How Does Content Drive Viewership?

Dave Holtz[1], Jeremy Yang[2], Michael Zhao[3]

Abstract

Why do some webpages receive massive numbers of pageviews? To determine how content drives viewership, we construct a unique dataset of all articles published by the New York Times (NYT) in August 2013. Our dataset is built from 2 major components, the NYT's internal web traffic data and article content data parsed from the NYT website. We use the internal web traffic data to accurately track the number of page views of each article as well as construct a set of robust control variables such as the desk and section of each article. To build content features, we use various machine learning and statistical natural language processing techniques on our parsed article content data and construct features such as article perplexity, sentiment, reading difficulty, and indicators that denote the presence of pictures, videos, etc. Additionally, we have access to the NYT's internal website traffic data. We feed all of our constructed features to into a predictive regression model. We find [MAJOR RESULTS HERE].

[1] dholtz@mit.edu
[2] zheny@mit.edu
[3] mfzhao@mit.edu

Contents

1. Introduction

In today's digital economy, many companies are very interested in attracting users to visit their websites in order to earn ad revenue. While many factors might motivate a user to visit a particular page, certainly one important factor is the content in that webpage. This paper explores the relationship between the content of a webpage and the number of page views it ultimately ends up receiving by constructing a unique dataset of all articles published by the New York Times (NYT) during August 2013. This dataset is built from two major components: the NYT's internal web traffic data and parsed NYT article content data.

Typically, a study such as ours tends to be very difficult to conduct as either accurate measures of viewership are unavailable[1] or the feature extraction of the content is too challenging (for example Youtube), or or both. Fortunately, our access to the the NYT's internal web traffic data allows us to exactly measure the number of page views an article receives. The web traffic data is rather rich and also includes internal meta-data that we use to build various control features. Moreover, since we are working with mostly textual data, we are able to take advantage of recent advancements in machine learning and statistical NLP to do feature extraction

[1] While oftentimes precise viewership data tends to be not available openly, oftentimes researchers use related observables, such as Facebook likes

on article text.

A similar study by Berger and Milkman (2012) [?] examines the relationship between content and word-of-mouth virality. They find that the emotional content of a NYT article is predictive of its virality. Using simple measures of an article's sentiment and emotionality, Berger and Milkman show that positive articles are more likely to show up on the New York Times "Most-Emailed" list. They also show that articles that evoke high physiological positive or negative arousal (such as awe or anger) tend to be more viral than articles that evoke deactivating emotions (sadness). We build on this study in two ways: first, we relate an article's content back to the number of page views it receives rather than its virality[2]. Second, we employ more sophisticated machine learning feature extraction techniques to see if they work any better over their simple measures.

2. Data

2.1 NYT Internal Web Traffic Data

Our NYT internal web traffic dataset is a record of all individual user activity on the NYT website covering the period of April 3rd, 2013 to October 31st, 2013. This activity data is stored as individual lines of json and includes who (if available) accessed what page at what time. Overall, it is over 20 terabytes in size and contains over 3 billion page views[3] Since the scope of this dataset is so large, we initially restrict this project to a single month, August 2013.

We limit our dataset to consist of pages that only contain articles or blogposts published during the month of August. We parse the data to obtain a list of urls, which need to be stripped of potential garbage. After cleaning up the url data, we are left with 6682 unique pieces of content. We then parse our dataset and aggregate the number of counts each url receives. In order to make the comparison between articles fairer since an article that's been out longer will have more page views on average, we only count the number of page views received up to 7 days after publication[4]. In total, our data consists of over 250 million page views. As seen in Figure 1 below, the distribution of page views is highly skewed with very have tables. After applying a log transformation (as seen in Figure 2), our distribution looks considerably more normal.

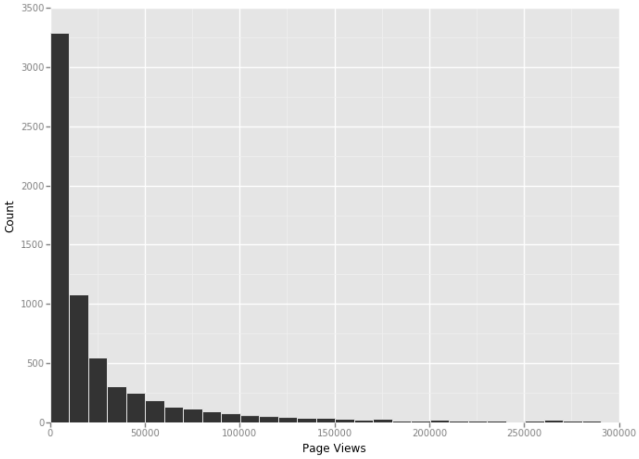

Figure 1. Histogram of Articles by Number of Page Views

Table 1. Page Views Distribution Summary Statistics

Total Page Views	248161455
Min	1
Max	2545288
Mean	37138.8
Median	10298.5
Std. Dev.	88972.9
Skewness	9.52191
Kurtosis	173.061
Observations	6682

In addition to aggregating the counts, when

2.2 Parsed NYT Article Content Data

[2] Which companies arguably care more about since word-of-mouth virality is usually a means to increase page views

[3] Not all page views are article views, for example, some events that are also tracked are searches, or user account settings.

[4] Given that page views tend to sharply drop off soon after publication since recency is quite important to the News, the number of page views obtained during the 7 days after an article is published represents the vast majority (usually well over 90%) of total page views an article receives.

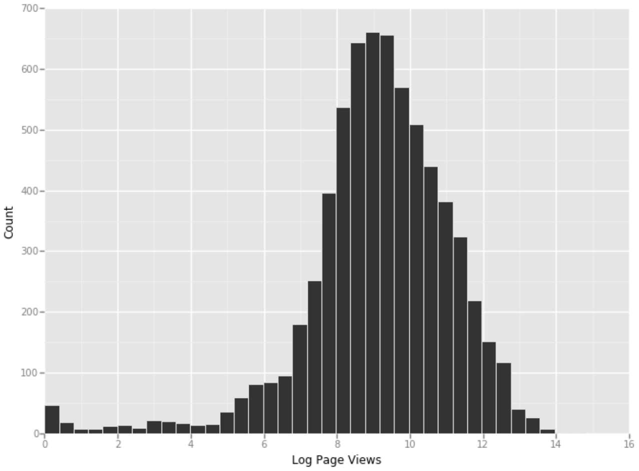

Figure 2. Histogram of Articles by Log of Page Views

Table 2. Log Page Views Distribution Summary Statistics

Min	0
Max	14.74975
Mean	9.122868
Median	9.239754
Std. Dev.	2.028668
Skewness	-1.270368
Kurtosis	3.800911
Observations	6682

3. Constructed Features

3.1 Subsection

Table 3. Table of Grades

Name		
First name	Last Name	Grade
John	Doe	7.5
Richard	Miles	2

Word Definition

Concept Explanation

Idea Text

3.1.1 Subsubsection

- First item in a list

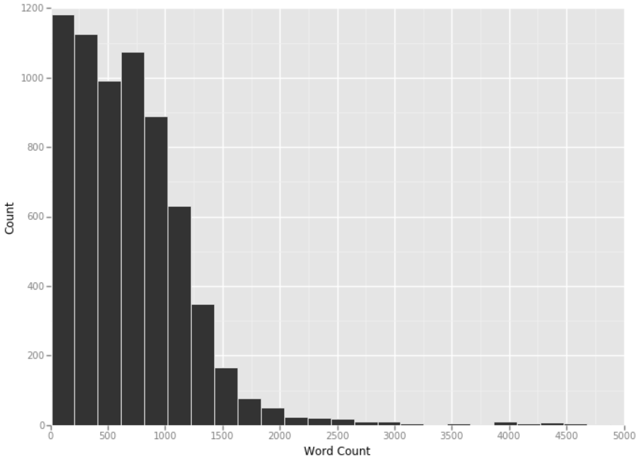

Figure 3. Histogram of Articles by Word Count

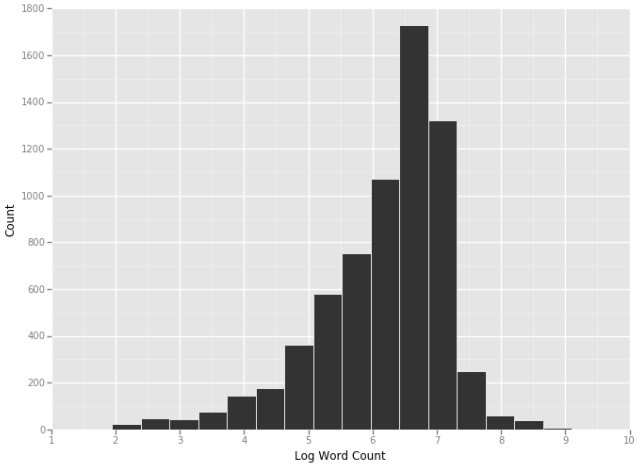

Figure 4. Histogram of Articles by Log Word Count

- Second item in a list
- Third item in a list

3.1.2 Subsubsection

3.2 Subsection

4. Predictive Regression Model

Acknowledgments

So long and thanks for all the fish [1].

References

[1] A. J. Figueredo and P. S. A. Wolf. Assortative pairing and life history strategy - a cross-cultural study. *Human Nature*, 20:317–330, 2009.