CONGRESSIONAL VOTE CLUSTERING
Foundations of Data Science and Analytics (DSCI 608)

Le Nguyen (383006341) - ln8378@g.rit.edu
Greeshma Ganji (362007736)(gg1849@rit.edu)
Nandhini Lakshman (nl7222@g.rit.edu)
Tolulope Olatunbosun (tao5634@g.rit.edu)

Problem Statement

In this project, we wanted to see if the current political divide in American politics shows up in voting record data, and furthermore see if we could build a model to classify which party a member of congress belonged to based on their voting record. We suspected that the vast ideological differences between the parties would be very apparent and show up as significant differences in voting records. From this suspicion, we assume the congressional data can be clustered into groups representing each party.

Research & Literature

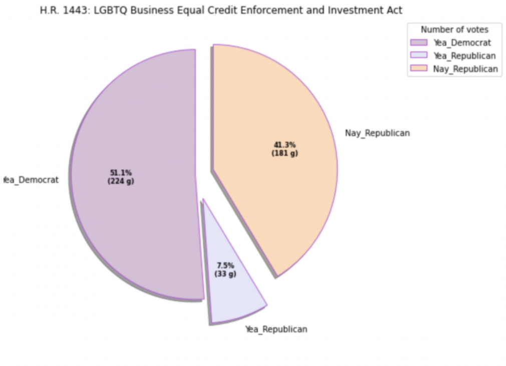
For this project, we looked at past analyses of this problem. Past work has primarily been done on clustering data from senators, while we clustered data from members of the house (govtrack, 2021)(Swallow, 2017). We can see from historical data, that the senate has been becoming more and more divided over time (Swallow, 2017). We suspect the house of representatives will follow the same pattern and be divided into distinct clusters along party lines. We can use some sort of clustering method to do this.

We had to find a proper model to do this and further a distance metric to define distances between congress people to cluster then. We found the Jaccard distance to be the most suitable distance metric to use as it is the distance between two sets (the sets of voting records) (Glen, 2016). Also through our research, we found the most appropriate clustering method to use was hierarchical clustering (Sharma, 2019).

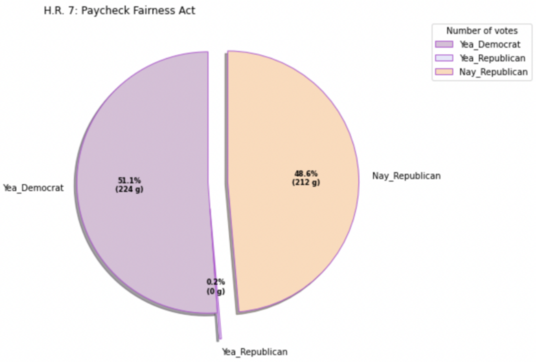
Analysis

Two lists were created for the number of times a Democrat and a Republican voted against their own party. The mean count was computed for each list so we could get a better picture of the average number of outliers. From the results, we observed a higher number of republicans who voted against their party compared to the number of democrats. We wanted to visualize the outliers in the form of pie charts for certain bills.

Upon visualization, for certain bills, we noticed a significantly higher number of Republicans deviating from typical Republican voting behaviour. The portions of the pie charts were much smaller for the Democrats who voted against their parties.
Generally, the parties seemed to be very divided in their voting behaviour.

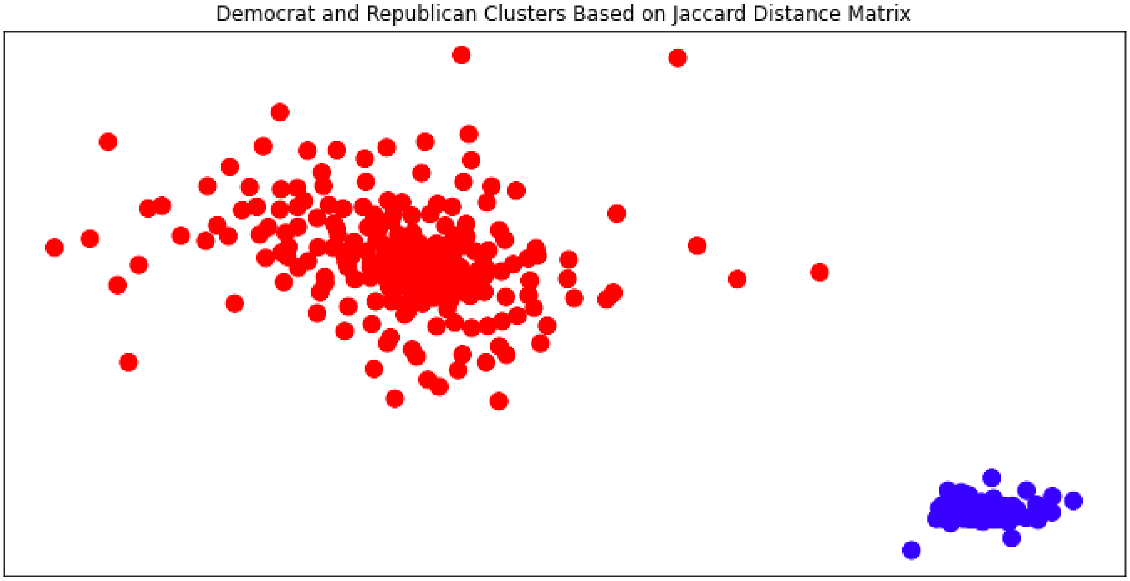


H.R. 1443: LGBTQ Business Equal Credit Enforcement and Investment Act



H.R. 7: Paycheck Fairness Act

After computing the Jaccard distance, we discovered relational distances between each politician with respect to their voting records. Smaller distances demonstrated similarity between voting choices across various bills as well as similar beliefs, being members of similar parties, or perhaps collusion [Scipy, 2021]. On the other hand, greater distance metrics represent distinct differences in voting preferences. The figure below illustrates each individual senator with respect to other senators. We found that members of the Democratic party generally vote with their party, and often with little variance. Contrarily, Republican senators represented varied voting preferences, though, distinct enough to stay within their respective clusters.



Democrat and Republican Clusters Based on Jaccard Distance Matrix

Working Plan

Le - Web scraping and cleaning the data for analysis and model development. Did some data analysis on the side as well.

Nandhini - Performed data visualization and data analysis on controversial bills to inform voting preferences and to detect anomalies in either party.

Tolu - Performed data analysis, and implemented Jaccard distance metric on the data to map preferences between politicians voting behaviors.

Greeshma -

Teamwork & Collaboration

In terms of teamwork and collaboration we did struggle a bit to have consistent meetings because we were all busy with our classes and had conflicting schedules. After the first few meetings to get the project idea set up we did not meet again for about a month. We had to start frequent meetings again when the project due date came near. This was a problem when it came to giving weekly updates because we were not consistently meeting to generate new work and have everyone else informed. Everything did come together in the end though and worked out.

In hindsight having more frequent meetings would have made the project a more comfortable process, but it is understandable that we got caught up in other work and classes. Perhaps having online meetings would have served us better with our busy schedules.

Individual Contribution

I took a strong role in the group project since the idea for the problem was mine. First, I had to clearly convey the premise of the problem and what I had in mind. A lot of this included educating some of the members of my group on how the American legislative system works and what the data we were looking at meant. Also, I had to explain what web scraping was and how it worked since that was something we did not cover in class.

I still took charge on the web scraping because I was most familiar with the tools. So I went into the United States government vote tracking website and looped through every congressperson for every bill and pulled out the relevant HTML elements. I did this using selenium, a web scraping python package. It was a straightforward process, I used a few functions in selenium to grab the HTML elements and store them in numpy arrays. All of the code used for this can be found in the git repository in its own notebook.

At this point the data was scraped but needed to be cleaned. I created a notebook dedicated to cleaning the data. Each cell in the notebook is used to clean a different column of data which can be found in the git repository. Once the data was cleaned, I handed it off to my teammates for analysis and gave them some time before jumping back into the project.

Once my teammates came back with their data analysis, it was time to create the model. Before we could do this, we needed to create some sort of distance metric to use between each politician in our data set. At first I simply used the total number of bills minus the total number of bills the politicians agreed on, but my teammate found a better distance metric in their research and we ended up using the Jaccard distance. From the Jaccard distance matrix, I used networkx, a python package for creating graphs, to plot the clusters.

Lastly, after doing some research I found that a hierarchical clustering model would work best for our problem given we already had a distance matrix. The Agglomerative Clustering model in sklearn natively takes in a distance matrix and clusters the points together. Once the model was created I checked its accuracy and found that it could classify each politician into their party 100% of the time.

Works Cited

1. govtrack. (2021). *Voting Records*. Retrieved from govtrack.us: https://www.govtrack.us/congress/votes

2. Swallow, E. (2017, November 17). *U.S. Senate More Divided Than Ever Data Shows*. Retrieved from Forbes: https://www.forbes.com/sites/ericaswallow/2013/11/17/senate-voting-relationships-data/?sh=4768ba54031d

3. P. Sharma, "A Beginner's Guide to Hierarchical Clustering and how to Perform it in Python," 27 May 2019. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/

4. Scikit-learn Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

5. pceccon. (2016, March 31). *Generating graph from distance matrix using networkx: inconsistency - Python*. Retrieved from StackOverFlow.com: https://stackoverflow.com/questions/36339865/generating-graph-from-distance-matrix-using-networkx-inconsistency-python

6. S. Glen, "Jaccard Index / Similarity Coefficient," 2 December 2016. [Online]. Available: https://www.statisticshowto.com/jaccard-index/

7. *Scipy.spatial.distance.jaccard*¶. scipy.spatial.distance.jaccard - SciPy v1.7.1 Manual. (n.d.). Retrieved December 2021, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.jaccard.html