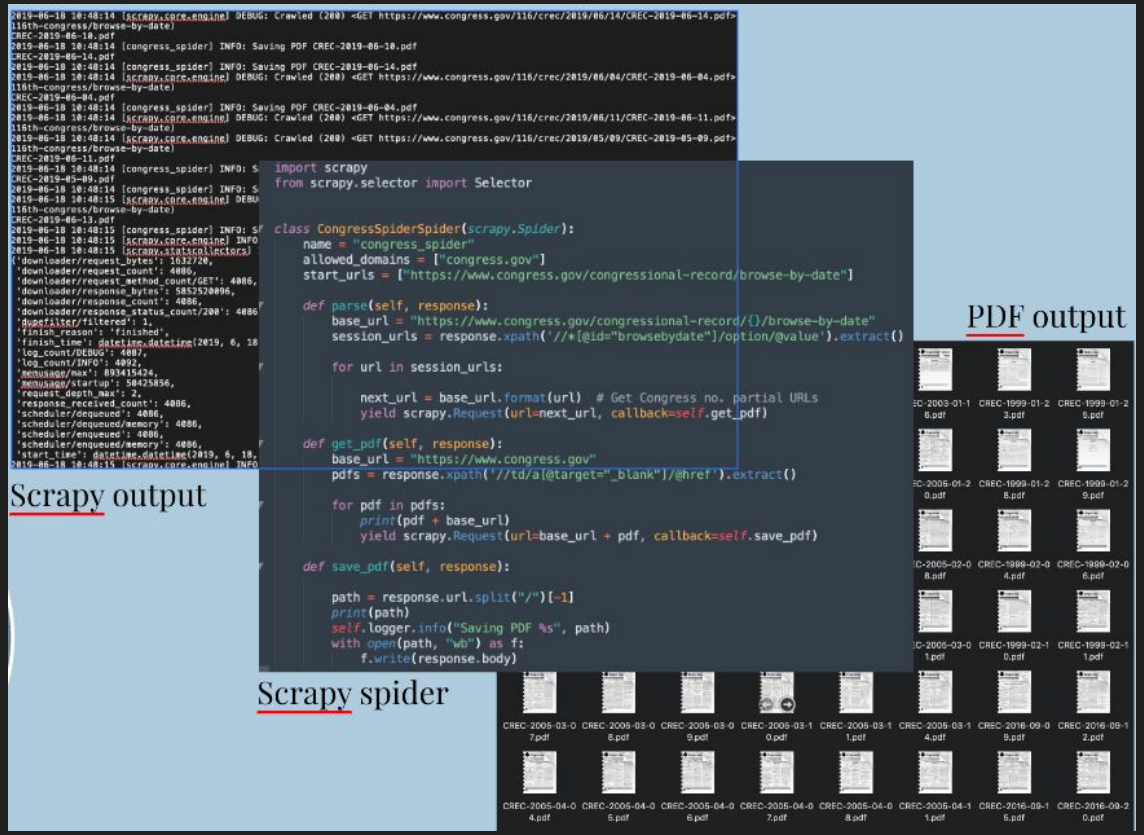
Scrapy

An open source framework for extracting the data you need from websites.

- Government Datasets
  - http://catalog.data.gov/
- Scrapy handles both navigation and extraction, complete solution
- BeautifulSoup only handles scraping
- Selenium is sometimes slower and more memory-intensive

Tools for staying on top of Congress

Scrapy



Tools for staying on top of Congress

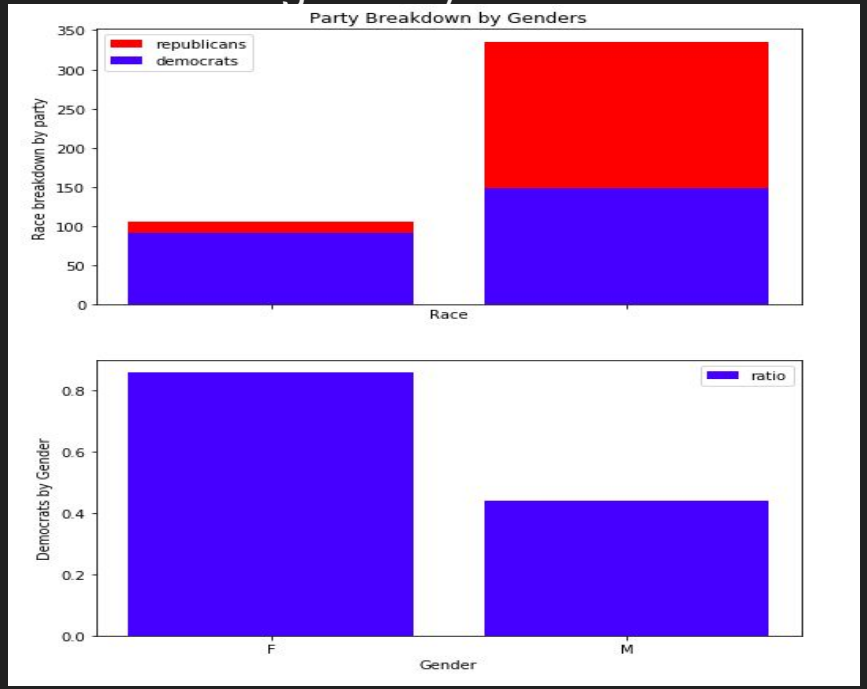
Tika

- APACHE Tika detects and extracts text and metatext from files
- Text extraction
- PDF -> txt

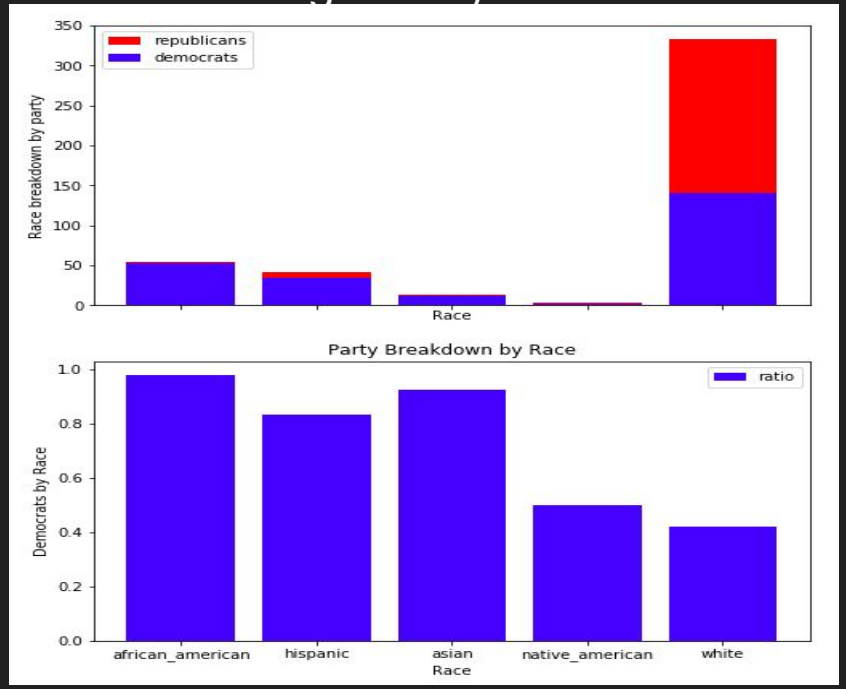
Congressional Demographics

Using API's we can visualize the Demographics of Congress

Congress by Gender



Congress by Race



Future Analysis

- **What bills will pass?**
- **Topic frequency, bill names, trends in policy**
- **How do sentiments change over the course of a career?**
- **Are there differences by gender? By race?**
- **Does sentiment on Twitter differ from sentiment on the floor?**
- **How does VADER compare with IFILL, SpaCy, etc.?**