**Analysis of United States Politicians Biodata and Media Favorability**

Ezra Cohen, Suma Dendi, Hawa Iman, Diba Demir

CodeBozu Fellowship

December 2021 - January 2022

**Ⅰ. Introduction**

The CodeBozu Computer Science Fellowship has been an extraordinary learning experience. Under a curriculum developed by Cornell University students, we practiced programming fundamentals, created databases and experimented with various methods of both data manipulation and visualization, and finally learned multiple python libraries, including Beautiful Soup, Selenium, and VADER. We submitted a total of five deliverables to the fellowship, ranging from a report on the ethics of web scraping to insights on data analysis. We worked collaboratively on Google Notebook and had weekly meetings to discuss updates, issues, and delegating tasks. This report will detail our work for the fellowship, the next steps for our project, and how we intend to share it publicly to benefit others.

**Ⅱ. Research on Web Scraping**

Before coding, we researched the ethics of web scraping to develop team guidelines on how to best approach the aspects of the project that required scraping. After reaching the consensus that web scraping is an incredibly powerful tool but can be malicious when misused, the following guidelines were developed:

1. For ethical collection of data:
   1. Data must be publicly available and not copyrighted
      1. To scrape data that is copyrighted, requires purchase, or logging into an account, such as academic journals, agreements should be made with the publisher or website owners
   2. Find the website’s robots.txt file and terms and conditions to learn how the website should be crawled and indexed through the format <domain\_url>/robots.txt
   3. Request permission from the owners of the website to collect large amounts of data or include a user-agent string that tells the owner what we are doing and how to contact us
   4. All original publishers and owners should be credited for their data
2. To mitigate repeatedly querying the server:
   1. Check if the website has an API that allows for easy access of the data without scraping
   2. Recursive scrapers should be tested first on a smaller dataset to ensure that they don’t send unnecessary queries
   3. Publishing data and findings publicly in a way that is easy for others to use, such as an API, so that others can utilize the data without scraping

**Ⅲ. Creation of a Database of United States Politicians Biodata**

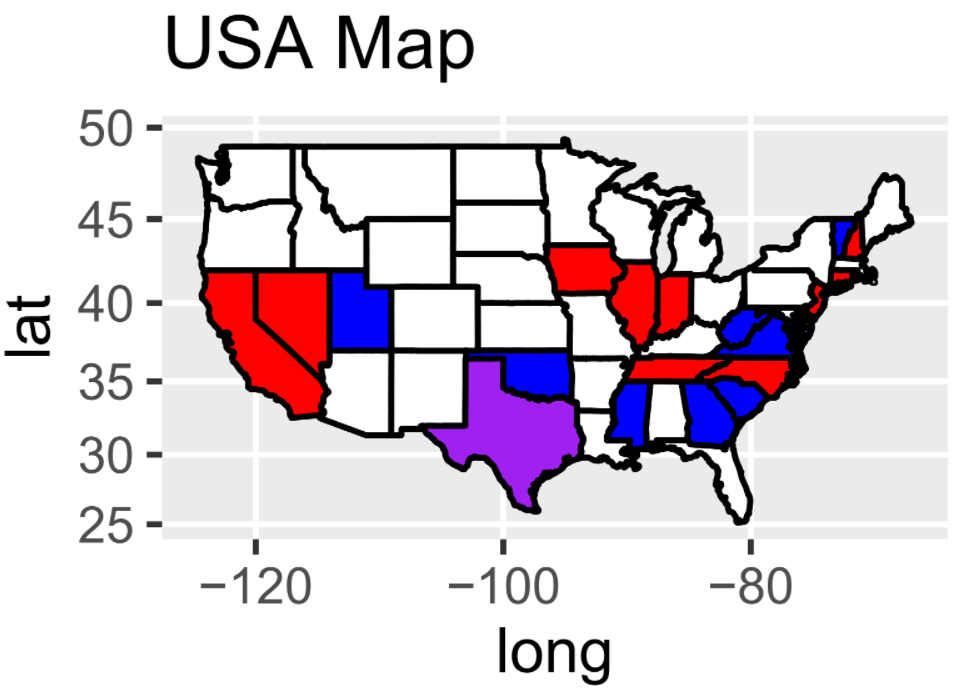
For the first part of the fellowship, we familiarized ourselves with web scraping by creating a database of politician biodata from Wikipedia.

**Procedure**

This procedure entailed two rounds of scraping. First, we scraped the Wikipedia pages 21st-century presidents of the United States 1, 21st-century vice presidents of the United States 2, 20th-century presidents of the United States 3, 20th-century vice presidents of the United States 4, and 19th-century vice presidents of the United States 5, to find all the links to web pages of each individual politician, and add their links to a list. Next, we looped through that list using a function that scraped biodata consisting of politician name, full name, birthday, birthplace, party, other professions, and type of office held, from each politician’s webpage. To do this, we used the Beautiful Soup 4 Python library and the Xpath method from Selenium to find the tags that contained the desired data and append the text within the tags to a dictionary. Some Wikipedia pages, however, lacked consistency with the location of certain data under certain tags. For example, most pages listed other professions of the politician in a tag named “Occupations,” while some used to text “Professions,” and others didn’t list other professions of the politician at all. To solve this issue, we created a series of if statements to account for inconsistencies in the tags of biodata that we noticed in certain politician webpages. After these fixes, we obtained a complete database of 92 politicians and their biodata organized in a dictionary. To better visualize our database, we converted it into a data frame using the Pandas Python library and downloaded it as a [CSV file](https://drive.google.com/file/d/1E7SzJSYg3bhe66xBSzhZiBROPxzJ37Mi/view?usp=sharing).

**Ⅳ. Insights from United States Politician Biodata**

We were able to extract and visualize a number of insights from our database of US politicians biodata.

1. Visualization of the primary political party of politicians born from each US state. For example, the state of California is red because it had a greater number of Republican presidents and vice presidents from our database born there, while the state of Maine is white because it had no politicians from our database born there.

The map is white in places where there were no presidents or vice presidents born from, blue in places where there were more Democrats who were born there, red in areas that had more Republicans, and purple in areas that elected an equal number of Democrats and Republicans. There were other colors like yellow and orange for if there was an equal amount of people who were in other politician parties elected as the number of either Republicans or democrats, black if there was ever an equal amount of any which there wasn't, and green if there were more people who were from other parties that were elected than either democrat or republican. Our code also included cyan to show up if any mistakes were made in the program, and thankfully there was no cyan anywhere on the map.

The word clouds below visualize the professions, in order of frequency, of Presidents in comparison to Vice Presidents from our database.

1. The five most common professions of American presidents were “lawyer,” “author,” “businessman,” “Military officer,” and “soldier.”

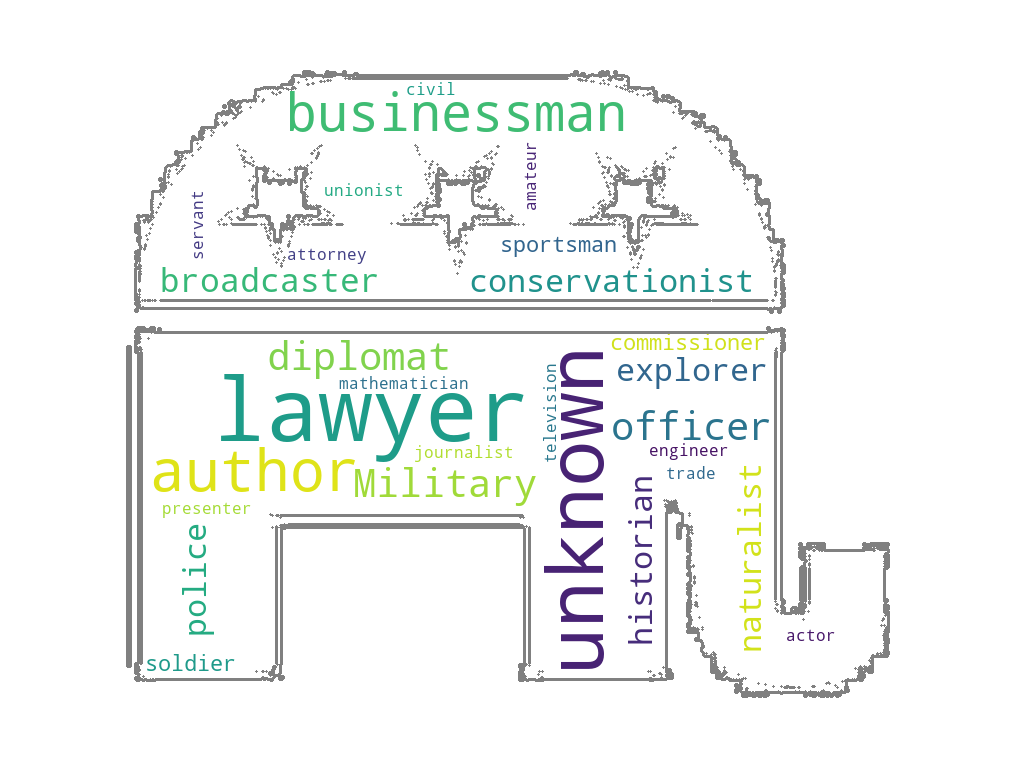


1. The five most common professions for American vice presidents (excluding the unknowns) were “lawyer,” “author,” “environmentalist,” “businessman,” and “journalist.”

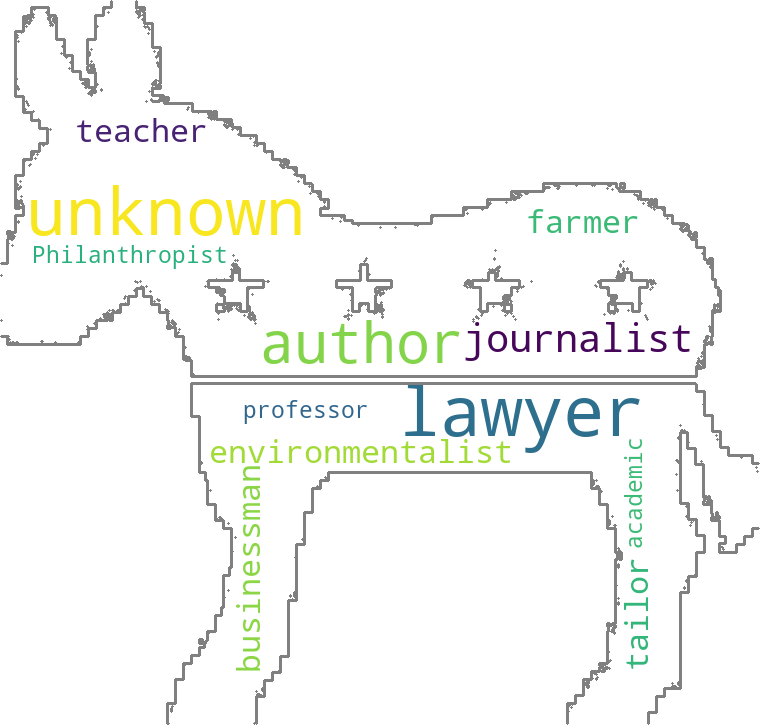


The word clouds below visualize the professions, in order of frequency, from politicians in the Republican and Democratic party in our database.

1. The top 5 most common occupations for politicians from the Republican party are “lawyer,” “author,” “businessman,” “diplomat,” and “military officer.”



1. The top 5 most common occupations for politicians from the Democratic party are “lawyer,” “author,” “journalist,” “teacher,” and “farmer.”



1. Most politicians from our database in the republican party were presidents, and most politicians from our database in the democratic party were also presidents.

**Ⅴ. Ranking Media Favorability of Presidents**

For the second part of the fellowship, we focused on scraping news articles and using sentiment analysis algorithms to determine the favorability of certain news sources towards different American Presidents.

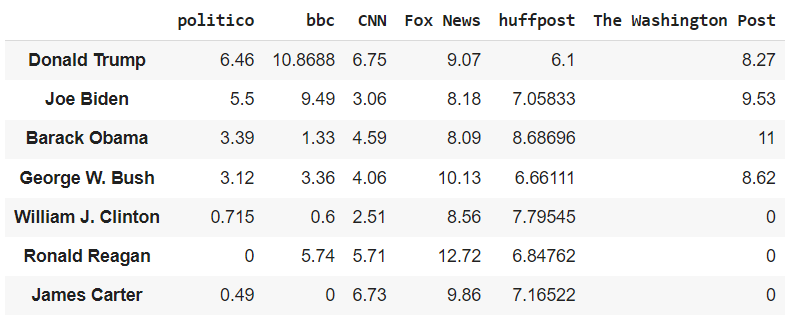
**Procedure**

The news websites that we analyzed were randomly selected from the “major news sources” table on the News Media in the United States Wikipedia page 6 and we got Politico, BBC, CNN, Fox News, HuffPost, and The Washington Post. The presidents that we tested were the eight most recent US presidents, excluding George H.W. Bush, for the purposes of avoiding confusion between George H.W. Bush and George W. Bush in our web scraping and sentiment analysis. Therefore, our final list of presidents to generate media favorability ratings consisted of Joe Biden, Donald Trump, Barack Obama, George W. Bush, Bill Clinton, Ronald Reagan, and James Carter. Next, to create a random, unbiased sample of articles to analyze from each media website, we searched for the president on each website and selected all of the articles on the first pages of the search results of each website. This also ensured a broad range of articles per media source so their overall score wouldn’t be biased by outliers, but some samples of certain sites had more articles to scrape than others, which could’ve led to some of our favorability ratings having a lower margin or error than others.

After scraping the article links from each website, the links were loaded into a for loop that ran sentiment analysis. We used the Selenium Webdriver to collect the dynamically loaded information from each link and search for the tags that contained the text that we would scrape. We then ran VADER sentiment analysis on each article and added its score to the news website’s mean. Some links had primarily non-text content like videos, but we were still able to gather captions and other small pieces of text to analyze and add to the mean score.

**Favorability Rating Results**

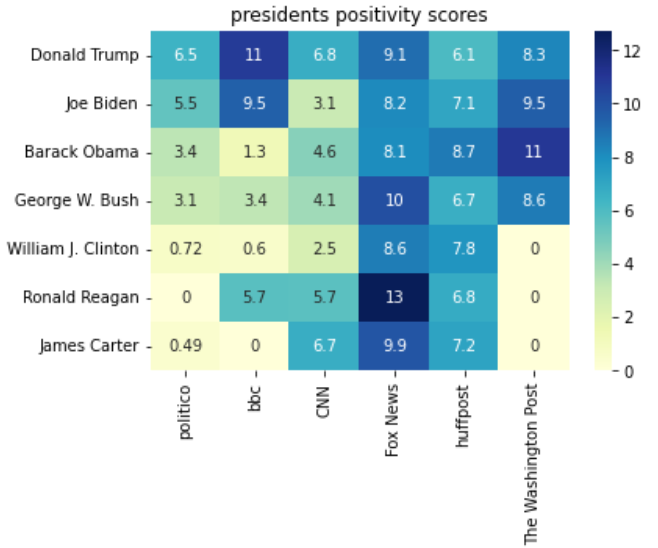
After obtaining favorability ratings for each president through each news websites, we organized the result into a data frame using the Pandas Python library:



The ratings were calculated on a scale of 0-13 with 13 being the most positive, and 0 being not positive. The implications of this scale will be analyzed below in “Potential Shortcomings in our Analyzer.” Of the ratings of 0 in our results, three of them, specifically the ones under The Washington Post, were generated due to our analyzer being unable to analyze their text which was dynamically loaded in with JavaScript, even despite our efforts with selenium. The other the ratings of 0, specifically the one for Ronald Reagan with Politico and James Carter with BBC, were naturally generated, meaning there was most likely very little content for those presidents with that news media, so the little text they had was primarily neutral. This makes sense considering both Ronald Reagan and James Carter were older presidents in terms of recency.

**Heatmap Visualization of Favorability Rating Results**

Next, for better visualization of our results, we organized the ratings into a heatmap using the Seaborn Python library with the x-axis containing the media source and the y-axis containing the presidents:

****

**Potential Shortcomings in our Analyzer**

While we scraped all the articles on the front page of each randomly selected website, the samples for different presidents may have contained an unequal number of outliers, resulting in some boxes in our heatmap being more accurate than others. Also, as we mentioned before, some articles were extremely brief or only contained captions belonging to a video. These most likely had a positivity score that was neutral, which would slightly affect the mean of the news website. Additionally, our VADER analyzer had a few shortcomings, including not being able to recognize double negatives or identifying the subject of the sentence to which a positive or negative phrase is attached, which may have skewed the favorability ratings of certain presidents. Solutions for these shortcomings will be detailed in Ⅶ. The Future of Our Project.

**Ⅵ. Insights from Ranking of Media Media Favorability of Presidents**

1. According to our analysis, each president had the following average media positivity score rounded to the nearest tenth place:
   1. Donald Trump: 8.0
   2. Joe Biden: 7.2
   3. Barack Obama: 6.2
   4. George W. Bush: 6.0
   5. William J. Clinton: 4.0
   6. Ronald Reagan: 6.2
   7. James Carter: 4.9
2. Our results show that the selected media sources have the following average positivity scores towards the Republican and Democratic political parties with our sample of presidents:

|  | **Republican Presidents** (Donald Trump, George W. Bush, Ronald Reagan) | **Democratic Presidents (Joe** Biden, Barack Obama, William J. Clinton, James Carter) |
| --- | --- | --- |
| **Politico** | 3.2 | 2.5 |
| **BBC** | 6.7 | 2.8 |
| **CNN** | 5.5 | 4.2 |
| **Fox News** | 10.7 | 8.7 |
| **HuffPost** | 6.5 | 7.7 |
| **The Washington Post** | 8.5 | 10.3 |

Out of the six media sources presented, Fox News had the highest positivity score of 10.7 out of 13 for presidents in the Republican party, while The Washington Post had the highest positivity of 10.3 out of 13 for presidents in the Democratic party. Politico was the least politically biased, with positivity scores for both political parties closest to neutral, a net difference between the positivity scores of the two political parties of only 0.7 out of 13.

1. Fox News had the highest favorability of presidents overall, with a positivity score across all seven presidents of 9.6 out of 13.

**Ⅶ. The Future of Our Project**

The next steps for this project would be to improve our media favorability ratings by eliminating more biases and fixing shortcomings. First, to decrease variability in the samples of articles for each president with different news websites, we would use a larger sample size for all websites and make the sample size the same across all presidents and websites. Next, the Natural Language Toolkit can be integrated into our code to prevent our VADER analyzer from misinterpreting the sentiment of a sentence in our scraped articles due to confusion among technicalities in English grammar such as subject and object, or the passive and active voice. For example, if something negative is said in a sentence about a different politician in an article about Joe Biden, Natural Language Toolkit would correctly identify that the negative phrase is directed to the other politician, and not Joe Biden.

After improving the accuracy of our ratings of media favorability towards various US presidents, we will share our findings publicly. We intend to publish this report, our deliverables, and full code to a GitHub repository for others to review, learn from, and utilize our findings. We hope our project inspires others, especially high schoolers, to recognize the power of programming and technology to solve problems, gather insights about the world we live in, and share knowledge with humanity.

**Ⅷ. References**

1. <https://en.wikipedia.org/wiki/Category:21st-century_presidents_of_the_United_States>

2. <https://en.wikipedia.org/wiki/Category:21st-century_vice_presidents_of_the_United_States>

3. <https://en.wikipedia.org/wiki/Category:20th-century_presidents_of_the_United_States>

4. <https://en.wikipedia.org/wiki/Category:20th-century_vice_presidents_of_the_United_States>

5. <https://en.wikipedia.org/wiki/Category:19th-century_vice_presidents_of_the_United_States>

6. <https://en.wikipedia.org/wiki/News_media_in_the_United_States>

**Ⅸ. Tools/Frameworks**

1. Python 3
2. Google Colaboratory (collaboration)
3. BeautifulSoup 4 (scraping data from HTML and XLM files)
4. Pandas (data analysis)
5. Mathplotlib (plotting and data visualization)
6. Numpy (mathematical operations on arrays)
7. VADER (text sentiment analysis)
8. Selenium (automating browser test cases)
9. Kora (tools for using libraries in Google Collaboratory)
10. LXLM (for Xpath locator)
11. PIL (opening, manipulating, and saving image file formats)
12. os (creating, removing, fetching, and manipulating a directory)