Predicting Popular Government: An Analysis of States

Justin Hines
jph2149@columbia.edu

May 11, 2012

1 Abstract

The term popularity in the modern of era takes on new meaning, especially considering when the exchange of information is instantaneous. Information is now a commodity with a viral life. The virality of information the internet can change public sentiment towards a variety of topics, products, and issues. In this fold has come the exploitation of information in attempts to increase its virality. The commodification of information by content producers has led to the exploitation of said information, at the cost of clarity, correctness, and quality, all to attract more end-users. Popularity now is a means of success based on the merits of "likes", upvotes, retweets, or clicks.

The demand of our increasingly digital world and the instantaneous form of information has forced our government to enter this realm and become a content provider. As such, the federal government's public image is subject to the virality of information. As a result, the government naturally tries to better its public image by catering content towards its end users. NASA is now popular, thanks to death, but also the virality of information spread on social networks and the great content being pushed by NASA.

But at what cost does this commodification occur, especially within the realm of government? Is it possible to garner some statistically meaning of the words in this context and how it relates to popularity. If so, it would seem reasonable that if such a statistical trend exists, we could exploit to predict what articles will be popular and how popular they will be.

Thus, the goal of our modeling will be to classify articles into popularity levels using a Bernoulli Naive Bayes Classifier. Such a model will be scalable for the large dataset described below, but should also provide some insight the word choice popularity distributed amongst the documents, which should prove interesting.

2 Data

The initial dataset for this project came from an achieved data set provided by usa.gov in conjunction with Bit.ly. The dataset, which can be found at http://bit.ly/K8thyx and additional information at http://1.usa.gov/JmGvZq, consists of a variety of data points about end-user clicks on a shorted governmental link. Most governmental websites publish information through these links as a means of collecting analytics. The data is stored as a pub/sub stream of JSON entries, one per line, that represent real-time clicks when the data is collected. In the achieved data, the data are stored similarly. An entry consists of a JSON dictionary with a variety of different data points, including with interest to us, a bitly global hash of link, the end link, and the location where that click came from.

The dataset includes entries from December 2011 to present day. For the sake of sanity a subset of the data was used, from January 1st, 2012 to April 30th, 2012.

While this collection is certainly interesting, it provides no real insights on how information is being shared by what states, and what information is popular where, and why. In addition, this information provides no analysis on the end sight beyond its link.

In order to clean the dataset, a script was written that would filter out an clicks not occurring the United States (this was defined if the data entity contained an originating information, if that location was not in one of the fifty United States). In addition, in order to collect more information about the end site, the site at the link destination was downloaded, and filtered for content only contained in the title or in paragraph tags. Any sites that did not have over fifty words between the title and in paragraphs was automatically excluded. Also, a counter was kept for each time a link was seen being clicked by an individual state, in addition to a global counter that kept track of every time a link was clicked, regardless of the state it originated from.

In order to compute the above, since the dataset is rather large with clicks being on the order of magnitude of 600k unique clicks per filtered, the work was divided and computed on an ec2 cluster, with each month being divided into 6 parts.

In order to apply some meaning on the popularity of content (a strict threshold count was excluded as referring state populations differ dramatically), links were sorted by popularity in terms of click frequency, and then divided into 5 equal labels, in order from least to greatest to highest, low, medium, high, popular, and highly popular. The author recognizes this a rather naive method of applying labels, as an article in highly popular and high may differ by a non-statistically significant amount, but rather both lie on a boundary. For sanity however, this method was applied.

The resulting data was stored in tabular separated value files, grouped by state, along with a global, with the first entry containing the global hash, location, number of clicks, and link content.

To run the data collection and refinement system

3 System

For simplicity, a Bernoulli model of word occurrence in documents was used to generate a Naive Bayes Classification of labels, where each document is modeled as vector with the probability, or weights, based on the probability of each word occurring in a document independently. A Naive Bayes classifier was chosen for its relatively cheap computation and scalability over other models for both training and predictions, as training is a simple linear count of word frequency and prediction. Similarly, predictions in a trained model, calculating the logs odds is linear.

Training and log-odds predictions were very similar to those presented in here http://jakehofman.com/ddm/wp-content/uploads/2012/03/homework$_0 2.pdf For the sake of not being repetitive, I leave it to the reader to read the art$

$Performance was evaluated grouped by state on a series of different \alpha and \beta for smoothing. For each state a 5x5 mat$

4 Analysis

Unfortunately, the reduced dataset and filtering showed that for the majority of states, the dataset was far too sparse to properly attempt to classification. Many states only had in the range of 100-500 unique articles, with the greatest frequency of clicks ranging from 10 to the small thousands. In addition, the concentration of clicks was heavy in the highly popular category (while somewhat expected, many only say an increase to more than 2 articles in popular to highly popular, making attaching labels to low, medium, and high relatively irrelevant. The only set that had enough unique articles to justify an classification and critique was the global set, and yet we have even low accuracies here. For highly popular articles we recorded of accuracy was around .239, which is fairly insignificant, with a total accuracy around .249. For this subset the words are rather interesting, as we can see that the frequency of endorsement is high, along with aid. The endorsement would choice could be perhaps be attributed to the recent republican primaries. Something also of note is the lack of words about NASA, as I thought personally we would have seen a spike since the recent closure of the flight program. Appendix A includes a listing of all the 10 most informative words for each state, accuracies per label and total in addition to the confusion matrix.

5 Conclusion

Overall, results were rather disappointing as I had anticipated a much dense feature space since the initial data set was so large, and seemed to infer a rich potential. However, the dataset was misleading in that the amount of clicks, or each line is extremely concentrated towards certain articles. For instance, we recorded only a shy 1938 unique articles. A shy 9690 unique articles after filter with roughly only 7752 clicks in the bottom 4 labels, with the most popular article, a photo gallery http://www.whitehouse.gov/photos-and-video/photogallery/march-2012-photo-day, got 5356 clicks, which amounts to 70 percent of the entire previous four labels. The top 15 articles amounted to 15302 clicks, roughly 50 percent of all clicks senn. Only the top 50 articles ever saw more than 4 clicks. In short, the government hasn't fallen victim to my dramatic Abstract, as no one ever visits governmental websites!

6 To Note

The cleaning of this dataset provided to be the most difficult aspect. On the first run through the bitly data set, I by-passed the note that I should keep a hash of articles I had seen so I don't re-download them and I only came to this realization after running the script for more than 24 hours. The second time I was running into memory issues as the amount of ram on my laptop was insufficient to store all the downloaded content. In order to reduce space, I ran it again used c pointers using c types, but I did notice until much later than when c pointers are written, they print the memory location the pointer as opposed to the string. It was a rather frustrating experience on my levels, and I hope you can provide some leniency on my tardiness. Please check my github repo at https://github.com/JustinHines/PredictingPopGov.