deletweet-conclusion

April 3, 2017

0.1 CONCLUSION

Leveraging machine learning techniques such as text mining, natural language processing, and k-means clustering, we were able to attempt answers to the 4 questions established in the introduction.

Using NLTK's tools for text mining and NLP - including their domain-specific social media tokenizer - we were able to pull out the most common words and themes of the tweets in the Politwoops dataset. We found terms and phrases common to politics in general, as well as words and names specific to the timeframe of the dataset (2011-2017). Along the way we established the importance of normalization of the text to the success of this process.

Next we took a close look at the nature of the dataset as not only a set of tweets by politicians, but as a set of tweets that were specifically deleted by politicians. Referencing previous research done into deleted tweets, we established a threshold to distinguish tweets that were most likely deleted for aesthetic reasons from tweets that may have been deleted for other regrettable reasons. We performed content analysis on this subset, in an effort to distinguish it from the dataset as a whole. With the same goal in mind, we also analyzed the times tweets in the subset were created and deleted relative to the dataset as a whole. And while some differences were noted, nothing concrete was found to explicitly distinguish the two. While ore comparison is needed against a set of tweets by these same politicians that were not deleted, it seems that the vast majority of the tweets have been deleted for aesthetic reasons such as misspellings, improper formatting, broken links, etc.

Using NLTK's Naive Bayes Classifier, as well as their previously classified tweet corpus, we performed sentiment analysis on the dataset, with close to 75% of the tweets being classified as positive, with the remaining classified as negative. Introduction of a third, neutral class to this process is needed to make the results more robust.

Finally, we showed a proof of concept hashtag recommendation system that uses K-Means and Jaccard Distance to cluster similar tweets together and recommend hashtags from tweets in the cluster to its neighbors. While this system is effective, it is extremely computationally expensive, and therefore not suitable for use outside of a research domain.