Proposal: Predicting New York Times Article Readership

Dave Holtz, Jeremy Yang, Michael Zhao

November 11, 2015

Part I: Description of Project

Our group has access to micro-level readership and sharing data from the New York Times from April 3, 2013 to October 31st, 2013. We are interested in using content and contextual data to predict how many page views a given article will receive. This task itself should be achievable using relatively straightforward regression techniques. However, the real challenge of this project lies in the feature extraction. A 2012 paper by Berger and Milkman [1] found that the virality of online content is related to the positive or negative valence of the content of the article. They framed the question as a classification problem, wherein they extract numerous features from an article and trained a model to predict whether or not that article would appear on the New York Times most e-mailed list. The feature extraction methods they use provide a great baseline for comparison, and also a great starting point for our research. We expect that with more granular data and more sophisticated feature extraction techniques, we can achieve even better predictive accuracy for a very similar (and arguably more important) task.

We are currently considering the following features (and methodologies to extract them) for inclusion in our model:

- Sentiment analysis (Naive Bayes text classification and tf-idf weighting)
- Contextual data (day-of-week and time of publication, article topic, article author, Flesch-Kincaid Reading Ease, etc.)
- Article uniqueness (determined using a Neural Network based language model with word vectors)

Note: Michael is also working on this project for Natural Language Processing (6.864). Dave and Jeremy are only enrolled in Machine Learning.

Part II: Timeline and Division of Labor

Timeline:

- Nov. 12 - Nov. 19: Scrape New York Times article text, Aggregate readership data and collect non-text contextual data
- Nov. 19 - Nov. 25: Perform sentiment analysis on text, begin training of neural network language model
- Nov. 25 - Dec. 2: Complete training of neural network language model, train readership regression model
- Dec. 3 - Dec. 8: Tweak readership model, write report, do additional fixes and optimizations as necessary

Division of Labor:

- Dave: Scrape New York Times article text, neural network language model, sentiment analysis, contribute to report, last minute optimizations

- Jeremy: Scrape New York Times article text, sentiment analysis, readership model, contribute to report, last minute optimizations
- Michael: Non-text contextual data and readership data aggregation, neural network language model (primary contributor), readership model, contribute to report, last minute optimizations

Part III: Risks

This project brings with it a few risks. They are listed below.

- The neural network language model proves very difficult to train, both in that it can be complicated to train for programmers, and require significant computational resources / time. Although we are confident we can get code to function, it may not work well given the amount of time necessary to both train the neural network and find appropriate parameter values.
- Data sanitization could more time-consuming than anticipated. We are using data that comes directly from the New York Times database, as well as text data scraped from the internet. This data is real and messy. It may take longer than expected to extract useful features from it, and certain features may prove infeasible to extract.
- Although we plan on using Berger and Milkman's result as a rough guideline, their prediction task was fundamentally different than ours, as their labels are categorical rather than numerical. We may find it difficult to draw direct comparisons between their results and ours, and may need to resort to a different (and simpler) baseline of comparison.

References

[1] Berger, Jonah, and Katherine L. Milkman. "What makes online content viral?." *Journal of marketing research* 49.2 (2012): 192-205. APA