Practice 3: Gutenbergr project.

Histograms

Mayra Cristina Berrones Reyes 6291

September 22, 2020

1 Introduction

The Gutenberg Project is an online library that offers its users free access to more than 60,000 free books in different formats that go from ebooks, html, plain text, etc. It began with creator Michael Hart in 1971 at the Materials Research Lab at University of Illinois [3].

In previous work, we explored some of the features of libraries for text analysis such as `gutenbergr, tidytext, tm,` and, `dplyr`. Going deeper in analysis we now explore the sentiment analysis of texts using the library of `tidytext` to perform other experiments using histograms as a tool.

2 Books

Again for this project we use a book from the acclaimed novelist Jane Austen [1], *Pride and Prejudice* [4] published in 1813. And to make contrast in the present study, we are going to be comparing results with another book from another famous female author from the 19th century.

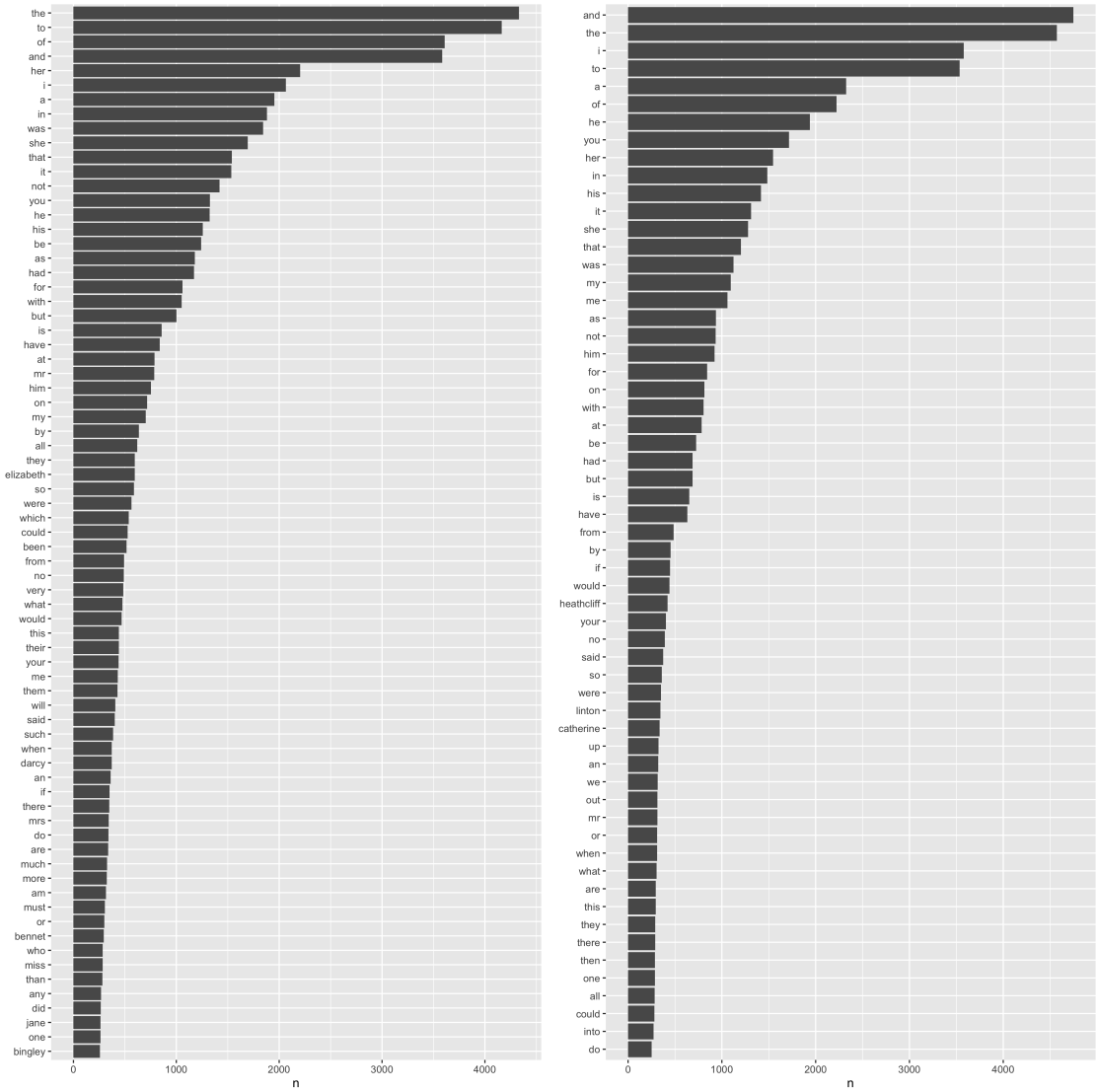
Wuthering Heights [8] is a novel by the author Emily Bontë [2] published in 1847. Her work is now regarded as classic in English literature, but in her time, the way she depicted mental and physical cruelty so attached to the love story she constructed in her book, challenged several Victorian ideas about religion, morality and class.

This two books are so alike in the way it changed public perception about class and the woman's place in society in their time, but their themes and tone of writing sound and feel to the reader so polarizing. So it gives us the focus for this experiment. Using the sentiment analysis of the library `tidytext`, we work to find the frequency of positive and negative words in both books.

3 Data exploration

To begin the exploration of the data, we start by tokenizing the words of each work. In Figure 1 we see that there is not much difference in the use of what is known as stop words. The main

distinct feature of this, is that in the work of Brontë Subfigure 1b we have more of this connective words than in Subfigure 1a which is the Austen book.



(a) Pride and Prejudice.

(b) Wuthering Heights

Figure 1: Frequency of all words in both books.

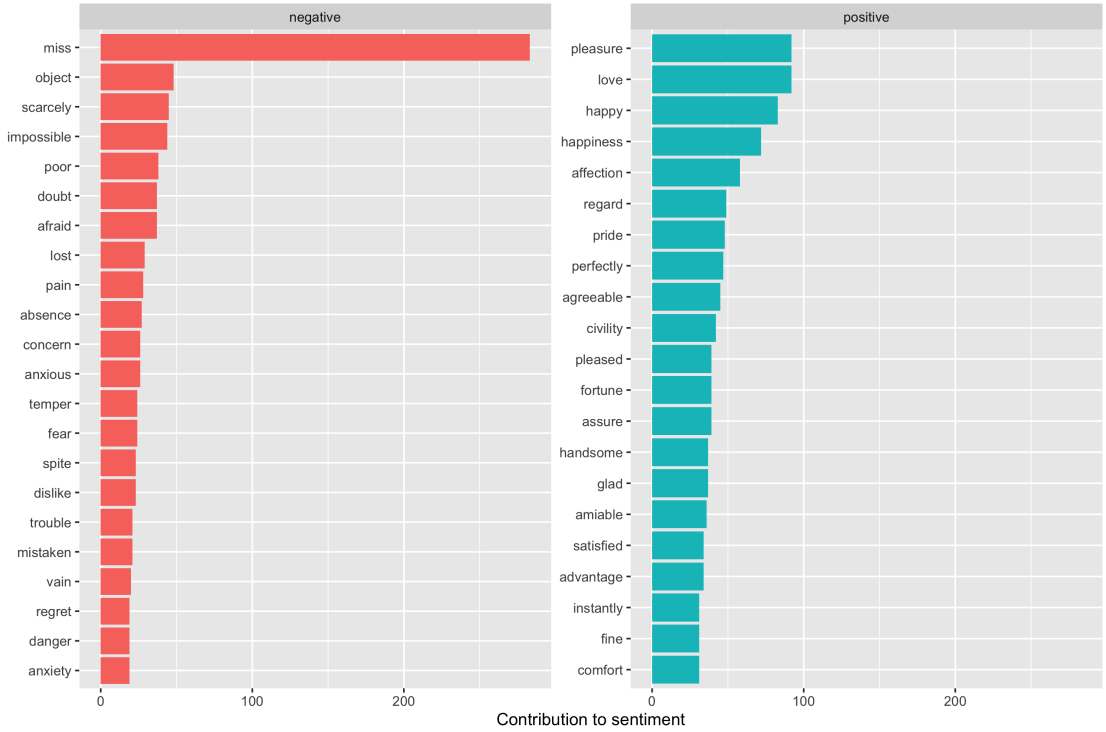
With this Figure 1 we can assess that Brontë is more prone to use connective words in her work. We can easily see that most of the words used on both books are similar in nature, except for the names of the characters. This can be associated with the fact that both authors are from the same era, and they both use British English.

Stop words however are not the main interest of the experimentation. The next step, in order to use the sentiment library is to create another variable without the stop words. This can be accomplished with the same library we used to tokenize the words on the downloaded document, only this time, we add the `anti join` variable. Inside we put `stop words` which is a data frame that contains pre loaded connective words in the English language.

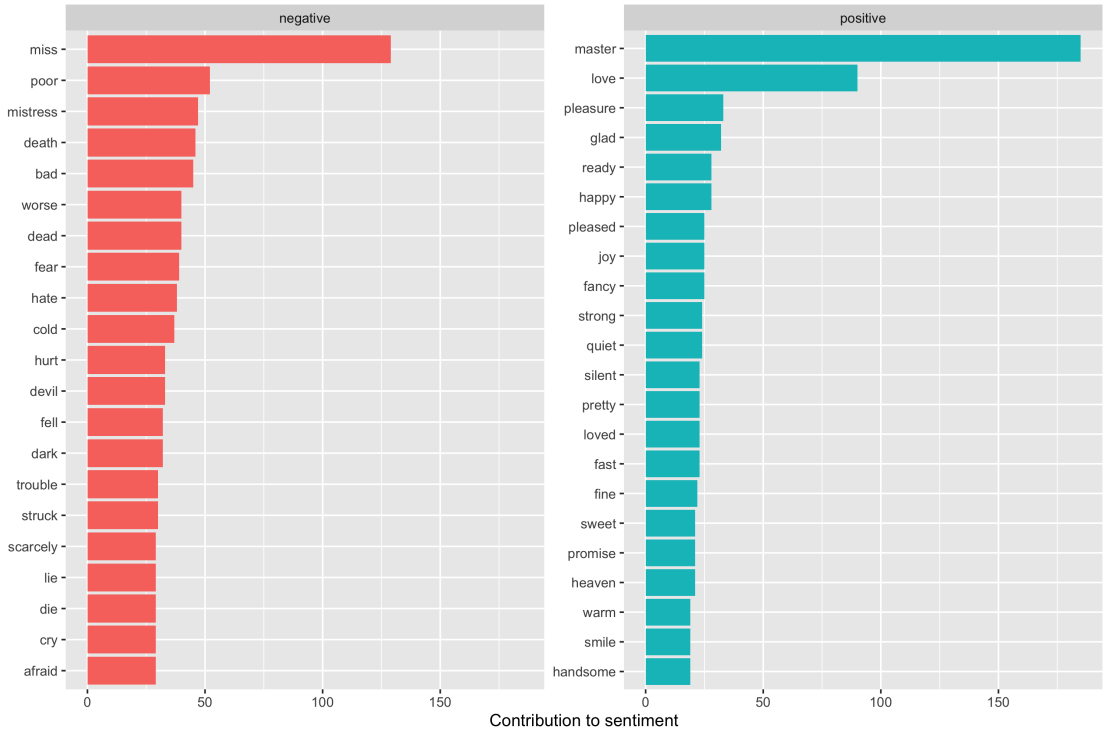
Having the data free of stop words, we can now use the library of sentiment to analyse both books. There are a few libraries of sentiment in the `tidytext` library, but in this case we choose the `bing` because it manages the words into positive or negative emotions. In Figure 2 the Subfigure 2a represents the positive and negative words of the book *Pride and Prejudice*, and the same is represented in Subfigure 2b for the *Wuthering Heights* book.

Right away when comparing the two Subfigures we notice that the negative words form the Austen novel have the feeling of being quite benign as opposed to the negative words in the Brontë book. For instance, in textitPride and Prejudice we have words like danger, poor, lost, absence, concern, dislike. And on the other hand in *Wuthering Heights* we have words like dead, hate, dark, hurt, devil.

In the side of the positive words, we see how the distribution of the frequency of the words is smaller in the Brontë side. Also, the top word *master* can be interpreted differently inside the book.

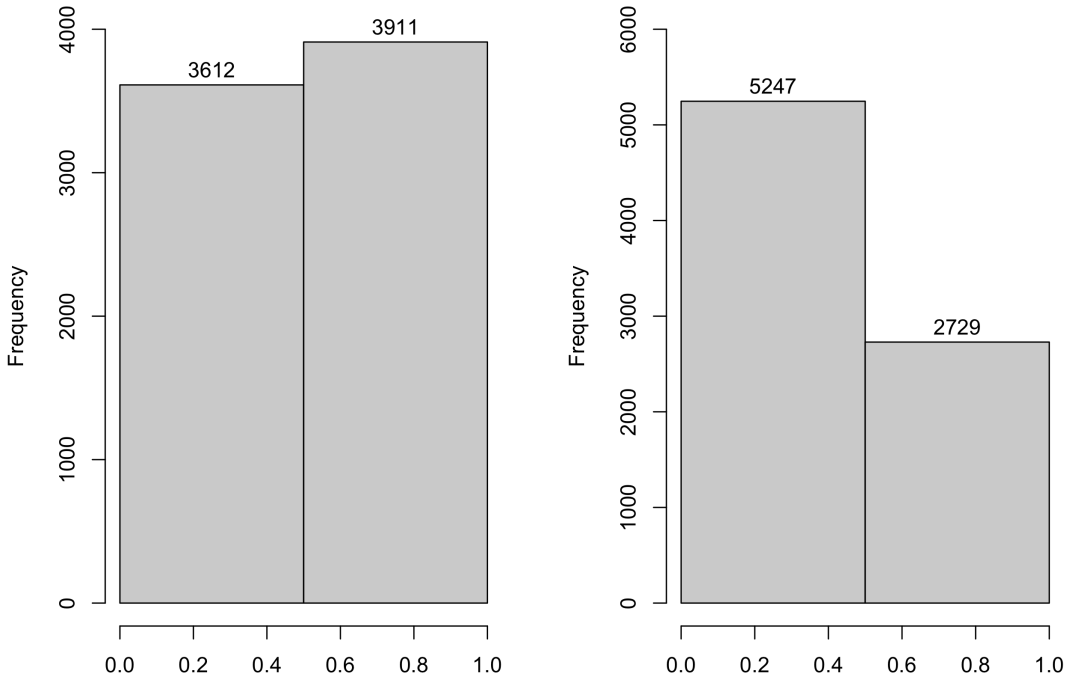


(a) Pride and Prejudice



(b) Wuthering Heights

Figure 2: Sentiment analysis of both books.



(a) Pride and Prejudice

(b) Wuthering Heights

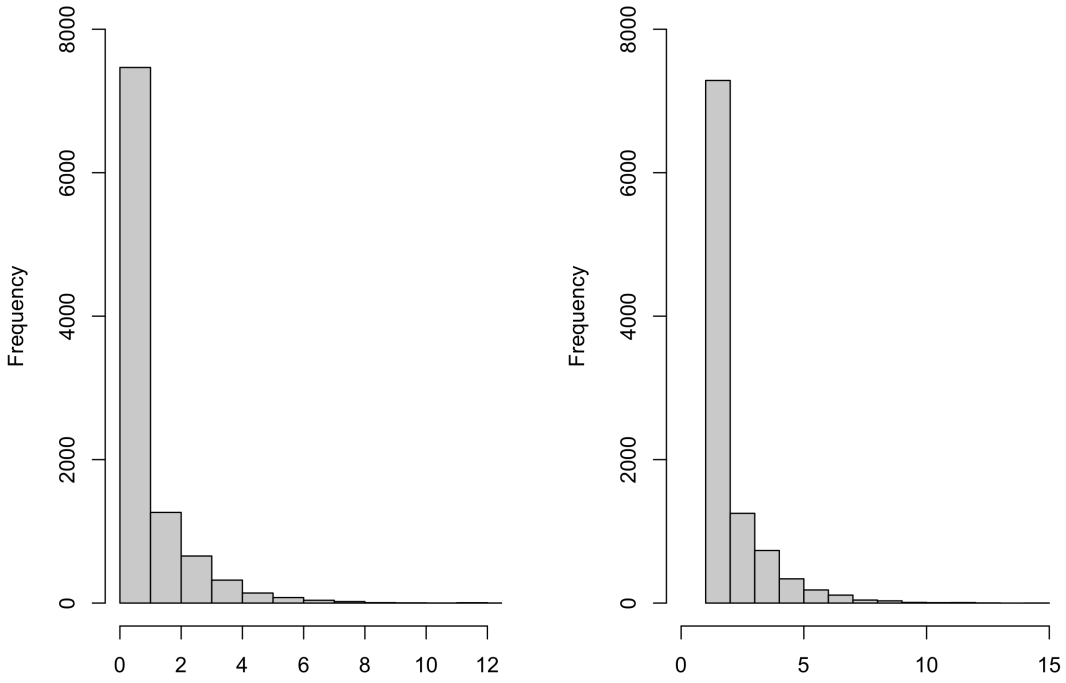
Figure 3: Frequency of positive and negative words. Negative are represented in 0 and positive in 1.

Having this outline of the sentiment analysis, we need to transform the data we have into something we can use for the histograms for the second part of this work. The reason we used the library `bing` for the sentiment analysis is because it divides the words into positive and negative emotions. With this, we transform them into binary data.

The first thing to know after this transformation is the distribution of positive versus negative words in both works. In Figure 3 we used histograms to represent that distribution. If we add the positive and negative words in each book, we find that the difference between total of words is not that great. *Pride and Prejudice* holds 7,523 words, while *Wuthering Heights* has 7,976.

The main and visible difference is that in Subfigure 3a the positive and negative words are almost even, with positive words being slightly higher. In the case of Subfigure 3b we see a stark difference, being that the negative words represent almost two thirds of the total of words.

In this case, Figure 3 goes appropriately with the previous knowledge of the books. While both authors are female, lived in the 19th century and both are novelists, the feeling of the book



(a) Histogram with rgeom tool.

(b) Histogram with made code.

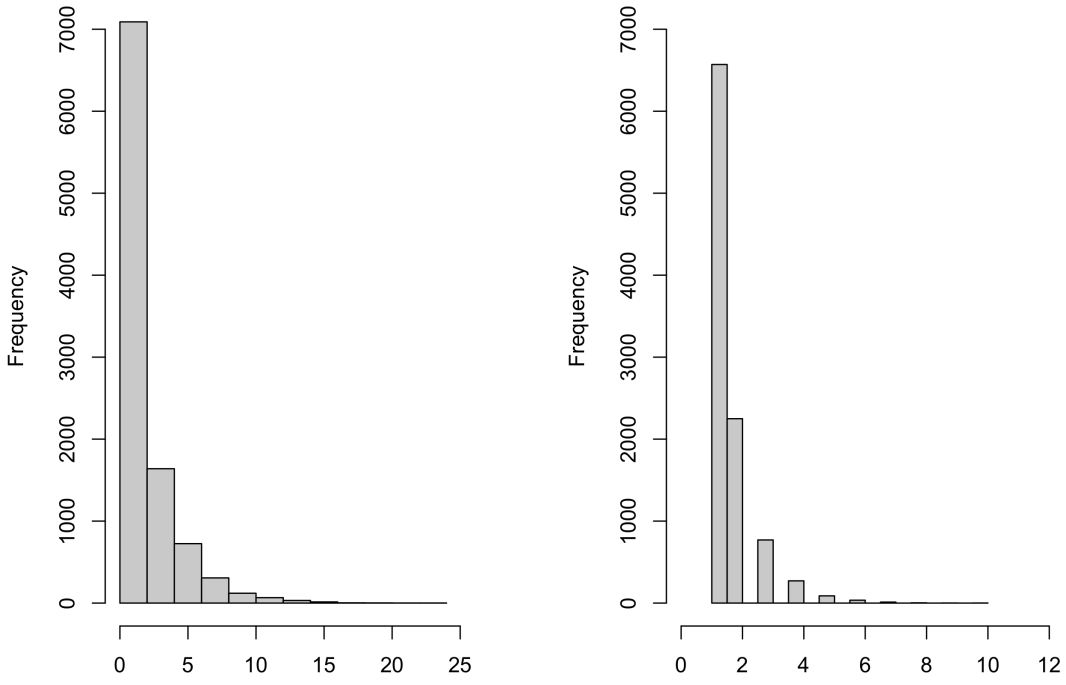
Figure 4: Comparison between the use of the tool `rgeom` and a made code on the book Pride and Prejudice.

is quite different. While in *Pride and Prejudice* our main protagonist Elizabeth is guilty of being biased and proud in her view of Darcy, and he in turn was spoiled and prideful as well, they are quite tamed compared with how volatile and prone to violence Heathcliff is. Also, the main protagonist Catherine is very selfish and cruel to everyone around her.

For this, the difference in characters behaviour, can be interpreted as one of the main reasons for the disparity in distribution of positive and negative words.

Having confirmed our first idea of the difference in tone in the books, we now use the information of the positive and negative words to perform distribution experiments. In Figure 4 we first perform an experiment with the tool `rgeom` [6] with the same variables we have in our data, to see if we can recreate something similar with our interpretation of the problem in a made code.

The `rgeom` tool takes a number of experiments and the percentage of success we have. In both cases we take as success if they have a positive word, so the percentage we give is the ones we calculated from our data. In the case of *Pride and Prejudice*, we have a percentage of 51%



(a) Histogram with rgeom tool

(b) Histogram with made code

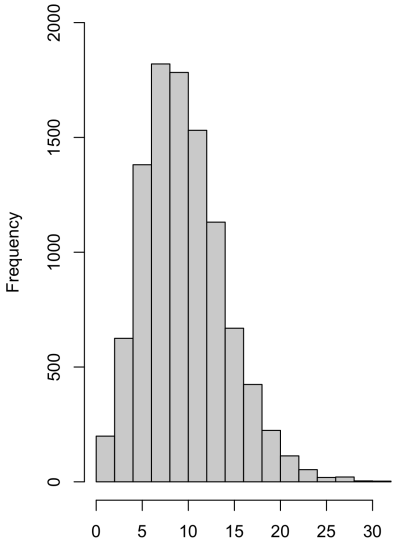
Figure 5: Comparison between the use of the tool `rgeom` and a made code on the book Wuthering Heights

positive words, and in *Wuthering Heights* we have a 34%.

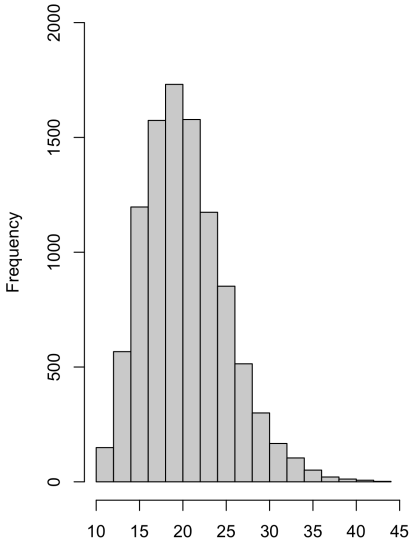
In Subfigure 4a we have the result of a 10,000 repetition of the experiment with a 0.51 probability of success. We replicated the experiments using our data of binary information, and with the tool `sample` we take a "word" and we see if it is a success (positive) or not (negative). As we can see, the behaviour of both plots is almost the same.

In Figure 5 we have the same experimentation, but with the book *Wuthering Heights*. In this case the difference between the subfigures are a bit more noticeable than in the previous experiment. In Subfigure 5a we have the result of a 10,000 repetition of the experiment with a 0.34 probability of success. Subfigure 5b uses the same made code than in Subfigure 4b changing only the parameters of probability and the data used.

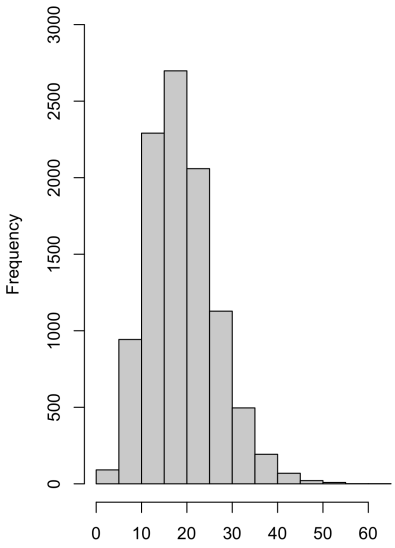
In Figure 6 we use now the `rbinom` [5] tool to further our experiments. This tool adds a variable inside the parameters. Same as with the `rgeom` tool, we have the number of experiments, then we have a number (k) which determines the amount of successful events before it can move



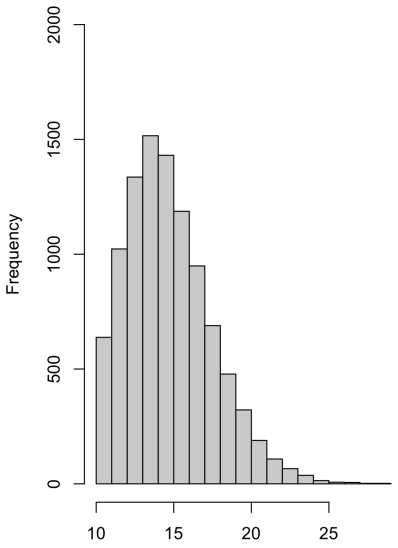
(a) Using rbinom, Pride and Prejudice.



(b) Using made code, Pride and Prejudice.



(c) Using rbinom, Wuthering Heights.



(d) Using made code, Wuthering Heights.

Figure 6: Comparison between the use of the tool `rbinom`

Table 1: Table of results from the `rhyper` tool with book Pride and Prejudice.

11	12	13	14	15	16	17	18	19
3	2	10	26	49	96	184	278	419
20	21	22	23	24	25	26	27	28
619	799	984	1036	1084	1101	892	818	596
29	30	31	32	33	34	35	36	38
429	276	143	84	43	24	2	2	1

Table 2: Table of results from the `rhyper` tool with book Wuthering Heights

20	21	22	23	24	25	26	27	28
2	2	8	27	41	86	145	258	417
29	30	31	32	33	34	35	36	37
566	786	977	1171	1126	1176	988	800	646
38	39	40	41	42	43	44	45	
375	213	109	50	17	8	5	1	

to the next iteration. Then it needs the parameter of probability of success.

In Subfigures 6a and 6c we use the parameter of 10,000 iterations, with k number equal to 10. The probability of success is 0.51 and 0.34 respectively. Then for the made code in both books we use the same parameters, but using our binary data of each book.

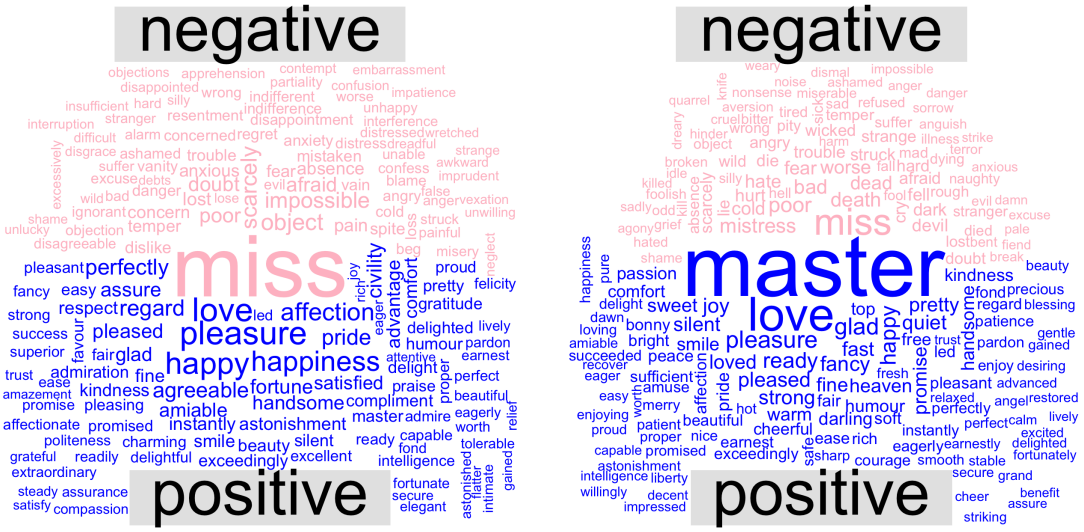
As we can appreciate in the comparison in Figure 6 the behaviour of the plots is very similar.

For the last experimentation with distribution, we take the tool `rhyper` [7]. In this case, it requires the amount of iterations, the amount of "white balls", the amount of "black balls" and the size of our sample. The white balls are represented as our positive words. The black balls as the negative ones. We set our iterations at 10,000 and our sample as 50.

Both tables have a gray and a white area. The gray area represents the number of bad words found in our sample of 50. The white area represents the number of times that many negative words repeated it self in our 10,000 iterations. In Table 1 we see the results of using the `rhyper` tool with the book Pride and Prejudice. Doing some calculations with our percentage of positive and negative words of this book, we have as a limit of negative words 24. So we add up all the frequencies after 24, and divide them for 10,000. The result of this is 0.4411.

Moving to Table 2 we have a different limit parameter. In this case is 33. So we add up all of the elements after the 33 and divide them by 10,000. The result is 0.4388.
.

Now, for our self made code, we use the same parameters for each book, 10,000 iterations, k of 10, and a sample of 50. The only variants are the positive and negative words for each book. Something else we added was the limit variable, which is the same as we used to make our



(a) Pride and Prejudice

(b) Wuthering Heights

Figure 7: Word cloud of positive and negative words in both books

calculation with the tables, 24 for Austen and 33 for Brontë. As a result we have, in the case of Austen a probability of 0.448 and 0.435 for Brontë. Comparing this to our results with the calculation of the tables we have 0.4411 and 0.4388, we have very similar results.

4 Other experiments

In the previous practice I struggled with word cloud plots. Here in Figure 7 I make a comparison of positive and negative words in both books. I find it quite funny that the main words for each book are miss and master.

5 Conclusions

With our first comparison of positive and negative distribution of words we were able to identify that Wuthering Heights has a more dark, more macabre tone of writing than Pride and Prejudice. The following plots with the R tools such as `rgeom, rbinom` and `rhyper` were interesting to develop in self made code. It was also curious to see the last experiment, and be able to prove again our impression on the difference of writing in both books.

In Tables 1 and 2 what sparked attention was the begining and end frequency of each. In Table 1 we begin with a small number and en with the number 38, with very low frequencies.

This tells us that there is a small amount of negative words in this data. And in Table 2 we begin with frequency of 20 and end up in the number 45. This is also aligned with the percentage of the positive and negative words we know of the book.

References

[1] Biografia de jane austen. hhttps://www.biografiasyvidas.com/biografia/a/austen.htm. Accessed: 2020-09-22.

[2] Biografia de emily bronte. https://www.biografiasyvidas.com/biografia/b/bronte_emily.htm. Accessed: 2020-09-22.

[3] Gutenberg project. https://www.gutenberg.org/about/background/history_and_philosophy.html. Accessed: 2020-09-22.

[4] Gutenberg project – pride and prejudice. https://www.gutenberg.org/ebooks/1342. Accessed: 2020-09-22.

[5] The binomial distribution. https://www.rdocumentation.org/packages/stats/versions/3.3/topics/Binomial. Accessed: 2020-09-22.

[6] Geometric distribution. https://rpubs.com/mpfoley73/458721. Accessed: 2020-09-22.

[7] The hypergeometric distribution. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Hypergeometric. Accessed: 2020-09-22.

[8] Gutenberg project – wuthering heights. http://www.gutenberg.org/ebooks/768. Accessed: 2020-09-22.