Topic Shifts Between Two US Presidential Administrations

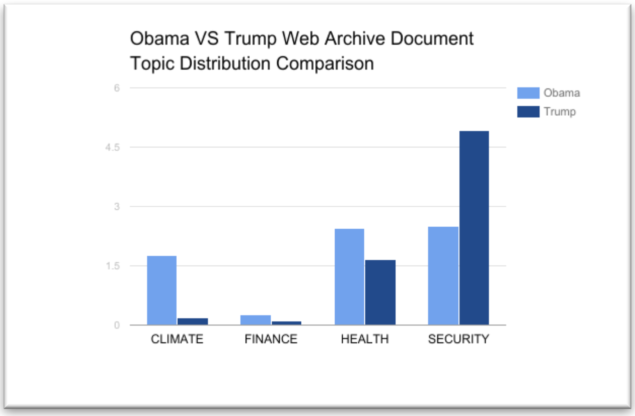
Ziquan Wang, Borui Lin, Ian Milligan, Jimmy Lin

Historians cannot write the history of the 1990s without the Web.[1] Accordingly, web archives such as those collected by the Internet Archive will become a critically important source for learning, understanding, and analyzing recent history. However, the scope of web archives means that historians will not be able to read pages one at a time – instead, they will need to turn to distant reading techniques in order to explore the information within a large number of web pages in a timely manner. A pivotal research area within the web archival field is topic extraction.[2], [3] In this presentation, we discuss our work with text classification on two crawls: the first from President Barack Obama's end-of-term *whitehouse.gov* site, and the second from President Donald Trump's beginning-of-term *whitehouse.gov*.

The two datasets present an interesting contrast, especially in that the Obama administration's site is considerably larger than Trump's (319.6GB vs 12.9GB). We downloaded each crawl from the Internet Archive. For each page within each *whitehouse.gov* domain, we removed boilerplate text such as HTTP headers and HTML tags and extracted the main text content to form a document. Each document was then split into sentences using a pre-built PTBTokenizer from the Stanford CoreNLP package. This generated a large set of unlabeled sentences. Each sentence was then categorized into either CLIMATE, FINANCE, SECURITY, HEALTH, or UNCATEGORIZED. We summarized a list of keywords for each of the four categorized topics by reading sample articles from *whitehouse.gov* and used Spark to examine each sentence and assign a label in parallel based on majority votes of keyword occurrences from each class. We subsequently used this as a seed in our work to create a model to predict further sentences that belong to those topics but did not possess the keyword. If a given page contained no sentences that were categorized as above, the document was more likely to be uncategorized; if it possessed a hit, more sentences are likely to be categorized. We then created a dataset with 306,676 labelled sentences with 18% being categorized and 82% being uncategorized, splitting it into 80% and 20% parts for training and testing. We then used a bag-of-words model to convert sentences into feature vectors, and experimented with Linear Classifier and Naïve Bayes Classifier, finding that the former clearly outperformed the latter.

We then applied this training model to predict the topic of documents in parallel. As seen in the chart, there is an overwhelmingly high percentage of security-related web pages in the Trump web archive. Besides security, percentage of health-related web pages is also very high. This can be attributed to President Trump's desire to replace the *Affordable Care Act* with his *American Health Care* Act. Similarly, the number of climate-related pages in the Trump web archive has fallen.

This paper presents our research of using a Big Data framework such as MapReduce/Spark and Machine Learning classifier library such as the Stanford CoreNLP library to do text classification on *whitehouse.gov* web archives. We have successfully built a model that can well predict the topic (from a pre-defined topic list) of a page.



[1] I. Milligan, "Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives," *Int. J. Humanit. Arts Comput.*, vol. 10, no. 1, pp. 78–94, Mar. 2016.

[2] Y. AlNoamany, M. C. Weigle, and M. L. Nelson, "Detecting off-topic pages within TimeMaps in Web archives," *Int. J. Digit. Libr.*, vol. 17, no. 3, pp. 203–221, Sep. 2016.

[3] G. Gossen, E. Demidova, and T. Risse, "Analyzing Web Archives Through Topic and Event Focused Sub-collections," in *Proceedings of the 8th ACM Conference on Web Science*, New York, NY, USA, 2016, pp. 291–295.