BU Department of Computer Science

bu.edu/cs

Frontier Culture and Modern Politics in the U.S

Kevin Chen, Jean Paul Vazquez, Zhiwei Tang

Introduction

The movement of frontier settlement was a main part of American history. Historian Frederick Jackson Turner argued that the movement towards the American Frontier fostered individualism, frontier thesis and identity. Factors which, as this project seeks to investigate, may have had a long-time impact on American culture and politics [1].

This project seeks to understand the defining features of the American Frontier by identifying "frontier language" and its persistence over time.

Our data consisted of two historical datasets: Frederick Jackson Turner's speeches, a folder containing the histories of several hundred U.S. counties organized by state as well as two modern datasets to compare with the historical datasets consisting of presidential nominees nomination acceptance speeches and political party platforms.

Term Frequency Analysis

Term frequency finds the amount of times each frontier word appears in a document. As for the term frequency analysis, all the historical texts were run through sklearn's CountVectorizer package with the vocabulary set to the list of frontier words given. The final result of this being a list of lists wherein each list corresponds to a single text and contains the amount of times each specific frontier word appeared in that text. Then, numpy was used to find the average amount of times each frontier word was used and then all these averages were summed together to obtain the average amount of frontier words are used within the historical documents confirmed to be exemplary versions of "frontier-esque" writing.

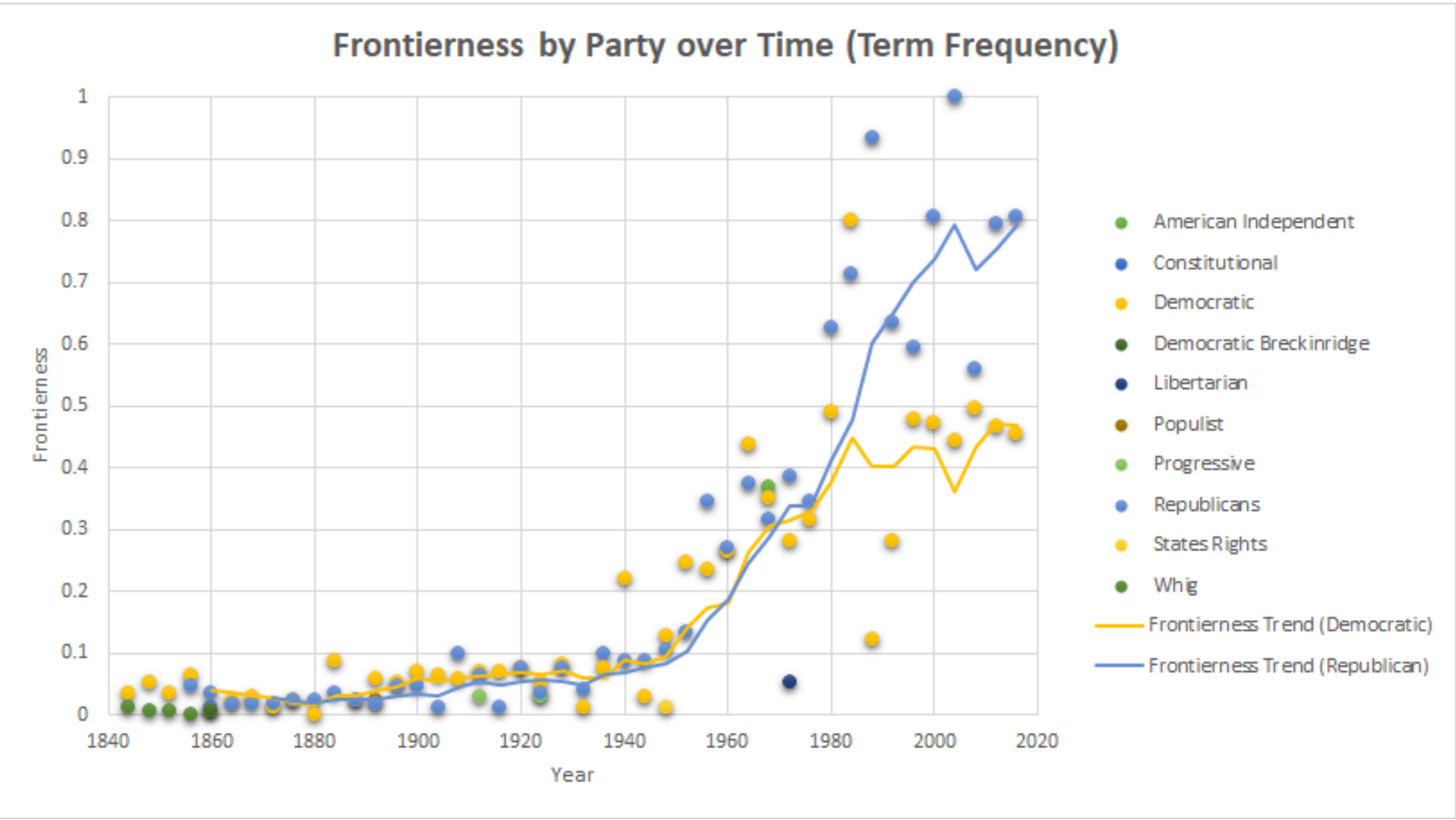


Figure 1. Frontierness by Party over Time (Term frequency)

References

[1] Bazzi, Samuel and Fiszbein, Martin and Melese Gebresilasse, Mesay, Frontier Culture: The Roots and Persistence of Rugged Individualism in the United States (October 2017). CEPR Discussion Paper No. DP12406. Available at SSRN: https://ssrn.com/abstract=3066018

[2] Rajaraman, A.; Ullman, J.D. (2011). "Data Mining". Mining of Massive Datasets (PDF). pp. 1–17. doi:10.1017/CBO9781139058452.002. ISBN 978-1-139-05845-2.

Term Frequency Analysis (continued)

The only issue that resulted from this was that within the modern datasets, even the ones which contained the largest amount of frontier words still had far less than the historical text average. In order to account for this, the same process of summing the averages was performed across all the modern texts and the texts were then arranged in descending order according to their difference between the historical sum of averages and the sum of averages within that text. These values were then run through sklearn's min-max scaling package so as to create a more useful range of values which was dubbed "frontierness".

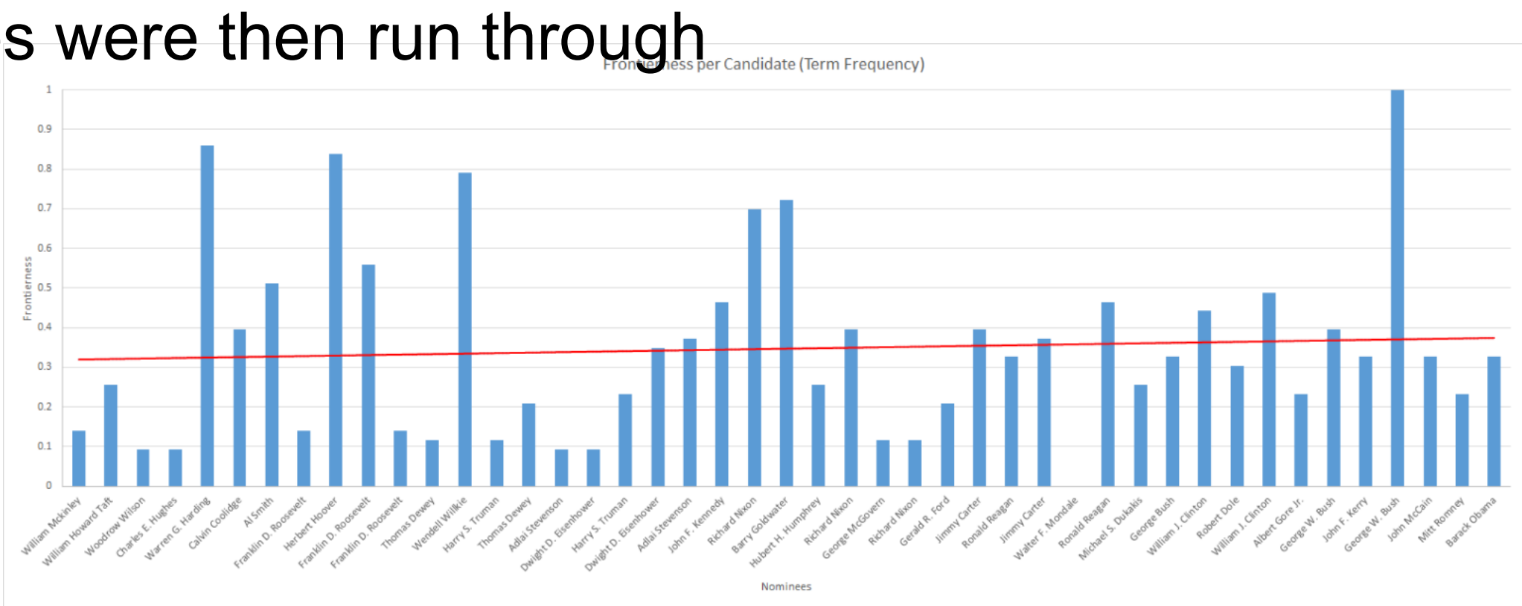
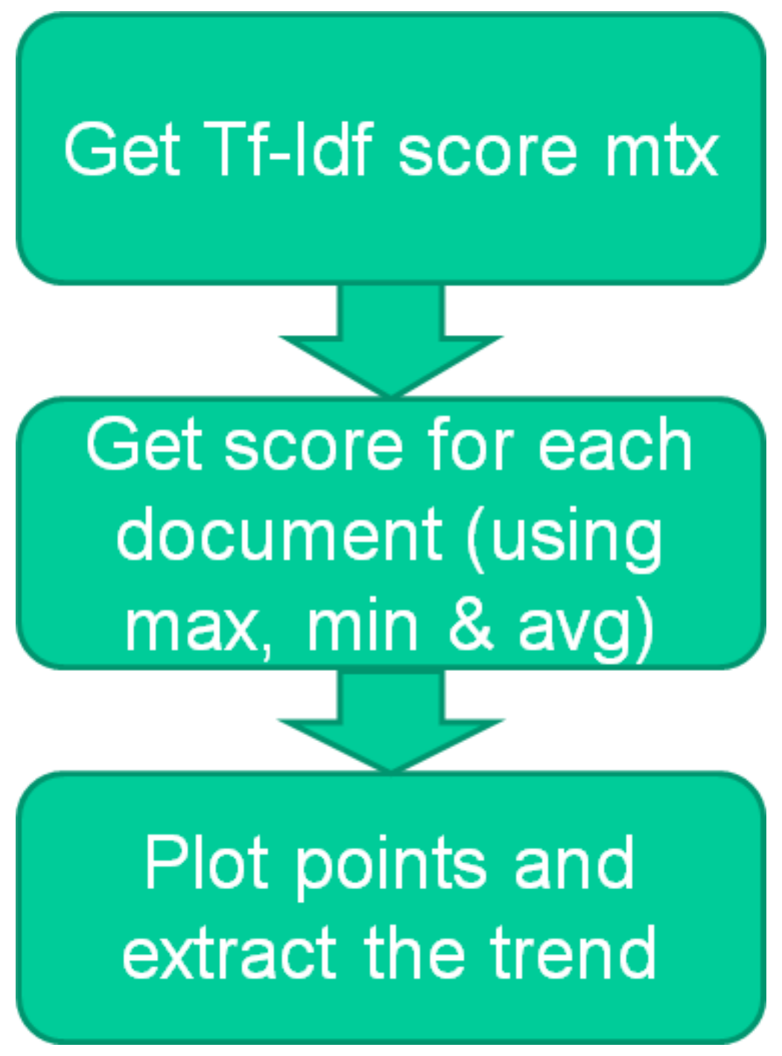


Figure 2. Frontierness per Candidate (Term frequency)

Tf-Idf Analysis

Tf-Idf score is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. [2]

For the Tf-Idf analysis, we built a matrix of Tf-Idf scores of all frontier words with respect to each party platform document of Democratic and Republican party as well as Frederick Jackson Turner's frontier literatures.



Get Tf-Idf score mtx

Get score for each document (using max, min & avg)

Plot points and extract the trend

Figure 3. Tf-Idf Analysis Implementation

To simplify the complexity, we reduced the dimension of the features by getting a score of the "frontierness" of each document. We used maximum, minimum, average(summation) to get the score for each document and compared the performance. Then by comparing the trend of the scores from different party platform documents in time-period basis, we can get a brief result for the popularity trend of frontier words over time among politicians.

Results*:*

The clearest trending comes from the averaging method in dimension reduction. We used move averaging method with a window size of 10 to get a smooth trend line. It generally has a rising trend.

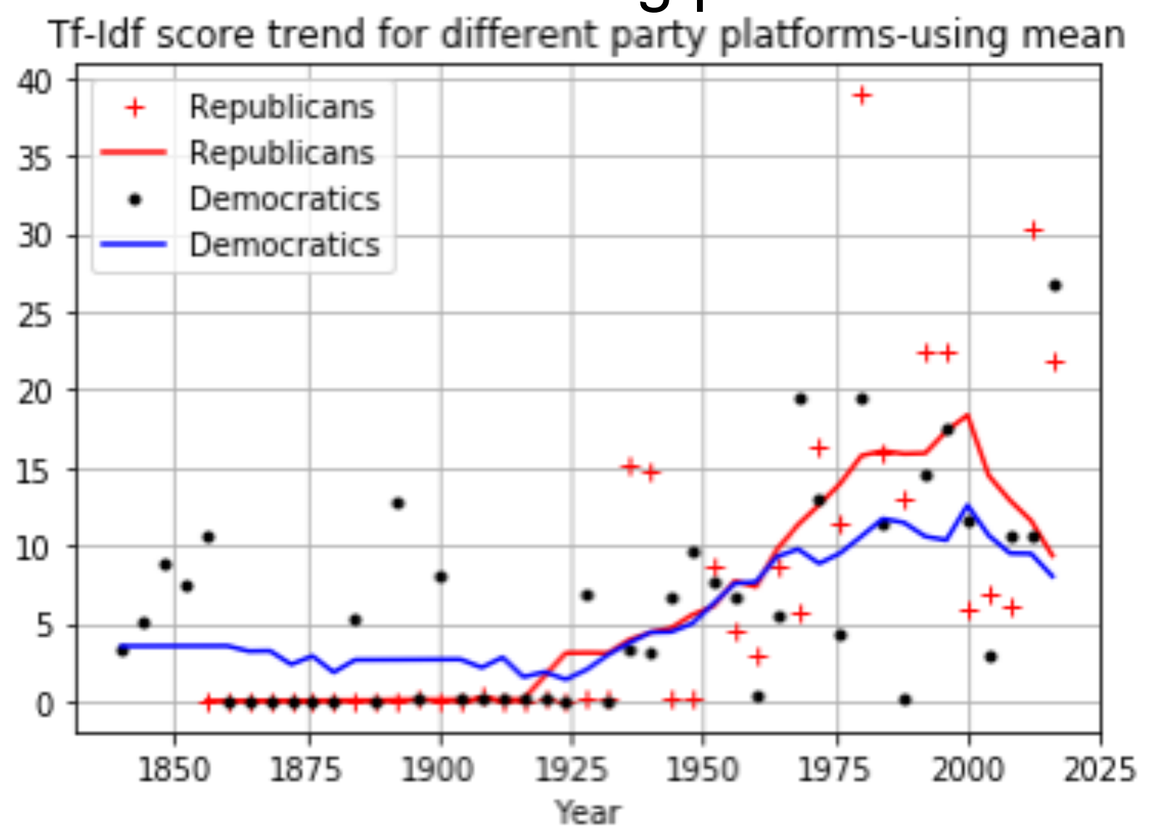


Figure 4. Tf-Idf score trend plot

Near Word Association

"Near word association" consists of scanning through the dataset and grabbing the 10 words closest to every frontier word across all texts and creating a new list of "pseudo-frontier words" from a certain subset of the 10 most common words within this "pseudo-frontier words" list. This last metric was suggested by Dr. Lapets as he found that, in his years of experience with thematic analysis and natural language processing, it is often the case that certain words are used in conjunction with certain other words more often than not. To verify a positive correlation between frontier words and its nearest words, we compared our tf-idf analysis of frontier words to an additional tf-idf analysis of the nearest words. We ran a final tf-idf analysis of frontier words' plus the 10 most common "pseuo-frontier words".

For the "near word association", the 10 nearest words of each frontier word were, by default, the 5 words before each frontier word and the 5 words after each frontier word. If there were less than 5 non-frontier words before or after a frontier word, then the nearest words on the opposite side would be added until there were 10 words nearest to the frontier word. If there was any overlap amongst nearest words, then those nearest words would only be counted once.

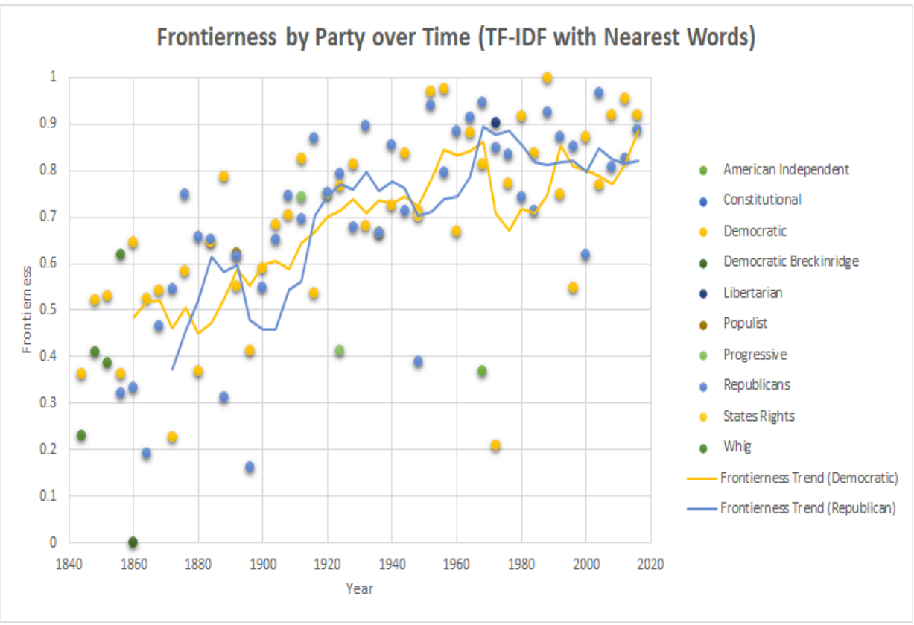


Figure 5. Frontierness by Party over Time (Tf-Idf with Nearest Words)

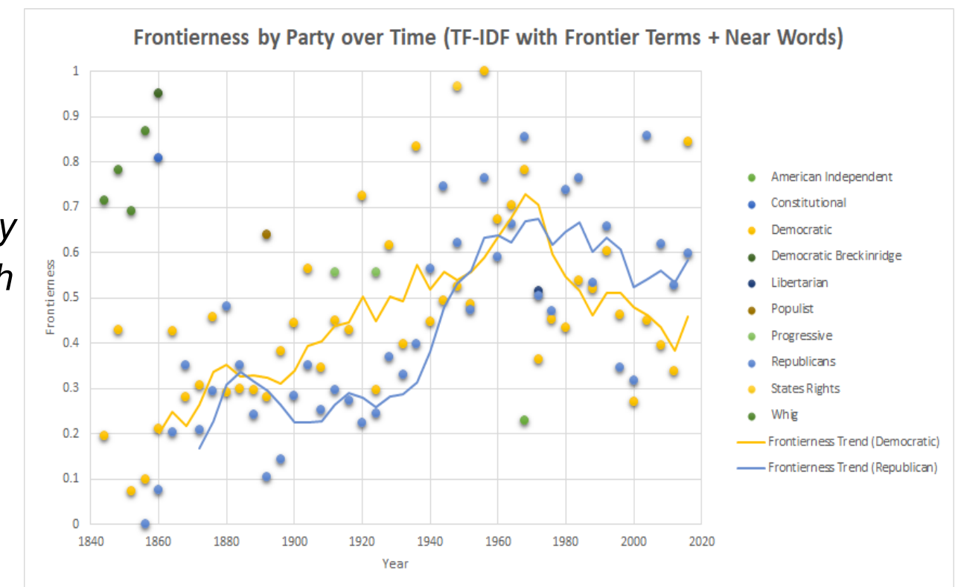


Figure 6. Frontierness by Party over Time (Tf-Idf with Frontier Terms + Nearest Words)

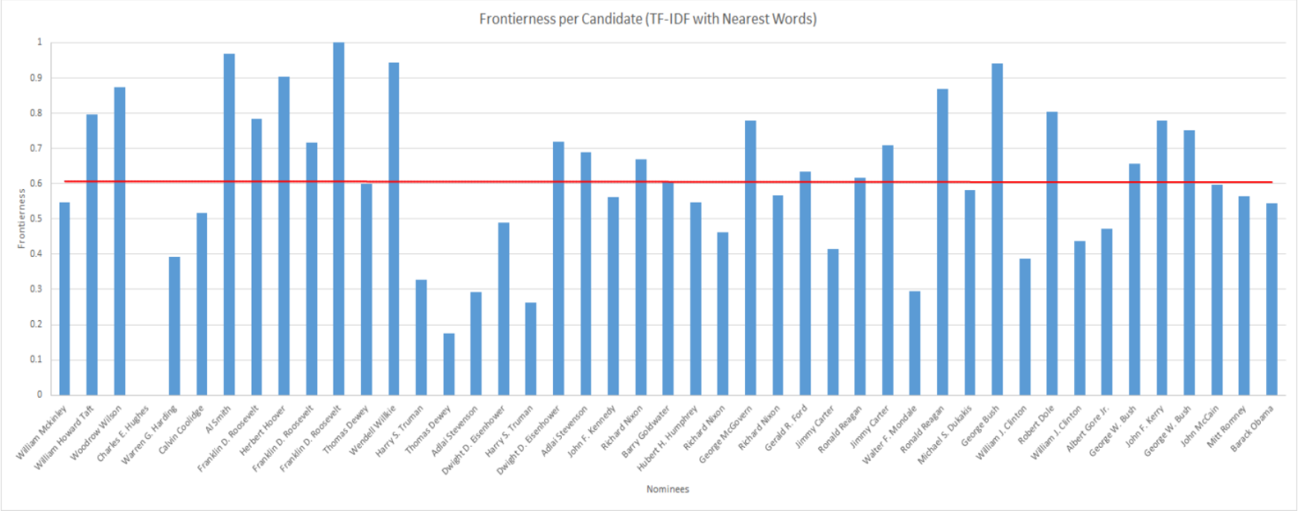


Figure 7. Frontierness per Candidates (Tf-Idf with Nearest Words)

Conclusion

For all three methods, the general trend of "frontierness" score is rising. From the plots of "frontierness" based on both term frequency and Tf-Idf analysis, the rising speed remains low for "frontierness" score before the 1930s and it rises much faster after that. This may lead to the conclusion that the presidential nominees as well as the Democratic and Republican party are possibly becoming more "frontier-esque" as time passes, especially after the 1930s.

@BUCompSci