





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



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

- Government Datasets
 - o 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



```
019-86-18 10:48:14 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.congress.gov/116/crec/2019/06/14/CREC-2019-06-14.pdf
 116th-congress/browse-by-date)
CREC-2019-06-10.pdf
 2019-86-18 10:48:14 [congress_spider] INFO: Saving PDF CREC-2019-86-10.pdf
 CREC-2819-86-14.pdf
 2819-86-18 18:48:14 [congress spider] INFO: Saving PDF CREC-2819-86-14.pdf
 019-06-18 10:48:14 [acrapy.care.engine] DEBUG: Crawled (200) <GET https://www.congress.gov/116/crec/2019/06/04/CREC-2019-06-04.pdf>
 2019-86-18 10:46:14 [congress_spider] INFO: Saving PDF CREC-2019-86-04.pdf
2019-86-18 10:40:14 [<u>scrapv.cpre.enoing</u>] DEBUG: Crawled (200) «GET https://www.congress.gov/116/crec/2019/86/11/CREC-2019-86-11.pdf
]lith-congress/foruse-by-delta
  819-86-18 10:48:14 [scrapy.spre.engine] DEBUG: Crawled (200) «GET https://www.congress.gov/116/crec/2019/85/09/CREC-2019-85-09.pdf>
 REC-2019-05-09.pdf
                                                    from scrapy.selector import Selector
2019-06-18 10:48:14 [congress_spider] INFO: S
2019-06-18 10:48:15 [SCCRPY.COCK.ROGINE] DEBU
116th-congress/browse-by-date)
IREC-2019-86-13.pdf
 peis-e6-10 le:48:15 (congress_spider) INFO: Sr ctass CongressSpiderSpider(scrapy.Spider):
2019-86-10 le:48:15 (scrapy.Spider) and e = "congress_spider"

name = "congress_spider"
                                                         allowed_domains = ["congress.gov"]
  'downloader/request_bytes': 1632720,
'downloader/request_count': 4086,
                                                         start_urls = ["https://www.congress.gov/congressional-record/browse-by-date"]
  'downloader/request method_count/GET': 4886,
'downloader/response_bytes': 5852520896,
'downloader/response_count': 4886,
   downloader/response_status_count/200': 4886
                                                                                                                                                                                           PDF output
  'dupefilter/filtered': 1,
'finish_reason': 'finished',
                                                               base url = "https://www.congress.gov/congressional-record/()/browse-by-date"
                                                               session_urls = response.xpath('//*[@id="browsebydate"]/option/@value').extract()
  'finish_time': datetime.datetime(2019, 6, 18
  'log_count/DEBUG': 4887,
'log_count/INFO': 4892,
'xenusage/max': 893415424,
                                                               for url in session urls:
   menusage/startup': 58425856,
   request_depth_max': 2,
                                                                     next url = base url.format(url) # Get Congress no. partial URLs
  'response received count': 4086,
                                                                    vield scrapy.Request(url=next url, callback=self.get pdf)
   scheduler/dequeued': 4086,
   scheduler/dequeued/memory': 4886,
  'scheduler/enqueued': 4086,
 'scheduler/enqueued/memory': 4886,
'start_time': datetime.datetime(2019, 6, 18, 2019-86-18 10:48:15 [scrapy.core.engine] INFO
                                                         def get_pdf(self, response):
                                                               base_url = "https://www.congress.gov"
                                                               pdfs = response.xpath('//td/a[@target="_blank"]/@href').extract()
                                                                                                                                                                                  C-2005-01-2 CREC-1999-01-2 CREC-1999-01-2
Scrapy output
                                                               for pdf in pdfs:
                                                                    print(pdf + base_url)
                                                                    yield scrapy.Request(url=base_url + pdf, callback=self.save_pdf)
                                                         def save pdf(self, response):
                                                               path = response.url.split("/")[-1]
                                                               self.logger.info("Saving PDF %s", path)
                                                               with open(path, "wb") as f:
                                                                     f.write(response.body)
                                                Scrapy spider
                                                                                                  CREC-2005-04-0 CREC-2005-04-0 CREC-2005-04-0 CREC-2005-04-0 CREC-2005-04-0 CREC-2005-04-1 CREC-2016-09-1 CREC-2016-09-2
```

Tools for staying on top of Congress

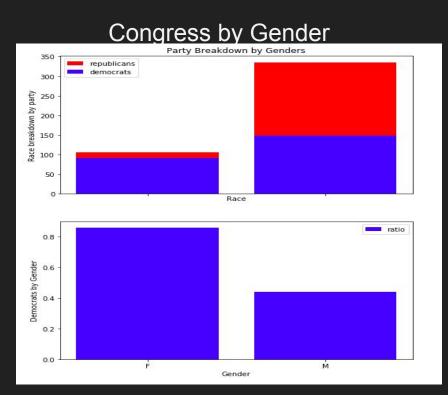
Tika

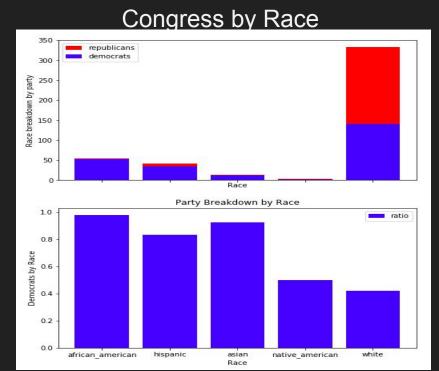
- APACHE Tika detects and extracts text and metatext from files
- Text extraction
- PDF -> txt



Congressional Demographics

Using API's we can visualize the Demographics of Congress





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.?