Where are the Women in *The New York Times*?
A cluster analysis of newspaper articles

EXAM NUMBERS: 88, 123, 128 AND 170

Københavns Universitet
August 31, 2018

Abstract

This paper describes the collection and analysis textual data from The New York Times API, using the powerful tools made available by Python.

Keywords: data mining , text analysis, social data science, python

Contents

Contributions:

- Exam no 88: section 1.1, 3.1, 4.4
- Exam no 123: section 1.2, 4.1, 4.5
- Exam no 128: section 1.3, 4.2, 5
- Exam no 170:section 2, 4.3, 6

The introduction has been written in collaboration between all contributors

Introduction

In 2008, a study found that less than 13% of Wikipedia editors were women, even though about half of readers were women (Ruediger Glott and Ghosh, 2010). This has been highly criticized given it results in a gender bias in terms of articles on the webpage: There are simply too few articles about significant women, compared to their male counterparts. In 2011 The New York Times published an article titled "Where are the Women in Wikipedia?". It argues that women feel not welcome in the environment citing "it's a popular stereotype that men are believed to know more 'hard facts', while women are better at nurturing and getting along with people" (Herring, 2011).

Analyzing The New York Times and inspired by research done by Wikipedia's contributors, we intend to explore gender dynamics in terms of published articles in the newspaper; how has the publication in The New York Times changed over the past 30 years? E.g., which topics are often discussed or has the distribution of topics changed over time? Who is the article's writer, which gender(?) and how does this affect what the article subject? Finally, has the intersection between topic and gender changed? These sort of questions will be examined.

Based on data from The New York Times' API, from 1990 to present, our paper analyses several aspects of the development, first analyzing the similarity between the articles, and using this information to draw conclusions concerning publication figures and the texts' complexity across gender.

1 Data collection

1.1 Data Source Description

The study investigates newspaper articles from The New York Times. This implies several practical and methodological advantages: First, an API key for the necessary data can easily be obtained, free of charge. Secondly, the API contains metadata on all published articles dating as far back as 1851. This is especially crucial given that our focus centers on the newspaper's development over time. Nevertheless, one data restraint exists; only few articles can be scraped in full length due to a pay wall. For this reason, we only have the headline and the snippet of each article (i.e. a short summary

of the article) to work with.

1.2 Obtaining the data

Given that The New York Times publishes hundreds of articles each day, it is only feasible to extract articles from a limited time span. Consequently, the period 1990-2018 is chosen. In addition, we cannot use data from every single day due to the same reason. Thus, we collect articles from every second Monday in the beforementioned time span. The weekday was chosen arbitrarily since January 1st 1990 happened to be a Monday. Before requesting the data from the API, we construct a list of the dates from which we intend to collect data. Secondly, we wrote a function to loop through each page of articles for each date in the list. We saved the data from each date in a .json file, which subsequently is merged

1.3 Ethical considerations in data collection

Since the data is obtained from an open API, potential ethical issues regarding the data's publicity do not seem present - all the information in the data is publicly available elsewhere. The only personal data in the sample is the names of the authors. However, the authors chose to publish the articles, thus, they voluntarily linked themselves to the respective article. Nevertheless, when parsing data at this scale, new information about the authors could become apparent, which would not otherwise be publicly available. Consequently, author names are excluded in the analysis, and is only used in the data processing to extract the gender.

While The New York Times API is open, one is required to register for an API key, and there is a limit of 2,000 requests per day. In order to obtain all the required data, much more than 2,000 requests was necessary. Therefore, we registered several times, using different email addresses. The request limit is indeed set up to avoid server overcrowding but considering that we are just collecting the data once, for research, and not on a regular basis, as an app or a webpage would, the ethical issue seems modest. In addition, we set up our data collection function to only perform one request per second.

2 Data Cleansing

Some data cleansing was required in order to progress with the data. This paper analyses headlines, snippets and gender of the author, why all the articles that did not have a valid headline, snippet or author name was sorted out. Also articles, where the gender could not be identified from the author's name were excluded. Our data cleansing is visualized in figure 1.

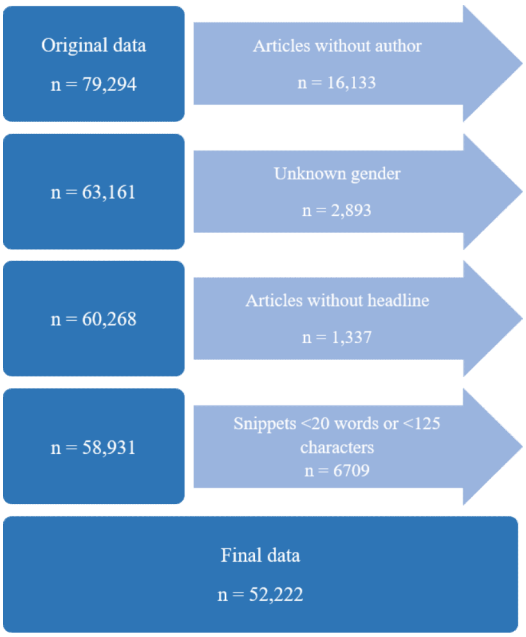

Figure 1: Data cleansing process

We utilized a pre-made lexicon to decide gender from the first names (GenderGuesser). The lexicon provides six values: 'male', 'female', 'mostly male', 'mostly female', 'androgynous' and 'unknown'. We chose to assign 'male'/'mostly male' as male and 'female'/'mostly female' as female. Given relatively few names were androgynous or unknown, these were excluded rather than trying to guess the gender. Additionally, given snippets are being used for deriving article similarities, we left out snippets of short length. We found that the max snippet consisted of 250 characters and we decided that for an article to be valid it should consist of either 20 words or 125 characters. Both word count and character count is included so that meaningful snippets with 19 words or less than 125 characters are not discarded. Combining both words and character count results in a better sorting. Otherwise, this would possibly distort the analysis. E.g. it was

noticed that several snippets contained nothing but the word "chronicle". Having excluded articles that do not meet our criteria, our sample is not necessarily representative of all things published by The New York Times since 1990. However, when examining a small sample of the articles we are sorting out, we found that they were mostly types of articles that we are not particularly interested in. For example, short comments, quizzes or crosswords. Thus, we will move on with the sample containing 52,222 articles.

3 Methodology

3.1 Tf-idf and *k-means*

To explore the development of articles published by The New York Times, we are interested to categorize the articles, to explore how many articles are published within various categories. We want to explore the similarity between the articles using the data in the snippets, because the metadata from The New York Times API does not contain categories that go across all the years. We will perform a cluster analysis in order to categorize the articles.

First, we tokenize and stem the data in the snippets and then calculate the term frequency-inverse document frequency (tf-idf). This method assigns high weights to words that occur often, and low weights to less frequently occurring words.The tf-idf is a product of the term frequency and the inverse term frequency.

$$\text{tf-idf (t,d)} = tf(t,d) \text{ x idf (t,d)}$$

The term frequency is the number of times a term occurs in a document. Whereas the inverse term frequency is the log of the total number of documents divided by the number of documents containing a certain term (Raschka, 2015, 261f).

$$\text{idf (t,d)} = log\frac{n_d}{1 + \text{df(d,t)}}$$

The tf-idf vectorizer is basically a frequency matrix, displaying the frequency of each term in each document (cf figure 2).

In our tf-idf vectorizer we have further specified a parameter to exclude words which are reoccurring in more than 80% of the snippets, assuming

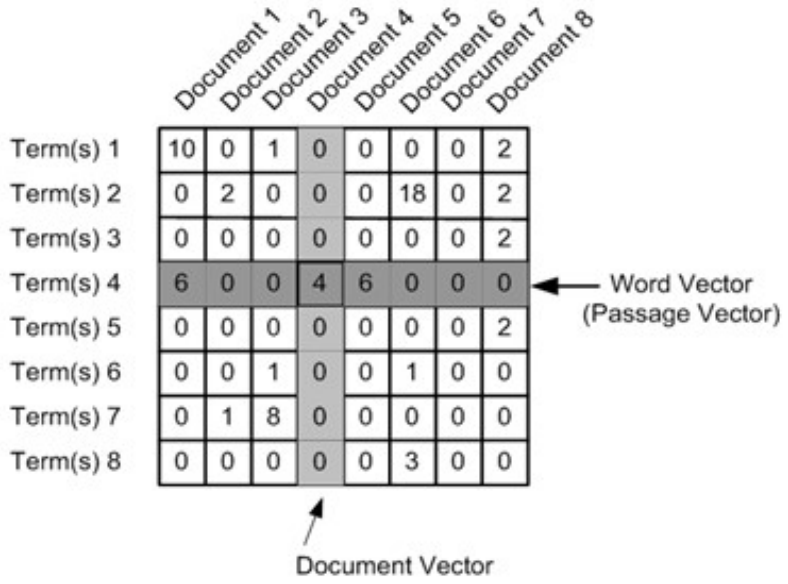

Figure 2: TF-IDF Matrix

that these words have no real significance as they are being used very often. We have also specified a list of stopwords, that are not included. Further, we have not included terms that are only present in less than 3 of the snippets.

The frequency matrix is then used to perform a cluster analysis of the snippets. We have chosen to use the *k-means* algorithm to structure the tf-idf matrix. We ran the *k-means* several times, with a different number of clusters each time. We started, rather arbitrarily with 10 clusters. When we found that some categories were very small, we tried with 8 clusters. However, 8 clusters were harder to distinguish from each other so we tried again with 12 clusters. The 12 clusters returned relatively distinguishable categories, but there was still a large category that seemed to contain many unrelated words. Thus, we also tried 14 clusters but then found that several categories that seemed related, were being split into two. Finally we ended up using the 12 clusters.

Rather than choosing a number of cluster based on comparing the results of different numbers, a method exist to compute the optimal count. The elbow method estimate the optimal number of clusters. The method is simple, however, we performed the method but our computer did not have the required memory to calculate the results. Therefore, we stuck to our original method of choosing the number of clusters.

Now, each article is linked to a cluster such that the clusters' sum of squares are minimized between two sample points, w_k and x. Formally, this is written as

$$d(x, w_k)^2 = \sum_{j=1}^{12}(x_j - w_j)^2 = ||x - w_k||_2^2$$

Here, index j represents the jth dimension, i.e. cluster and h is the estimation number (before global optimum is located) (Raschka, 2015, 350ff).

Now, *k-means* repeatedly recalculates, thereby learning by batch updating, the mean of the clustered observations as the words are reassigned to other clusters, thus, changing the center point until the optimal centroid (average) is identified, i.e. the minimizing the sum of squared errors/quantization error:

$$SSE = \sum_{i=1}^{n}\sum_{i=1}^{k} w^{(i,j)}||x^{(i)} - \mu^{(j)}||_2^2$$

where μ^j refers to the representative point (centroid) for cluster j, while $w^{(i,j)}$ is a binary variable equaling 1 if the sample x^i is in cluster j. Otherwise, $w^{(i,j)} = 0$. So, the smaller the SSE, the tighter are the clusters. As a result, the data underlying structure is revealed. In more formal terms, the algorithm's optimization problem is solved by a so-called Newtonian gradian descent Bottou and Bengio, 1995 in which the temporal estimations are called 'prototypes', k. The updating element regarding w is given by

$$\Delta w = -\epsilon_t \frac{\delta d}{\delta w}$$

Here, ϵ_t represents the learning rate.

4 Results

4.1 Descriptive Graphics

To explore The New York Times' articles' development, we counted number of articles published on each day (i.e. every second Monday), then calculating the annual means.

The first plot shows the development in the amount of articles published each day. Firstly, we see a large spike articles published around 2001-2004.

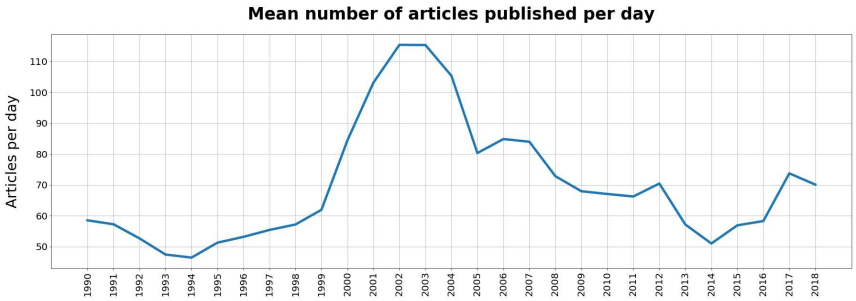

Figure 3: Number of published articles

Afterwards, the amount of published articles returns to around the same level as in the 1990s. A small upwards trend is seen. Overall, articles published per day has increased since the 1990s.

To further explore the development, we split articles by gender. The following figure shows the development in the proportion of articles written by women. While overall share of articles written by men far exceeds

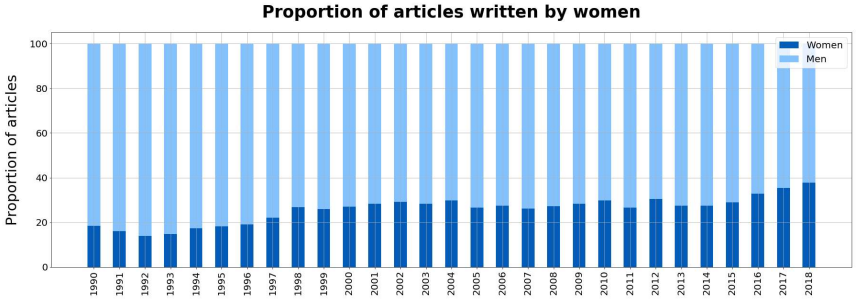

Figure 4: Proportion of articles written by women

women's share, it is also clear that the proportion of articles by women generally has grown over time. The growth, however, is not completely stable and constant, waving a bit during the first years of the 1990s and early 2000s.

4.2 K-means

Table 1 and 2 shows the top words for each cluster. Based on the words we have categorized the clusters according to the content. We found that one cluster was much larger than the rest, and contained words that were difficult to categorize (i.e. *Other*, n=23,609). To further explore the contents

Table 1: Top words in each category (part 1)

Cluster	New York	Law	Sports	U.S.	Business	Science/health
	New York City	Court	Game	United States	Company	Study
	Mayor	Supreme Court	Team	Official	Executive	Research
	Michael R. Bloomberg	Rule	Season	Team	Chief	Suggest
	School	Appeal	Coach	Government	Corporate	Risk
	Official	State	Player	Olympics	Billion	Report
	Council	Justice	League	President	Business	Cancer
	Rudolph W. Giuliani	Federal	Yankee	Trade	Market	Disease
Count	2393	656	5291	1824	3109	1490

Table 2: Top words in each category (part 2)

Cluster	Crime	Election	International Politics	Bush	Post Bush	Other
	Police officer	Senate	Government	Bush	President	Show
	Kill	Campaign	Official	President	Clinton	Work
	Arrest	Democrat	Federation	White House	Obama	Business
	Shot	Republican	Charge	Administration	Vice President	Market
	Woman	Candidate	Nation	Clinton	Trump	Way
	Police department	Party	Prime minister	Republican	Former	Ago
	Brooklyn	Election	Security	Iraq	Hillary	Come
Count	1370	2941	6382	1041	2116	23609

Note: Column 2-5 can be categorized as Politics, we examine the subgroups. Column 4 and 5 are merged as Washington.

of this specific cluster, we have performed a separate *k-means* on this cluster. The top words are displayed in Table 3.

The second *k-means* analysis returned some categories that were not visible in the first analysis, i.e. *Entertainment* and *Travel*. Some words that seem very related to clusters in the first analysis, have been clustered in the second, i.e. *Business*. But there is still one large category which is non-distinguishable. Performing the second *k-means* proved to have little effect.

The categories vary in size but they all contain more than 500 entries (i.e 1% of the sample). The categories discovered in the second cluster analysis contain enough entries, that we consider it relevant to include them in the categorization of the articles.

Table 3: Top words 'Other' category

Cluster	Business	Other	Entertainment	Entertainment	Travel
	Market	Show	Work	Advertisment	Made
	Bank	York	Film	Television	Travel
	Percent	Way	Artist	Aygenc	End
	Billion	Nation	Movie	Network	Airline
	Price	Home	Festival	News	Flight
	Stock	Ago	Art	Magazine	Airport
	Corporation	Open	Film festival	Account	Business
Count	2358	16578	1131	1855	1687

Note: From column 2 is easy to see that is even possible to go further in inspection and that there are some topics repeating.

4.3 Exploring development in categories

As our *k-means* analysis revealed some categories in the articles, we have grouped the articles in each category per year. Figure 5 shows the development in terms of how much space each category takes up in the total amount of articles published. One point worth noting is that the category

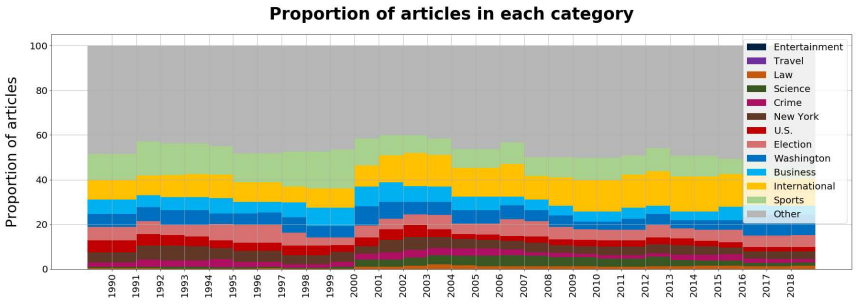

Figure 5: Articles in each category

of *Sports* seems to take up much less space after the millennium change than it did before. Similarly, *International Politics* is taking up more space from 2001 and forward. This might suggest The New York Times have

become less local oriented and instead more globally focused during the past 30 or so years. The change could be related to a more widespread diffusion of internet connection and to a more widespread fruition of NYT website across the world. Another category which has changed across time, is category *Science* and *Health*. In the 1990s, few articles featured science and health topics. After the new millennium, this significantly increases before quieting down again during the past couple of years.

Looking into specific categories and considering women's proportion of written articles, some interesting aspects are revealed. While we have examined plots of all the categories, we decide only to include the most important elements in the following.

Figure 6 shows that the proportion of articles about politics by female journalists is steadily improving. While only about 10% of political oriented articles were written by women in 1990, it is nearly 50% of today's articles. However, we see that there is a small but compelling decrease in 2015-2016. occuring around the presidential elections - exactly like in 2004.
It appears that when Democratic candidates faced seemingly difficult elections, The New York Times (a *liberal* newspaper that has always supported the Democratic candidate), fills the political section of male journalists. The opposite happens when the round is perceived as easy: in 2008 (Obama), 1996 and 1992 (Clinton) more women were writing about politics.

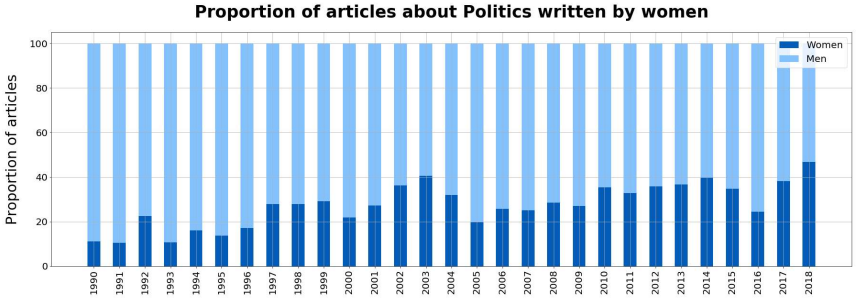

Figure 6: Politics articles written by women

Figure 7 shows the category where female writers take up the least space is *Sports*. Very consistently, the vast majority of *Sports* writers are male. Interestingly, we see that women are suddenly writing 20% of *Sports* articles in 2018. Thus,

Finally, Figure 8 shows another interesting category, when it comes to female writers: Law. While there are some years where none of the articles

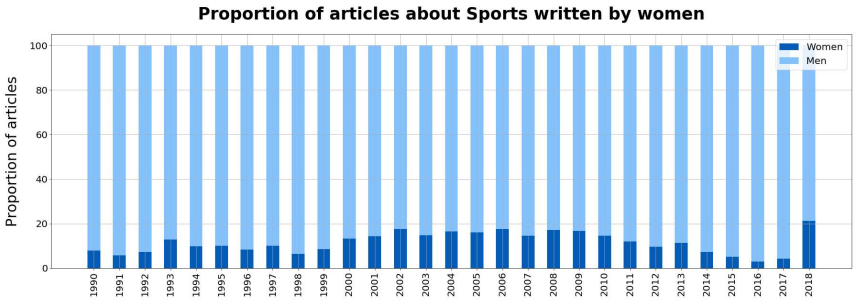

Figure 7: Sport articles written by women

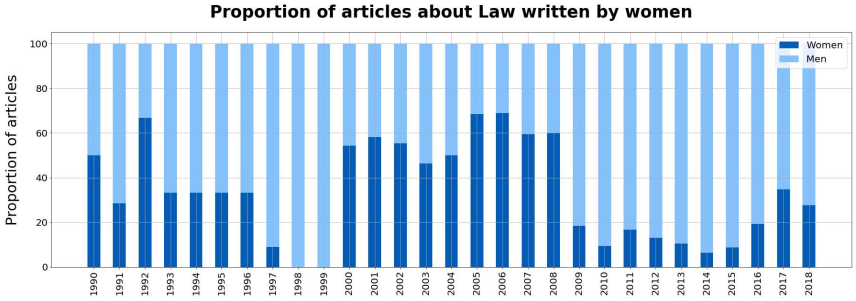

Figure 8: Law articles written by women

are written by women (1998 and 1999) there are other years, where the majority of the writers are female. Perhaps the reason why it seems to vary a lot, and women write many of these articles, is that this type of article often is very case oriented. One story will produce many articles. Thus, the same woman may have written many of these articles, resulting in this very large proportion of female writers in the law category.

4.4 Measuring Complexity

Another interesting aspect we inspect is the one related to complexity. Recalling what we cited in introduction, we want to explore whether the complexity of articles in The New York Times is dependent on the gender of the author. As previously mentioned, we did not have access to the whole corpus of full articles, so we scaled the snippets' complexity. To evaluate complexity of the text, we utilized the Dale–Chall readability formula, scaling the most commonly used 3000 English words according to

its lookup table and return a grade. The formula used in this method is

$$0.1579 \left(\frac{\text{difficult words} * 100}{\text{words}} \right) + 0.0496 \left(\frac{\text{words}}{\text{sentences}} \right)$$

We chose this formula, because it contains a dictionary of specific complex words, rather than rely on the length of the words in the text. In Figure

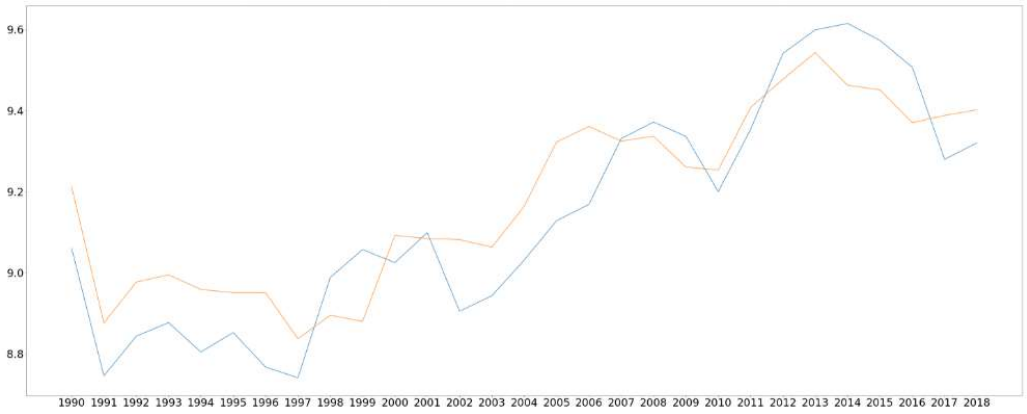

Figure 9: Average readability over years

9 the result of this analysis is displayed. It is possible to see that along years the readability of snippets has gone through a positive trend in readability, with some peaks followed by downward corrections, still remaining in the top difficulty area of the scale. The full table for interpreting the coefficient is provided in Table 4. Looking at the gender differentiation, where the

Score	**Difficulty**
90-100	Very Easy
80-89	Easy
70-79	Fairly Easy
60-69	Standard
50-59	Fairly Difficult
30-49	Difficult
0-29	Very Complex

Table 4: Dale–Chall readability scale

orange line is for male and the blue one for female, it is possible to see that women in 1990 were, on average, writing easier articles than men.

However in 1997 and 2011 men passed to use an easier language, and it remains so up until 2016, when male language get back being the more complex one.

Is there important to state, though, that the difficulty levels from male and female writers are never excessively far one from the other, and that we are still moving in the higher level of complexity for this method of evaluation.
It goes without saying, moreover, that we are only analyzing the snippets of the articles and that the analysis of the full text of the article could possibly bring completely different results, but definitively more reliable. Snippets are summaries of articles, so it will obviously contain more difficult, significant words than the full article.

4.5 Measuring Sentiment

Still wanting to explore gender differences, we performed a sentiment analysis on snippets. We wanted to explore various stereotypes about women, such as the one presented in our introduction: *women are more nurturing and better at getting along with people*, why they might be less likely to write negatively about other people. Another stereotype about women is that they are more emotional than men, thus another idea is, that women might write with more sentiment than men. To explore this, we have used the Vader sentiment dictionary in the *nltk* package, to assign an overall sentiment score to each snippet. The scores are standardized between 1 and -1, where 0 is neutral, 1 is very positive and -1 is very negative.

Figure 10 shows that The New York Times is, on average, a rather neutral newspaper, although there is a period between 2002 to 2007 (the very years of Bush's second mandate), where the sentiment seems to be more negative. When looking across the genders, there does not seem to be a major difference.

Going beyond yearly average and detailing more, in Figure 11 we have plotted the mean sentiment score per day for both genders.
This does indeed provide more insight. It is immediately noticeable, that women are writing with more sentiment than men, both negatively and positively. This tendency is very large during the 1990s, after which the sentiment for men and women converges through time and fluctuate in the same sentiment level for both genders. The 2001-2006 negative peak for women is visible also here, but is the only actual divergence from a path of calmness in language. There is one negative peak that stands out a lot in

Figure 10: Sentiment means over years

2014 for the women, although this is likely based on a few articles. But the graph does seem to contradict the stereotypes of women who want to be likeable, and not express themselves negatively.

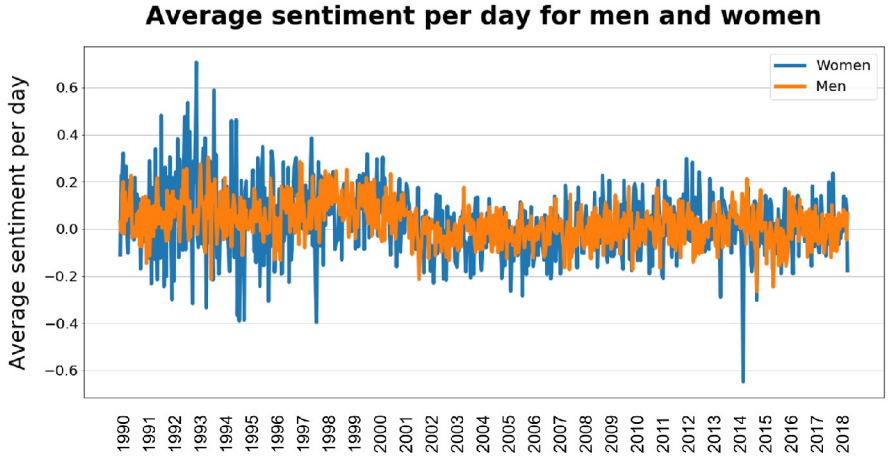

Figure 11: Sentiment per day

5 Discussion

Due to the New York Times' API request limitation, as well as the timeframe of this project, our study rests on a sub-optimal starting point, as we have not been able to collect all the data available for the time period we are examining. In addition, computational limitations simply did not permit the data amount to be any more extensive. As a compromise, data from every second Monday throughout many years, seems better than having data from every day for a shorter period. The fact that the data only contains metadata about the articles and not the articles in themselves, may also be a technical limitation. However, considering the timeframe and limited computer memory available, using the snippet is likely a better solution, than using full length articles, which would require much more processing time and memory. Again, we arrive at a compromise between the complexity of the data and the time period to examine. Finally it could have been very interesting to examine a much longer time period, since there was data available since 1851. However, if we were to do this, we would have to limit our data collection to very few days per year, which would make the data much less representable. In conclusion, it seems fair to presume that the actual collected data amount suffices in order to derive representative conclusions about the newspaper.

About methodology, it must be noted that although *k-means* is effective at identifying clusters with a spherical shape, the cluster amount is set by the researcher, which inevitably leads the possibility of a not realized better result. To counter this, we ran the process with several different pre-set number of clusters. Finally, it can be added we didn't ran into issues with an empty cluster which otherwise might show up in k-means.

Furthermore, it is important to note that *k-means*, as any other cluster analysis, is a means of generalization, that will not reflect the true variation in the data. Furthermore, when we decide to name a cluster which contains certain words, we are interpreting a meaning behind the common words in the cluster, again generalizing without noting variation in the data. As an example, as we have used both unigrams, bigrams and trigrams in the cluster analysis, the New York cluster included both Bloomberg and Michael Bloomberg. Some articles containing the word Bloomberg may be about finance, while others may be about the former mayor of New York. Because of the scope of the data, we are unable to distinguish between these.

Finally, we have to address the quality of our clusters. The quality is

questionable, since we find this large cluster, which we have named "Other" because we are unable to distinguish what defines it. This could indicate that the clusters are not clearly separated. To evaluate the quality we could have used silhouette analysis, but once again our computers did not have the capacity to perfom such an analysis. Thus we cannot conclude on the quality of the clusters.

6 Conclusion

Overall, times have changed for The New York Times, although not too much. Women are writing an increasingly larger part of the articles. However, men are still writing the vast majority.

Women do not seem to write about traditionally manly topics such as sports, but the share in more serious topics such as politics. The political articles proportion published by women took a dip around the presidential election in 2015 and 2004 - both of which a Republican candidate won the election.

The analysis regarding snippets' complexity shows that The New York Times is a highly academic newspaper, albeit it has become slightly more easy to read since 1990. When examining the difference in complexity across gender, it seems that men and women, more or less, are writing articles of similar complexity. The reliability of this analysis, though, is questionable since the only part of the article that was analyzed were the snippets. This limits the scope for evaluating the complexity of a whole newspaper article.

Finally, the sentiment analysis revealed that the stereotype of using a more emotional language proves to be confirmed in the first year of our dataset coverage, but then the language for both genders converged to a more neutral one with the only exception of the years of Bush's second mandate. It is interesting to notice, also, that the switch to a more convergent linguistic register for men and women coincides with the same period in which The New York Times ceased to have a more local point of view and started to see itself like a worldwide landmark for information.

This paper has uncovered a problem with the gender dynamics at The New York Times, but future research could focus more on the reasons why. Furthermore, it could be interesting to do similar research looking at different papers with other political standings or papers from other countries.

References

Ruediger Glott, Philipp Schmidt and Rishab Ghosh (2010). *Wikipedia Survey – Overview of Results*.

Herring, Susan (2011). *Where are the Women in Wikipedia*.

Raschka, Sebastian (2015). *Python Machine Learning*. Packt Publishing.

Bottou, Leon and Yoshua Bengio (1995). "Convergence properties of the k-means algorithms". In: *Advances in neural information processing systems*, pp. 585–592.