– Independent Work Report Fall, 2016 –

Discovering Princeton's History: A textual analysis of collegiate newspaper headlines

David M. Liu

Adviser: Dr. Brian Kernighan

Abstract

Building upon previous efforts to digitize the Daily Princetonian textual archive, this project provides an interface for quantitatively analyzing historical, linguistic, and cultural trends in the archive. The purpose of the project is to make the trends embedded in the archive more digestible and available to the general public. For implementation, the project drew on work in sentence-level sentiment analysis and n-gram visualization, combining the two into a single interface. The dataset consisted of processed copies of all Daily Princetonian newspaper headlines since the paper's inauguration in 1876. Early work revealed that sentiment analysis was not effective for visualizing the archive's trends, let alone the bias. The solution was to visualize the trends using n-gram distributions that also included sample headlines chosen using sentiment analysis. The addition of sample headlines to the n-gram increases the ability of the viewer to trace the trends back to their original sources in the text. Using the final visualization interface, which is available at 54.203.14.54:5000/?keywords=men,women, *especially notable trends were discovered for the search terms "oriental" and "vietnam", for example. These visualizations effectively reflect the University's historical and cultural past in a quantitative manner and can be used by university scholars, Daily Princetonian writers, and the general public.*

1. Introduction

Motivations and Goals

Newspapers have traditionally lagged behind other forms of media in terms of technological advancement and development. While social media, television, and online news sites are inherently digital, newspapers are challenged to innovate in the 21st century when their primary medium relies on print and paper. On the other hand, the historical longevity of newspapers can be seen as a strength and opportunity; more so than their modern counterparts, newspapers consistently and thoroughly document histories of nations, towns, schools, and college campuses. Even today, in the technological age, newspapers still publish on a daily basis, recording and documenting history day in, day out. As a result, from a data mining and analysis perspective, newspapers are an untapped source of meaningful, historical data that has yet to be studied computationally.

The data set studied in the present project, the online archive[1] of the *Daily Princetonian* – Princeton University's 140-year-old daily newspaper – is no different from the above generalizations. It was not until 2012 that the paper's archive became digitized. Prior to the nearly $300,000 endeavor, university researchers would have to sift through physical bound copies of the paper or even parse the saved microfilm itself. Once the archive had been completed following five years of fundraising and laborious digitization, University Archivist and Curator Daniel Linke said, "This is going to change how Princeton looks at itself. It [the archive] will really give it a window into the past that we haven't had [before]", per a 2012 *Daily Princetonian* article[2].

[1] http://theprince.Princeton.edu/
[2] http://www.dailyPrincetonian.com/article/2012/05/prince-to-complete-digital-archives-dating-back-to-1876

In turn, the primary goal of the present study is to take the next step forward following the *Prince*'s digitization and begin to analyze the data itself. Specifically, the aim of the project is to visualize the archive's textual data so that other researchers, students, staff, and faculty can better understand the university's historical, cultural, linguistic, and political trends. As a digital humanities project, the present study will apply modern tools in language processing and visualization to a field – college journalism – that has been underutilized from a data science perspective. By visualizing the data, the trends within the archive will become more apparent, digestible, and understandable to the casual browser.

Overview of Project

After I decided to visualize the *Daily Princetonian*'s archive, many practical challenges and design questions soon arose. The most important question facing the project involved defining the trends that the visualizations sought to illustrate. Unlike some other data visualization projects, the need for large quantities of data was not an issue, as the archive offered hundreds of thousands of articles to analyze. Instead, a specific metric for quantifying the textual data needed to be chosen. In the ideal case, the visualizations would provide clear, understandable, and meaningful insights regarding the archive.

With endless visualization possibilities, my personal interest became one of the major deciding factors when making project design choices. Initially, I made several efforts to track the degree of bias in the newspaper over time, with the goal of exposing any losses of objectivity. However, the task of measuring bias proved to be beyond the practical extent of the present study, as discussed in later sections. As an alternative, n-gram visualizations proved to be much more feasible and informative, yielding clear, understandable trends in the archive data.

The remainder of the paper will begin by discussing some of the prior, related work that motivated the later implementation choices in developing the visualizations. Thereafter, the bulk of the paper will discuss the results of the various visualizations, evaluating their effectiveness in clearly depicting historical trends. Finally, a select few example visualizations will be discussed, motivating future work as well.

2. Problem Background and Related Work

Few prior studies have shared the same goal of visualizing newspaper archives, so it was difficult to explicitly develop a project based on existing work in the field. At the same time, development in the fields of language processing and OCR analysis offered significant research to guide the present study. The following section will describe the developments in sentiment analysis and Google's N-gram project and explain how those studies guided key design choices for this project.

Sentiment Analysis and Classification

Broadly speaking, sentiment analysis is a burgeoning field of language processing that aims to compute numerical polarities – typically positive and negative – for input pieces of text. Over the past decade, sentiment analysis has been applied predominantly to a field known as opinion mining, which determines public sentiment on specific entities [1]. The quintessential example of opinion mining involves processing movie and product reviews to quantitatively assess public opinion on these consumer goods. Other studies have mined online news and blog sites to monitor public approval of famous celebrities, companies, and locations [2]. This project will focus less on opinion mining and tracking sentiments of specific entities and more on sentiment classification of text, a broader and more general research problem.

With multiple prior implementations of sentiment analysis, from machine learning models to bag-of-words approaches, a key design choice involved choosing between using an existing sentiment analysis tool versus developing an in-house solution [1]. Because the majority of public sentiment analysis libraries are designed for modern corpuses, typically social media text, ideally I would have developed a library tuned specifically for the *Prince* archive. However, having studied the amount of human labor required to create a complete sentiment analysis library, I determined that an existing library would be a better option. Using an existing library would allow the project to focus more on the goal of visualizing the archive as opposed to making contributions in the field of sentiment analysis theory.

The sentiment analysis library chosen for this project is named VADER, or Valence Aware Dictionary for sEntiment Reasoning [3]. VADER was chosen as a suitable sentiment analysis library because it is provided in a popular language processing library called NLTK; using VADER would ensure a certain degree of reliability and practicality for a project that focused more on visualizing the archive rather than developing a sentiment analysis library. At the same time, the details of VADER's implementation were also suitable because the library utilizes a modified and improved bag-of-words approach that also considers context and sentence structure. Though this approach is far simpler than designing a machine learning model, the VADER developers showed that their simpler design actually yielded better performance when analyzing the sentiment of New York Times editorials – a similar dataset to the present study [3]. Additionally, the simplicity of VADER allows for more efficient computation, whereas machine learning models must be trained for each corpus [1].

To understand the strengths and limitations of sentiment analysis, in this project, it is important to delve into VADER's implementation details. Namely, VADER assigns positivity

and neutrality scores based on two major criteria: a "Gold Standard Vocabulary List" and five additional modifying heuristics. The vocabulary list in VADER's implementation reduces down to a manual mapping between words in the English language and a positivity valence ranging from -4.0 to +4.0. While simple, assembling the list was far from trivial as the researchers recruited human raters from Amazon's Mechanical Turk service to complete the mapping. Following a tedious process of human rating, the final vocabulary list included verified sentiment mappings for 7,500 individual tokens, including words, acronyms, and even emoticons. A small excerpt of the mapping is shown in the table below. By relying on VADER's vocabulary list, the present study could avoid the time-consuming process of mapping the English language.

"VADER" Gold Standard List (Excerpt)	
Word	**Positive Valence**
"okay"	0.9
"good"	1.9
"great"	3.1
"horrible"	-2.5

Table 1: Sample from the VADER Sentiment Library vocabulary mapping

Whereas a bag-of-words approach would have concluded with the vocabulary list, VADER also rates sentiment scores based on five additional heuristics that the developers themselves identified. These heuristics are as follows: punctuation, capitalization, degree modifiers (e.g. *extremely),* contrastive conjunction (e.g. *but*), and negation detection. While a few of the heuristics are not relevant to sentiment analysis of newspaper text, such as "punctuation", these five additions elevated VADER's performance by also considering the context of inputs. These heuristics also attest to the difficulty of sentiment analysis. For example,

the addition of a seemingly simple word such as “but”, can completely alter the sentiment of the text. In short, VADER is an imperfect yet well-tested sentiment analysis library for application development [3].

Putting these factors together, VADER calculates a positivity, negativity, and neutrality score for each input sentence. The first step of the process involves translating each word in the sentence to its corresponding valence, ranging from -4.0 to +4.0, as specified in the “Gold Standard Vocabulary List”. Words not contained in the list are assumed to be neutral and have a valence of zero. Next, the valence of each word is adjusted according to the aforementioned heuristics, either increasing or decreasing the value. Finally, the individual sentiment scores are calculated as fractions of the total valence. The total, or denominator, is defined as the sum of all positive valences, the absolute value of the negative valences, and the number of neutral words. By definition, the total of the three scores will always be one. For further clarity, these expressions are written in the equations below:

$$Total = \sum positive\ valences + \sum |negative\ valences| + (num\ neutral)$$

$$positivity\ score = \frac{\sum positive\ valences}{Total}$$

$$negativity\ score = \frac{\sum |negative\ valences|}{Total}$$

$$neutrality\ score = \frac{num\ neutral}{Total}$$

N-grams

N-gram visualizations are composed by plotting the relative frequencies of query words over time. For example, to calculate the n-gram frequency of the word “women” for the year

2000, one would divide the number of occurrences of the word "women" by the total number of distinct words in 2000 for the chosen corpus. Although the technique is simple and straightforward, unlike generating machine learning models for sentiment analysis, n-grams did not become popular until Google exploited the technique in 2011 [4]. What made Google's approach different and well-known was the size of its corpus, 5 million digitized books spanning centuries. According to the original paper, Google's textual corpus contained roughly 4% of all books ever published. By leveraging such a large dataset, Google's n-gram visualizations were more complete. An example of Google's visualization tool in use is shown below in Figure 1.

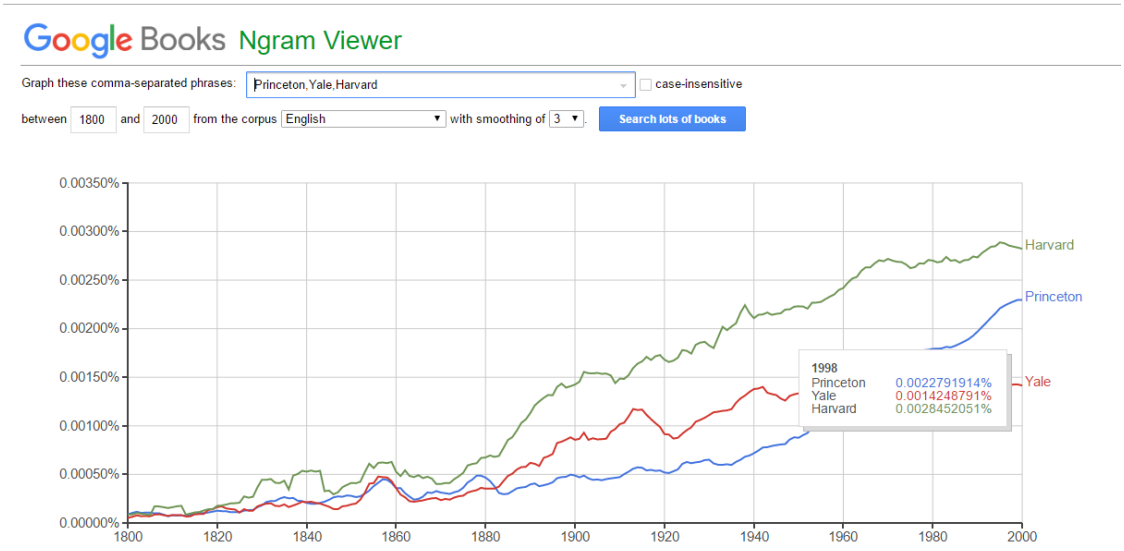


Figure 1: Example of Google's N-gram Viewer

Compared to Google's n-gram dataset, the present study only utilizes a few million words, or 1/10,000th of Google's dataset size. Thus, the n-gram visualizations from the present study will exhibit much more variability. Additionally, the vocabulary of query terms will be much smaller; whereas one could arbitrarily search for the term "gypseous" in Google's n-gram, such a search will yield null results in the present study. On the other hand, one advantage of the present dataset is specificity. For Google's project, the books comprising the corpus spanned

uncountable subjects, fields, and even languages – it would be extremely difficult to trace trends in the visualizations to their original source in the text. Additionally, disproportionate representation from certain genres and subjects could bias the Google results. In contrast, the entire corpus for the present study centers around a single college campus, thus the resulting trends, though possibly fewer than Google's, will all directly link back to the Princeton campus.

Perhaps the greatest takeaway from Google's pioneering paper on n-grams is not the technique itself but rather the conclusions and observations the researchers were able to draw from the data. From the beginning, the n-gram researchers describe the project as a study of "culturnomics", which means to "observe cultural trends and subject them to quantitative investigation", very similar to the present study's main goal [4]. The Google paper later elaborates, citing potential trends in "lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology" [4]. In fact, using OCR'ed copies of five million books, the Google researchers were able to track the usage of words such as "snuck" vs. "sneaked", cases of influenza over time, and even the differences between American English and British English. Due to a much smaller corpus, the present study will not be able to analyze nearly as many topics, but will draw significant motivation from the examples set forth by the Google n-gram project. A successful implementation would reveal culturomic trends in the Princeton campus and greater college student population.

3. Approach

Initial Setbacks

Motivated at first by personal interests, the project initially focused on applying sentiment analysis to the *Daily Princetonian* archive. With the grunt work already completed in

VADER, I was excited to observe the resulting trends when applied to a large dataset. Heading into the beginning phases of the project, my two major research questions were: First, would a sentiment library designed for modern media text perform well when applied to archival data spanning a century's worth of publications? Second, if so, what meaningful trends would emerge from such an application of sentiment analysis?

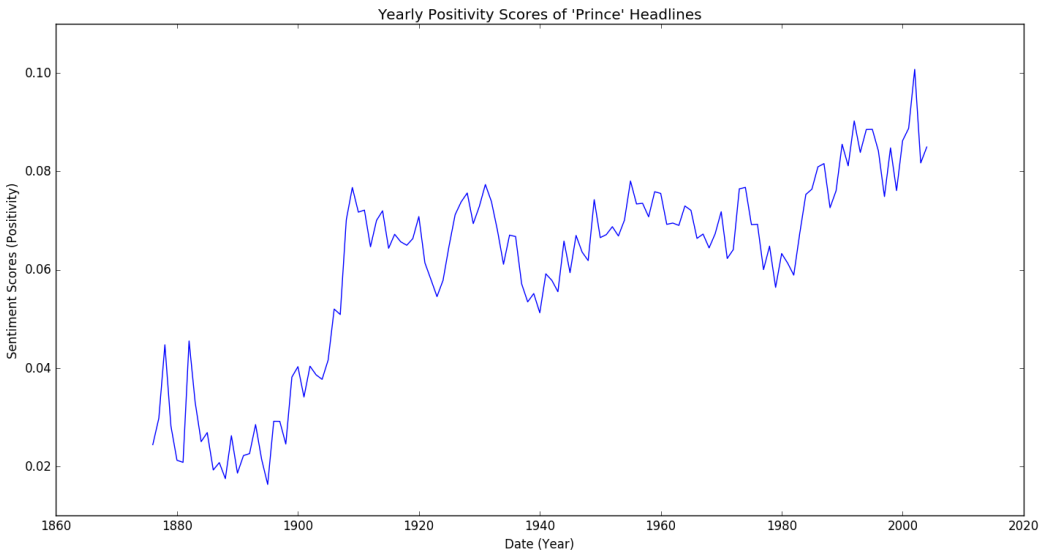


Figure 2: Average positivity sentiment score for headlines by year

The first few plots of sentiment scores were designed to ambitiously measure the degree of bias in the *Daily Princetonian* headlines over time. I studied only headlines because VADER is not equipped to score input text longer than a few sentences. Previous studies have, in fact, performed sentiment classification on entire documents, however these methods usually require greater assumptions and have more inconsistency. For example, a previous study that classified entire movie reviews assumed that each review, or "document", contained only one opinion holder (the viewer) and one opinion target (the movie). Because such assumptions could not be

made for the *Prince* dataset, I used sentence-level classification instead. Because headlines are succinct abbreviations of the article, the information loss is minimized.

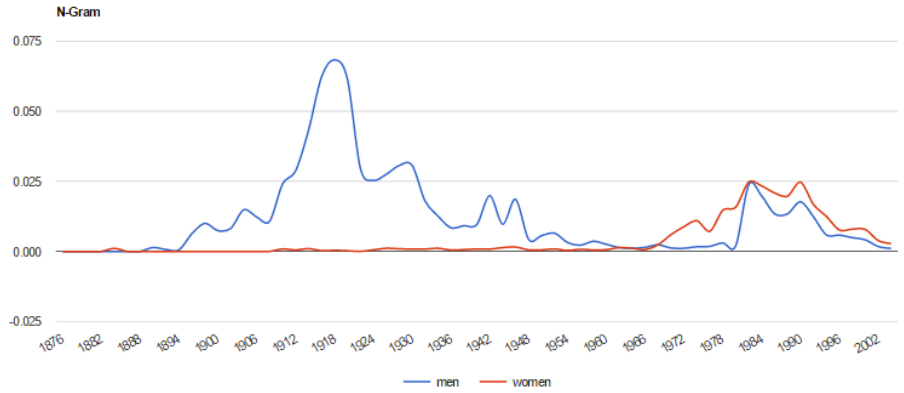
To calculate the bias, I presumed that a neutrality score – approximately the fraction of neutral words in a sentence – could serve as a bias heuristic, equating neutrality with objectivity. However, a deeper reading into bias detection debunked this initial hypothesis. A counter example to the above hypothesis could be the headline, "Princeton student accused of plagiarism, facing suspension". Though such a headline would receive a very negative sentiment score, it could also be a fair, objective headline. Dillon Reisman explains that detecting bias in newspaper headlines is especially difficult because these small pieces of text are very "information dense" and measuring bias is often subjective [5].

Furthermore, the initial basic plots of sentiment scores over time, themselves, proved to be unsatisfactory as trends were arbitrary at best. An example positivity score plot is shown above in Figure 2, with no discernible and verifiable trend present. The results of these initial experiments will be discussed later in the Evaluation section.

Combining Sentiment Analysis with N-grams

In search of a different metric to visualize the '*Prince*' data set, I eventually turned to the n-gram visualization approach. I had erroneously assumed that because the n-grams were simpler from a computation perspective, they would yield less meaningful "culturnomic" trends. In reality, the simplicity proved to be a strength. By simply analyzing the usage frequencies in the same headline dataset as before, trends became immediately apparent and digestible. For example, using n-gram visualizations it was possible to measure the frequency of words such as "woman" and "female" in the *Prince* dataset; in the end, the linguistic trends, shown in Figure 3 below, successfully matched the historical contributions of women following Princeton's gender

integration in the late Sixties. Unlike the sentiment analysis visualizations, the n-gram approach provided verifiable results that accomplished the original goal of elucidating trends in the text.



Most Positive Headlines

Year	Score	Link
1986	0.811	women want improvement
1991	0.806	surprise: intelligent, pretty grad women
1960	0.787	the great confidence men
1927	0.722	treat us like men

Most Negative Headlines

Year	Score	Link
1912	0.683	strikers trouble station men
1972	0.667	group fights prejudice against women
1984	0.667	men fail to keep margin low, destroy weak bruins, 73-40
1993	0.667	women complain of harassing calls

Figure 3: Linguistic trends for "men" and "women" match historical context

At the same time, the sentiment analysis work was not all for nothing. While sentiment analysis was not effective in illustrating macro trends over a century of data, it was effective on an individual basis, producing logical scores for individual headlines. So, sentiment analysis was used to supplement the n-gram visualizations. One of the major shortcomings of bare n-gram plots is the inability to understand the underlying causes for the apparent trends. While it is possible to parse overall changes in usage frequencies and observe peaks and slopes, it is more difficult to understand why these trends exist. For a given n-gram visualization, the goal, then, was to select a few of the key, original headlines to accompany and explain the visualization. To select which headlines to use, those with the most extreme – most positive and negative – sentiment analysis scores were chosen as effective samples, as evidenced in Figure 3 above. The

sample headlines component of the project represents an improvement on the Google n-gram project, allowing users to trace the trends back to the original sources. Finally, the final visualization interface was packaged into an online, user-interactive UI for distribution, which will be described in further detail in the following section.

4. Implementation

As a data mining project, a significant portion of the implementation centered around first collecting the necessary textual data and then developing an understanding of the dataset before creating the final visualizations. The implementation section below will walk through the process by which textual data was scraped from the archive's site, how the data was studied and analyzed from a development perspective, and, finally, how the project was packaged together into a user-friendly, public interface.

Scraping

Securing an accurate copy of the archive textual data proved to be the largest practical challenge at the beginning of the project. Though the original archive site was designed for casual browsing of individual papers from the past, it was not immediately obvious how one would efficiently and programmatically isolate the text from the archive site. Interfacing with the archive's front-end could limit the efficiency of visualization generation, with the bottleneck being the network latency. Additionally, there was a greater probability of error as scraping the front-end could potentially inject superfluous text into the dataset. A second possibility involved accessing the archive's backend database storage and obtaining copies of the source XML data files themselves. These files were generated from the OCR vendors themselves and are stored on the Princeton University Library servers. However, this approach would inevitably involve a large amount of database manipulation and data parsing. The source files were also suboptimal

because they contain a large amount of data that is irrelevant to the project, including the physical spacing of characters on a scanned newspaper page.

Because the data analysis relied only on the article headlines and not the entire article text body, a solution was to scrape the table of contents for each newspaper issue. This approach was different from scraping the archive front-end directly because the table of contents for each issue are generated through independent AJAX calls from the browser. By emulating these table-of-contents AJAX calls through Python's Requests library, it was possible to obtain clean copies of the article headlines.

To avoid all future network latency due to the archive site, all of the article headlines from 1876 to 2004, the span of the archive, were collected in a single batch and stored in a local JSON file. Inside the local file, headlines were stored in a key-value dictionary with the key being the date of an issue and the value being a list of all headlines for the given issue. Of note, a few known rogue headlines, such as "Advertisement" or "Untitled", were filtered and not included in the JSON file. Additionally, I casted all headlines to be only in lowercase to improve n-gram results. Because the dataset was already much smaller than Google's, lowercasing the headlines consolidates n-gram matches. For example, "Women" and "women" feed into a single n-gram visualization. After the complete data mining process, the resulting JSON file contained over 390,000 headlines in total, representing roughly 25MB of data.

Developing Visualizations

Because the project utilized a pre-made sentiment analysis library, the process of generating visualization charts was straightforward from a computational perspective. Everything from the preliminary "accuracy" charts, which will be discussed more thoroughly in the "Evaluation" section, to the sentiment and n-gram charts, simply involved iterating through

all of the headlines stored in the JSON file. In nearly all of the charts, with exceptions noted later, headlines were grouped by year. For example, to generate the n-gram charts, I calculated two major metrics for each year: the total number of headlines and the number of headlines containing a given n-gram. The majority of charts generated during the development process were created using Python's graphing library *matplotlib*.

Though the visualization algorithm ran in linear time with respect to the number of headlines, the performance of the scripts was reasonable. Regardless of visualization type, plots were generated within a few seconds. Comparatively, collecting the headlines themselves required over an hour of runtime. Thus, the cost of network latency far outweighed the runtime cost of iterating through all of the headlines in the dataset.

User Interface

The goal of the present study is to make historical and cultural ("culturomic") trends embedded within the *Daily Princetonian* archive more accessible to the general public and casual browser. Thus, the final component of the project involved opening the visualizations to public access through a user-friendly interface. To achieve this goal, the visualization scripts were wrapped into a Flask server, also developed using Python. A user of the server would submit keywords through URL parameters; the server would then respond with the customized visualization page. For example, to analyze usage of the terms "men" and "women" in the archive, one would access the following URL: 54.203.14.54:5000/?keywords=men,women; this result is shown in Figure 3 from the "Approach" section. When $n > 1$, the individual words of the n-gram query are separated by underscores, which are interpreted as spaces. Of note, the headlines listed in the UI are links that direct back to readable scanned copies of the original issue. An example of such a scan is shown in Figure 4 below.

Figure 4: Example scan of an original *Daily Princetonian* article

To improve general aesthetics, the front-end visualizations were produced using Google Chart's Javascript library, and the bare-bone CSS package Skeleton improved the appearance of the visualization site. The Flask server is currently hosted on an Amazon Web Services t2.micro instance. As visualization requests are processed in the order they are received, the server response time can reach upwards of one minute when 5-10 visitors are browsing the site. A future improvement would migrate the server to a higher-performance hosting service.

5. Evaluation and Discussion

Accuracy

After scraping the text from the archive site, it was important to verify the accuracy of the scraping results before proceeding with the visualizations themselves. Sources of data error in this project included everything from OCR inaccuracies to scraping mistakes, such as scraping sentences from the article text as opposed to the headlines or failing to scrape the archive in its entirety. Because the OCR accuracy was beyond the control of the project, the accuracy tests

described below are mostly sanity checks on the completeness of the dataset. It is important to note that developing these initial sanity checks improved my general understanding and knowledge of the dataset, which motivated the Examples that will be described later.

The first sanity check involved visualizing the number of issues that were published in each year, as reflected in the archive scraping data. Figure 5 below shows the number of issues plotted by year, once again using Python matplotlib library. From the figure, we see an expected dramatic increase in issues in the year 1890, which was the year the newspaper switched from a bi-weekly to daily publication. Additionally, the figure shows that the a disproportionately larger number of issues were published between 1900-1940. Thus, we should expect the visualizations to greater emphasize this time range. Since then, publication rates have steadied to a modern level of 150 issues per year.

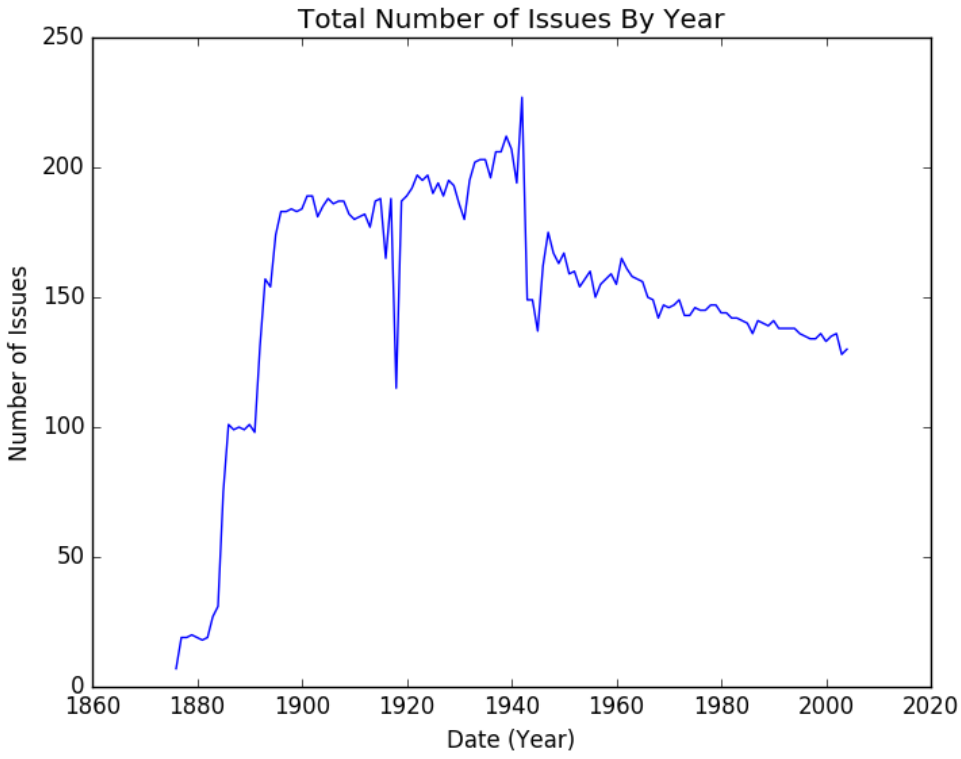
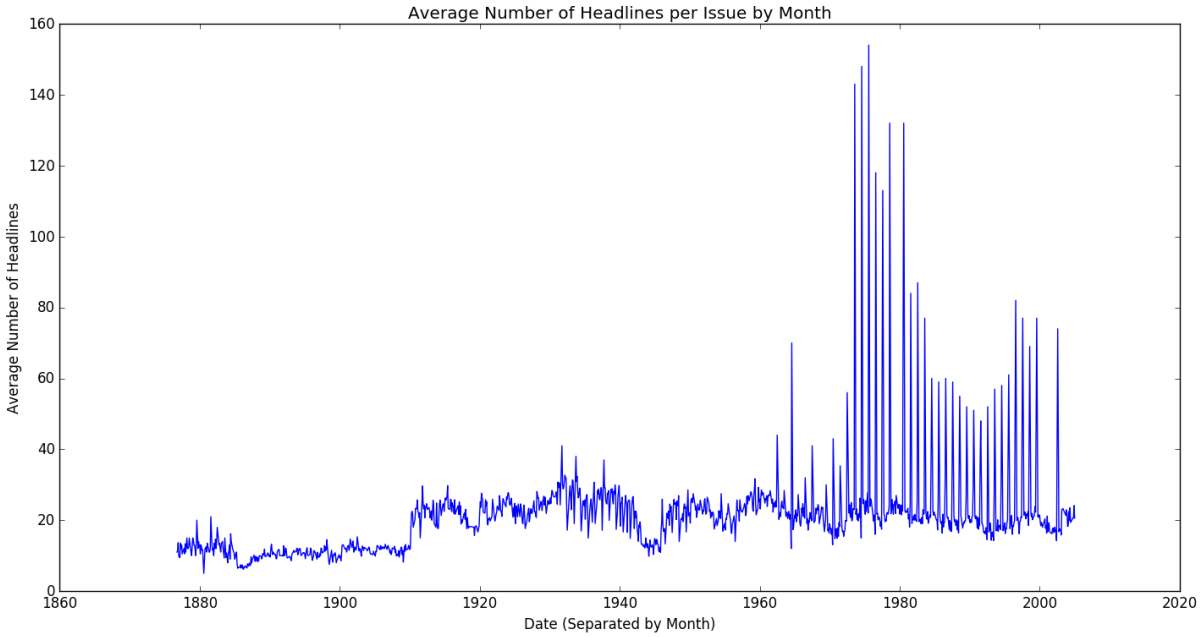


Figure 5: Accuracy check of the number of issues by year

Next, Figure 6 below plots the average number of headlines that were published in an issue on a monthly basis. This chart was originally created to check whether each issue in the scraping data contained a logical number of headlines. For the most part, the figure shows that the majority of issues averaged around 20 headlines, which matches expected results. However, the figure also includes seemingly unrealistic peaks beginning in 1973. While these peaks were initially interpreted as scraping errors, it was eventually discovered that the newspaper would in fact publish extremely long, composite "Class Issues" around reunions each year. The discovery of the Class Issues series is an example where sanity checks on the data led to a better understanding of the archive's content.

Figure 6: Average number of headlines per issue, by month

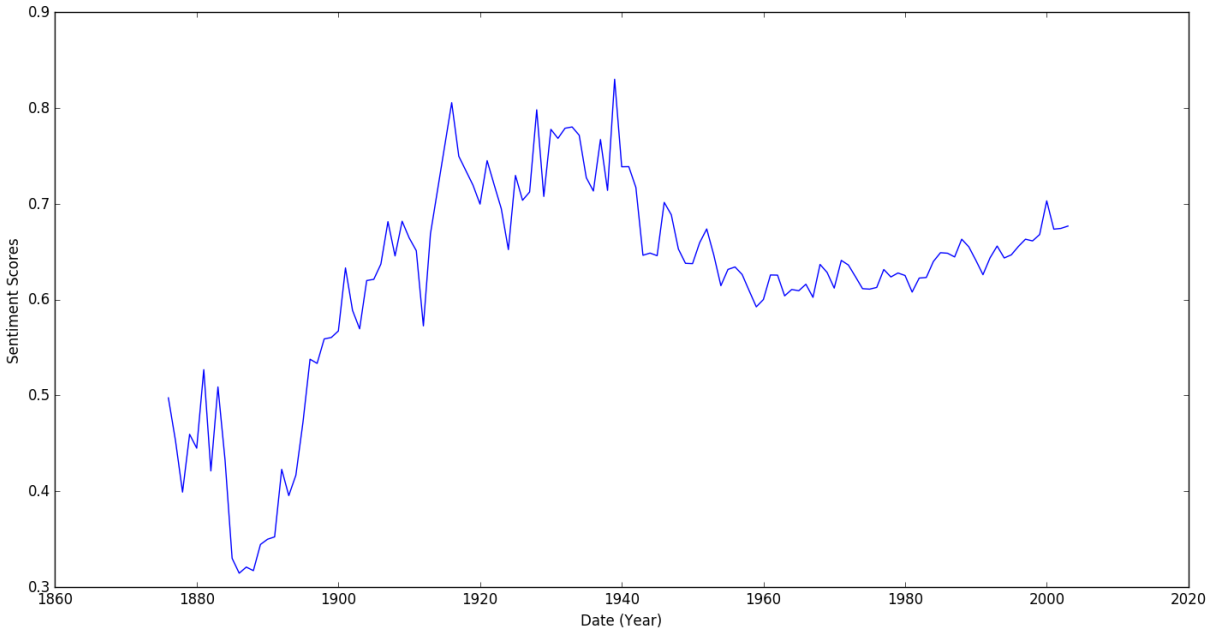


Sentiment Analysis Results

The following section will walk through the results from the sentiment analysis visualizations, including both the shortcomings and the strengths of the approach. The greatest

weakness of the sentiment analysis plots was the inability to measure and observe "culturnomic" trends. Should trends have been present, it was not possible to link them back to their source motivators. For example, Figure 7 below was one of the most comprehensive sentiment plots created during the project; the plot shows average neutrality scores by year, as calculated using VADER. As mentioned in the Approach section, prior studies suggested that the neutrality score would be insufficient in measuring bias. Nonetheless, it was hypothesized that the sentiment plot would still reveal trends in the attitudes of college-age students. While the figure does suggest notable increases in neutrality around 1900, for example, it was not possible to derive explanations without making unsubstantiated conclusions.

Figure 7: Comprehensive neutrality score plot spanning entire archive



On the positive side, the sentiment analysis results did confirm that VADER, though designed for modern text, could be effectively applied to older 20th century textual data as well. To test this issue, the most positive and negative headlines of at least 30 characters in length

were collected, as shown in Figure 8 below. In the figure, headlines are accompanied by their date of publication and corresponding positivity or negativity score. The results suggest that VADER correctly scored these headlines. Thus, although it was not possible to measure trends on a macro level using sentiment analysis, the technique was successful on a headline-by-headline basis. Subjectively, these superlative headlines comprised an interesting sample of the larger headlines data set.

Figure 8: Most positive and negative headlines in the archive

```
Most Negative Headlines
19320324        contemporary comment fight, fight, fight! 0.802
19811021        race relations: ignorance, insults, insensitivity 0.802
19870220        dorfman criticizes chile's violence, condemns government repression 0.805
19240228        lies, damn lies, and statistics 0.806
19830405        sexual harassment: ignored, problems persist 0.81
19960306        suspicious fire destroys restaurant's garage 0.81
20030421        mistakes make capital punishment unacceptable 0.813

Most Positive Headlines
19780519        nine pranksters share award for best joke 0.801
19621128        blaik prize winner: solid, articulate 0.813
19720911        schier wins top scholarship prize 0.815
19991028        respecting a daunting challenge 0.818
19820916        colonial gains strength with increased support 0.839
19090317        definite inter-club agreement. 0.841
19790413        promotional wisdom: everybody loves a winner 0.845
```

Strength of N-grams

Compared to the sentiment plots, the n-gram visualizations successfully revealed "culturnomic" trends embedded in the *Prince* archive text. The greatest strength of the n-gram approach is that all of the results are relative. Meaning, it was possible to study both obscure terms as well as more ubiquitous ones because the frequency of each term is only compared to itself. Additionally, because all of the terms were evaluated using the same, simple method, it was possible to validly compare the n-gram plots of several terms at the same time. Finally, because the n-gram metric is represented as a fraction or a percentage, this approach is more

resistant to noise in the data than the sentiment scores. For example, if one were to add superfluous, meaningless text to the end of each headline, a possible scraping inaccuracy, the n-gram calculations would still reflect the same trends. On the other hand, the sentiment approach would be impacted by extraneous noise in the dataset, measuring the sentiment of the noise in addition to the clean data.

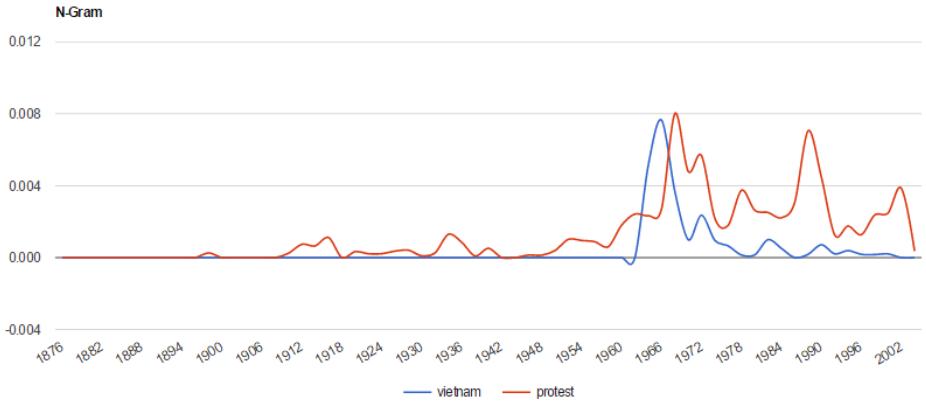
The n-gram approach was also favorable because the visualizations are user customizable. Whereas the sentiment plots are static and generated once, the n-gram visualizations are subject to the user's interests. The flexibility of the n-grams also increases the breath of the project. Instead of only analyzing trends in a particular area, such as bias, the user is able to study patterns in everything from history to linguistics. While the project initially intended to focus on a specific, pre-determined subset of topics, including gender and racial discrimination, the openness of the n-gram interface proved to be a strength.

Finally, the n-gram visualizations were overall better suited for a large, historical dataset. Sentiment scores, on the other hand, are not time dependent and can be run on a small sample size of text. In contrast, n-gram visualization leverages the size and time-scale of the data to its advantage. Along this train of thought, the n-gram visualizations would actually be more complete if article body text was used in addition to the headlines, but this would undermine the runtime performance of the UI and nullify the contributions of the sentiment scores, which cannot be applied to paragraph-length input texts.

Examples

In support of the project's ability to visualize "culturomic" trends in the *Prince* archive text, we will now walk through two major examples that highlight this claim. The first, shown in Figure 9 below, illustrates the usage frequency of the words "vietnam" and "protest" in the

Prince archive, with a few example superlative headlines shown. This first example effectively captures both social and historical trends related to the Princeton campus. As expected, the n-gram distribution for "vietnam" reaches a peak in 1966, roughly the peak of the historical war itself. Interestingly, the usage frequency of the word "Protest" parallels that of Vietnam. Though the amplitude for the "protest" n-gram is larger, its fluctuation is the same.



Most Positive Headlines

Year	Score	Link
1966	0.592	thanks from vietnam
1982	0.503	vietnam not enemy, falk, others contend
1966	0.467	peace organizations plan vietnam march
1968	0.461	object: record protest with votes peace freedom group starts drive
1965	0.444	student liberal organization to hold rally supporting peace in vietnam

Most Negative Headlines

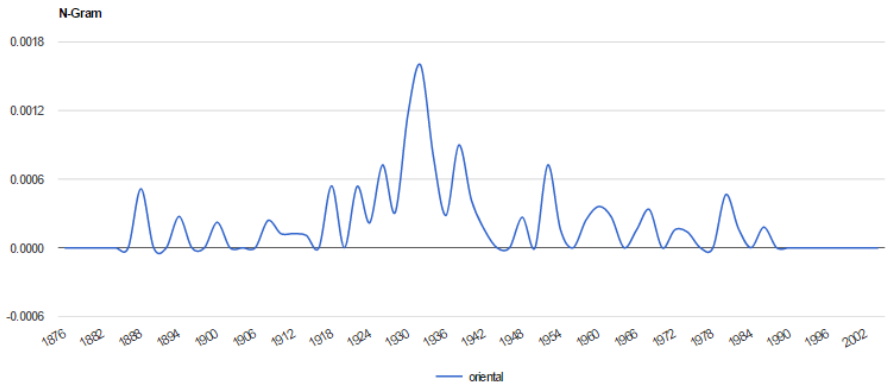
Year	Score	Link
1962	1.0	bomb protest
1978	1.0	protest
1979	1.0	protest complaint
1987	1.0	protest
1974	0.821	protest and repression

Figure 9: Final UI showing results for search terms "Vietnam" and "Protest"

One could argue that these n-gram distributions only restate known facts about the Vietnam War and college activism during the time. However, similar to the Google n-gram project, these charts better *quantify* the trends in a visual manner. For example, from the charts, one can determine in that in 1966, nearly one in a hundred headlines in the *Daily Princetonian* related to the Vietnam War.

The second example illustrates a more social linguistic trend within journalism and society. Namely, in Figure 10 below the results for the term “oriental” are shown. From the n-gram distribution, one can clearly see that usage of the term reached a peak in the Thirties, but has diminished since then. Now, the term “oriental” has become an artifact of the past. Compared to the previous example, the trend is less popularly well-known, but just as prominent, nonetheless. Additionally, the sample headlines are particularly useful for the “oriental” keyword, showing instances of the term’s usage from a linguistic perspective. The superlative – most positive and negative – headlines show that “oriental” can refer to anything from oriental languages, to oriental flu, to oriental luxury, with a time period associated with each of these usages.

Figure 10: Final UI showing results for search term “Oriental”



Most Positive Headlines

Year	Score	Link
1986	0.783	oriental romance
1949	0.457	'assurance of peace' essential to iran, states oriental ruler
1926	0.385	oriental comic opera one of triangles best "samarkand," eastern romance of love and intrigue fascinates opening night audiences. 2nd performance tonight novelty of scenic effects contributes to general satisfactory tone of whole performance.
1940	0.3	oriental languages 407 course performs neat disappearing act
1919	0.275	bishop roots discusses problems of far east says america can help oriental relationships by ensuring fair treatment in courts here.

Most Negative Headlines

Year	Score	Link
1968	0.394	student contracts rare oriental flu
1953	0.286	yale social life combines oriental luxury and darkest secrecy
1932	0.278	dr. dennett to explain oriental problems at 8:45
1927	0.266	speaker attributes rising oriental hatred to western fickleness in breaking promises
1975	0.204	upsurge in oriental applications provokes 'encouraging' deviation from national trend

6. Conclusion

To summarize, this project provides quantitative visualizations of historical, linguistic, and cultural trends reflected in the *Daily Princetonian* archive. The addition of "superlative" headlines also allows viewers to trace these trends back to specific articles – a sample of the larger corpus. The project advances efforts to not only digitize but also computationally analyze the data preserved in the University's vast archive. Specifically, the interface developed in this project can be used by scholars interested in Princeton-specific history and those investigating trends related to college campuses in general. Additionally, the project's results may be useful to the newspaper itself. Even the "sanity" accuracy checks can guide the newspaper's future leadership in how the paper has functioned in the past; for example, the visualizations in the Evaluation section clearly show the instances of "Class Issues" better than a plain, textual search engine. Furthermore, the visualization UI can even help guide today's newspaper writers as they research occurrences of similar events in the past.

Future work on the project should seek to improve the overall user's ability to discover culturomic trends of interest. In the current state, the discovery process is unreliable in that many search terms yield inconclusive trends. As such, it would be worthwhile to develop a metric that measures the quality of an n-gram distribution. Such a metric could be paired with a recommendation system that suggests similar terms to query. One could utilize the WordNet project to implement such a recommendation system. Finally, the project would also benefit from an expanded dataset, perhaps one that spans even more college campuses.

Acknowledgements

First and foremost, without the counsel and guidance of my adviser Dr. Brian Kernighan and teaching assistant Meagan Wilson (GS), this project would not have been possible. The two devoted generous time and effort to provide candid feedback throughout the semester. Additionally, their genuine interest and specialties in the field of Digital Humanities served as additional personal motivation in improving the project.

Finally, I'd like to thank Center for Digital Humanities staff members Dr. Jean Bauer, Dr. Clifford Wulfman, and Matthew Ritger (GS). The three provided logistical support for the project, assisting with brainstorming and redirecting to the appropriate resources. Having the support of this community made the semester project both challenging and rewarding.

References

[1] Liu, Bing, and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data.*, 415Springer US.

[2] Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. *Icwsm* 7 (21): 219-22.

[3] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

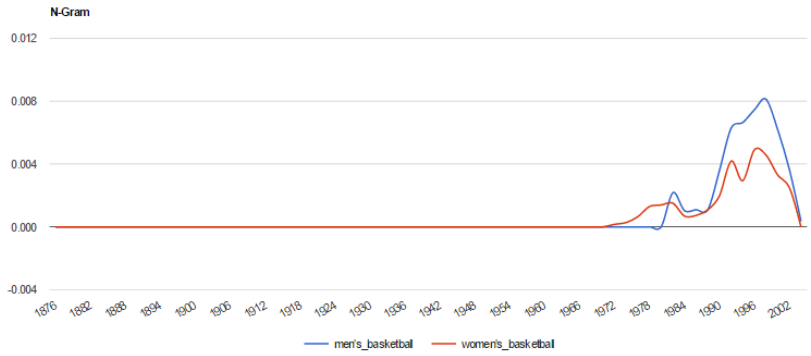
[4] Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. *Quantitative Analysis of Culture Using Millions of Digitized Books*. **Science** (Published online ahead of print: 12/16/2010)

[5] Reisman, Dillon. . *All the News That's Fit to Change: Insights into a Corpus of 2.5 Million News Headlines*, Edited by Joel Reidenberg. Princeton University: Center for Information Technology Policy, 2016.

Appendix

Because the article body could not contain all of the notable visualizations, the remaining examples have been compiled here for the reader's interest. Each of the visualizations below is accompanied by an interpretation of the results and their greater significance.

Figure 11: Discrepancy in coverage between gendered sports



Most Positive Headlines

Year	Score	Link
2002	0.709	men's basketball hopes persia's 80-foot miracle will spark success
2001	0.573	women's basketball gains momentum from exciting first win of season
1997	0.524	all-ivy men's basketball honors
2002	0.508	women's basketball hopes to continue climb to respectability

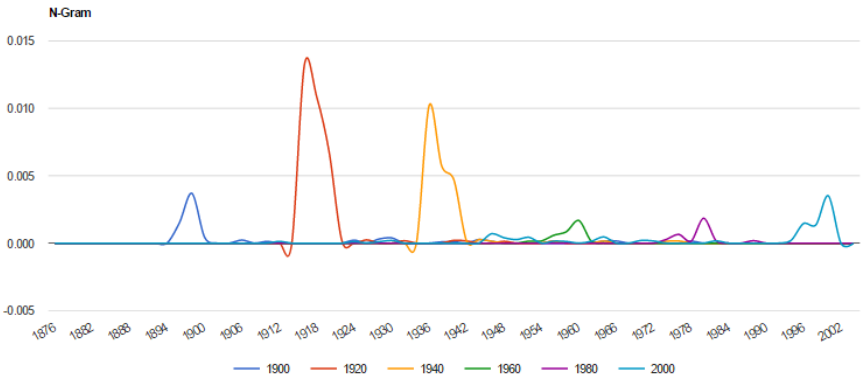
Most Negative Headlines

Year	Score	Link
1992	0.588	women's basketball struggles in disappointing campaign
1994	0.577	men's basketball falls victim to poor shooting in defeat
1996	0.524	sports shooting problems plague struggling men's basketball
1997	0.504	women's basketball suffers 10th defeat at st. peter's

The above n-gram and headline visualization shows coverage of both men's and women's basketball. From the n-gram, we can see that neither term was used before the Seventies because the gender was not ambiguous before Princeton became co-educational. Furthermore, searches for basketball, soccer, lacrosse, and swimming all reveal that women's sports received greater coverage than men's sports in the Seventies, whereas the trend either flips or narrows in modern times. It is also possible that the women's sports did not necessarily receive more coverage, but were identified with the "women" keyword more often whereas

men's sports did not require a gender associated. In either case, this example, in particular, illustrates the comparison power of n-grams where n > 1.

Figure 12: Analysis of the perception of time



Most Positive Headlines

Year	Score	Link
1900	0.592	1900 football championship.
1937	0.552	purnell wins 1940 post
1976	0.5	welcome class of 1980
1936	0.487	1940 stickmen win by default

Most Negative Headlines

Year	Score	Link
1936	0.467	1940 amateurs fail to appear
2001	0.467	campus crime swells in 2000
1919	0.438	issue 1920 preliminary war records to-day
1937	0.412	rain postpones 1940 golf

This quirky example is directly inspired by Google's original n-gram experiments. The researchers studied society's focus on the present and dismal of the past by visualizing the mentions of specific year numbers, such as "1883". The Google researchers concluded that society rarely looks in the future, years are not mentioned before they occur, and then attention on past years decays over time. A similar, though less pronounced, phenomenon occurs in the *Daily Princetonian* data set, with one caveat: from the visualizations and sample headlines it's clear that the years usually refer to graduation years. For example, the peak for the n-gram "1920" actually takes place in 1916. This example is a reminder to understand the nature of the underlying corpus before making direct inferences from the n-gram visualizations.