Cosine Similarity Analysis of FDR Documents and the Democratic Party Platforms

For this project, I decided to focus to see if the goals set out by political parties from the Party Platforms created at their national conventions are echoed in the later works of the presidents who later represent the parties in office. In this project, I specifically looked towards President Franklin Delano Roosevelt and the Democratic Party because, since he was in office for 12 years. My goal is to see how much Franklin Delano Roosevelt was affected by his party during his approximate 12 year presidency by specifically trying to answer the questions of: How closely did President Franklin Delano Roosevelt align to the goals set by the Democratic Party Platform during his 4 terms in relation to the content he submitted to the press, policymakers, and the public?

Ideally, my dataset would include the entirety of FDR's speeches, campaign documents, presidential documents in date-order, and each of these would have a coded response to a corresponding Party Platform for comparison. This ideal dataset relates to my problem because I would like to measure how each document is close/not close to the ideals laid out by the Democratic Party at the Democratic National Convention. To find real data, I was able to find an interesting site that holds a collection of many documents/speeches/etc. from FDR' s campaign until his death called The American Presidency Project, hosted and curated by professors at the University of California - Santa Barbara. This dataset, of course, is not an expansive list of every single one of his campaign and presidential documents. However, given that FDR was president for 9 years, I can assume that this database would have a good enough amount of data/documents and texts and will allow me to still be able to do an analysis using Natural Language Processing.

The data source for this project is from a database of Presidential documents that are open-source through the American Presidency Project of UC Santa Barbara. In the context of this project, the database has in total 2,582 documents ranging from January 22nd, 1932 to April 13, 1945, which covers the range of Roosevelt's first announcement of his intent to join the 1934 presidential campaign to the last speech he wrote before his death on April 12, 1945. Notably, every document is coded in terms of the context of the document, whether it's a transcript of an interview or letters sent by the President to another party. In terms of the context of these documents, Roosevelt is the author/participator in each document. Given that I am solely using this database to parse the text of each document, I am not completely aware if it contains FDR's full collection of communications within presidential campaigns and the presidency. However, I will frame my analysis and conclusions with the consideration that this collection may not be expansive.

To process this data, I had to use web scraping so I used the requests library, BeautifulSoup, datetime and pandas in order to find a list of all the links, dates, and texts of the files corresponding to the texts from the American Presidency Project website, with the end result being a dataframe with the date of the document, text, category, term_code, dnc_code, days_since_dnc.

category:

- 'press': all documents coded as Press/numerically as 12 from the website, as a placeholder for President to Press communication
- 'policy': all documents coded as Executive Orders/ 58, Messages/ 84, State of the Union Addresses/ 45, as a placeholder for President to Policymaker communication

- ‘public’: all documents coded as Fireside Chats/53, Inaugural Addresses/ 46 as a placeholder for President to the General Public Communication

Term_code (based on Date)

- ‘1’ for First Term
- ‘2’ for Second Term
- ‘3’ for Third Term
- ‘4’ for Fourth Term

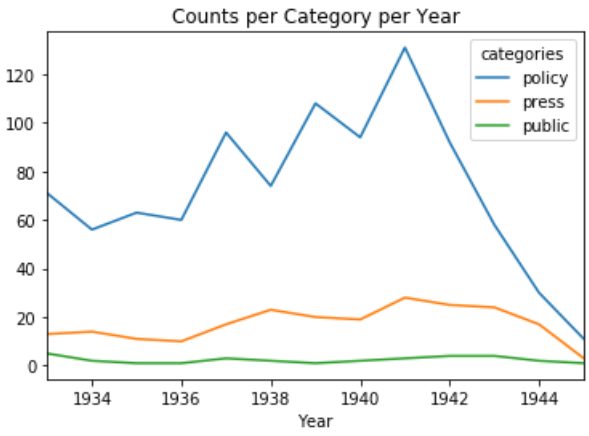
Dnc_code (based on Date)

- ‘1’ for post 1932 DNC Party Platform release
- ‘2’ for post 1936 DNC Party Platform release
- ‘3’ for post 1940 DNC Party Platform release
- ‘4’ for post 1944 DNC Party Platform release

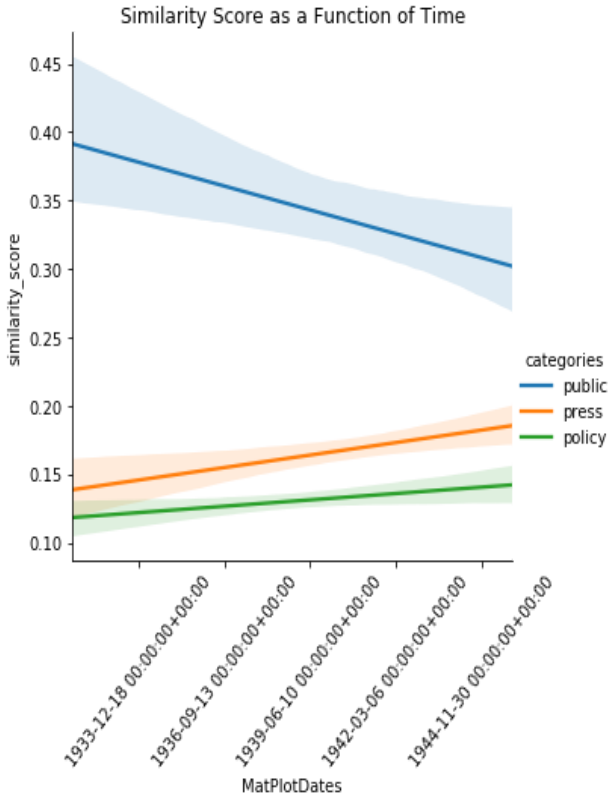
days_since_dnc

- For each dated document, I calculated how many days past it was released from it’s corresponding immediately preceding DNC Party Platform Release in order to have a feature to compare all documents.

From this basic dataframe, I created this basic line chart of the number of documents in the corpus that correspond to that year per category, and was able to find, logically, that most of FDR’s documents in the corpus are policy related, while the press and public compose a smaller percentage of the corpus per year, with the amount of policy related increasing to a peak in 1941.

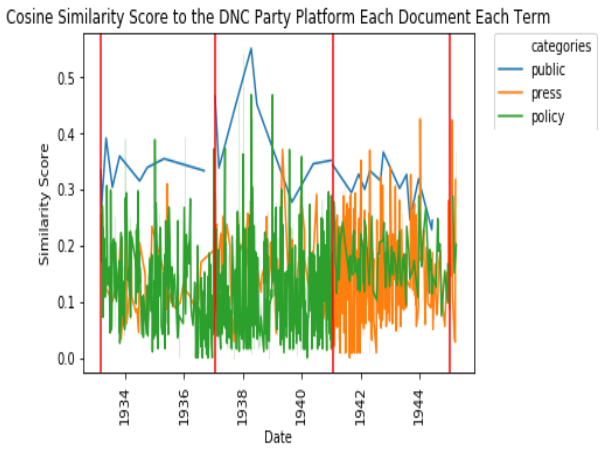


To make answer the question of 'How closely did President Franklin Delano Roosevelt align to the goals set by the Democratic Party Platform during his 4 terms in relation to the content he submitted to the press, policymakers, and the public?' I attempted to measure 'closeness' through the use of cosine similarity and tf/idf. For my idea of closeness, I wanted to see how close a given document was to the party platform in terms of how often it related to words not commonly used in the DNC party platform that preceded it, and used categories, dnc_code and term_code features to see how this related per category, per term and per DNC party platform. From this, I was able to create a regression chart with confidence intervals shown here.



From this I was able to conclude that, overall, publically FDR was more aligned with the goals laid out by the DNC in the respective party platforms in speeches and documents received by the public in comparison to the works and speeches

received by the press and policymakers. Aside from a regression, this trend can also be seen by looking at each document with it's similarity score, (red lines indicating the start of the new terms), we can also see that However, given that there were less document classified as public compared to both press and policy, with the latter of those two being the highest amount comparatively, I also acknowledge the potential bias towards this smaller part of the corpus.



However, when I tried to validate these assumptions and the use of this model to define closeness between the documents and their respective DNC party platform, I found that my model may not have been the best choice for this type of analysis. To validate my model, I added 4 new documents, the Party Platforms from the Republican Party National Conventions from 2000-2012 and found that, for nearly ½ of the FDR documents, they had higher similarity scores to RNC Party Platforms compared to their respective DNC party platforms, thus also leading me to conclude that this method may not have been the best for this analysis

To conclude, I find, that overall, I am unable to parse whether given the model I used. Although the model, before validation, seemed to point towards at least the idea of a relationship between what FDR stated to the public moreso than what was stated to the press or to policymakers/through policy, through attempts to validate my model I am now unable to say conclusively if there is, and conversely, is not a relationship between FDR's texts during his presidency and the DNC Party Platforms.