Text Mining Visualizations

of

The Little Mermaid

Joi Chu-Ketterer

DSC 680

Bellevue University

May 28, 2020

lchuketterer@my365.bellevue.edu

Instructor: Catharine Williams

Introduction

Literary analysis is a powerful research technique of exploring and analyzing a text's characters, motifs, symbolism, and overall intention. Often times, literary analyses are presented in essay form and are fueled by the reader's personal interpretations of the text's themes. While this approach has proven insightful through the ages, the advancement of technology can help make the process efficient. One of the main challenges in literary analysis is the time it takes for readers to fully analyze and understand a text. Additionally, most readers seldom retain all of the information they just read. The average person reads about twelve books a year [Jarmea 2018], but across fiction and non-fiction titles, they are only retaining 10% of what read [D'Souza *Psychotactics*]. Text mining is a tool readers can use to quickly understand the basic insights of a text as a process of turning unstructured data into structured data. While applications can apply to academic literary analysis, text mining has vast applications within business and marketing as well. Often times, text mining is the precursor to text analytics, which is the process of creating models and algorithms to achieve a goal [*Educba*]. This report focuses on creating visualizations for insights uncovered from text mining Hans Christian Andersen's *The Little Mermaid*, written in 1836. Throughout the report, examples of how the same methods can be used for business are included.

Business Understanding

Hans Christian Andersen's *The Little Mermaid* is one short story from his *Andersen's Fairy Tales* collection. The story follows a little mermaid as she falls in love with a human prince. The readers join her on an adventure to become human in attempt to win over his heart and earn an immortal soul. She fails to win his heart, and choses to sacrifice herself, which ultimately leads her to receiving an immortal soul. Text mining can be used to explore the short story in ways the reader may do themselves.

Text mining is a specific approach used to analyze unstructured data and convert it into structured data. Often times, data is extracted from PDFs, which stand for portable document formats. While a PDF may seem like an ordinary image file, it differs in its ability to preserve document formatting [*GCF*

Global]. For this reason, PDF version of documents are used for text mining to uncover insights from the unstructured data within.

Analysis

Data Extraction

The data for this project was extracted from a PDF version of *The Little Mermaid,* made public through Global Grey Publishing. The *pdftools* package was used to import the PDF into R, where the *pdf_text()* method was used to extract the text.

Method and Data Preparation

Once the data was imported, several text mining techniques were used to uncover patterns within the story. All analysis was conducted in R, with visualizations created in R, Tableau, and Canva. The majority of data preparation of the unstructured data was on data cleaning to remove punctuation or '*\n*' within the text. The following packages were used to clean the data:

- *tm*
- *stringr*
- *qdapRegex*
- *dplyr*
- *tidytext*
- *tidyverse*

Results

Several cleaning and parsing techniques were used on the text for different analysis. Three main approaches were taken, breaking down into *Counts, Sentimental Analysis,* and *Gender Analysis.* Within *Counts,* quote ratio, word frequency, sentence length, and punctuation count were analyzed. Within *Sentiment Analysis,* overall sentiment and sentiment contributions were analyzed. And finally, within *Gender Analysis,* trailing words in bigrams were analyzed. This report includes the most relevant visual aids created; additional graphs that were created are included in the Tableau and R Markdown files.

Results - Counts

Quote Ratio

Understanding the ratio between spoken words to total words can be useful in understanding how a story is structured. For instance, within the entertainment industry, understanding the total count of spoken words of a book may help a screenwriter gauge how adaptable a novel is for a screenplay.

The ratio of quotes to text in *The Little Mermaid* was calculated by counting the number of words between quotation marks and the number of words in the text overall. For this reason, no initial cleaning was done to the text, as to preserve the punctuation. Once the text was split by the quotation mark, the strings of characters were saved into a dataframe, and then cleaned for additional analysis. There was a total of 1,849 spoken words within the text and 9,251 total words altogether, resulting in about a 20% spoken word ratio. With such a small portion of spoken word, it can be interpreted that the story focused more on presenting the individual character's emotions or focused on describing sceneries. In either case, the story was heavily descriptive. In the case of a screenwriter trying to adapt the story into a movie, they may welcome the detailed imagery and scenery descriptors and freedom to write original quotes.

Word Frequency

Knowing the most common words within a text can uncover a lot about its themes and motifs. Within business analytics, word frequency clouds are often used to show the focus of an article. Quickly being able to identify what an article is about can help with document categorization or market research. For instance, being able to differentiate between an article that predominantly uses the words *pigment, gouache, opacity* is not relevant for a tech company's market research into existing IoT's.

To prepare the data for word frequency, all punctuation was removed from the text, and each word was tokenized. With a collection of all the words used within *The Little Mermaid*, stop words were removed. Stop words are identified as commonly used words that hold no value but are essential for proper grammar [Hope 2019]. Examples of stop words include *the, a, an,* and *in.* Once these were removed the 'true' most used words in a text can be uncovered to help identify it's focus.

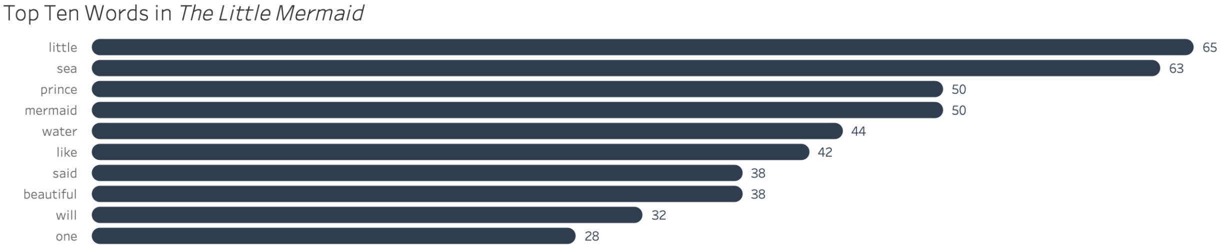

Figure 1. Top ten words used in *The Little Mermaid*

Without reading a single word of the story, looking at *Figure 1* lets the reader know *The Little Mermaid* is a story about a mermaid and prince, and a majority of the story happens underwater and at sea. Similar to how word frequency can help businesses determine which articles are relevant for their research, the average reader could use word counts to determine if a story would be interesting to them. Someone who primarily loves non-fiction stories can quickly pass on *The Little Mermaid* and spend their time reading texts that better align with their interests.

Sentence Length

Varying sentence lengths dictate the cadence of a story, and the pace in which the reader consumes it. The longer the sentences, the more fatigued the characters and readers feel reading them. Conversely, short and curt sentences can leave the reader feeling anxious. Visualizing the variance in sentence lengths provides readers a general sense of the pacing of the story. In a business or academic setting, visualization sentence lengths can help a write proofread their document before sending it into submission. Especially at a younger age, run on sentences are hard to detect on their own, but would be easy to spot in a graph.

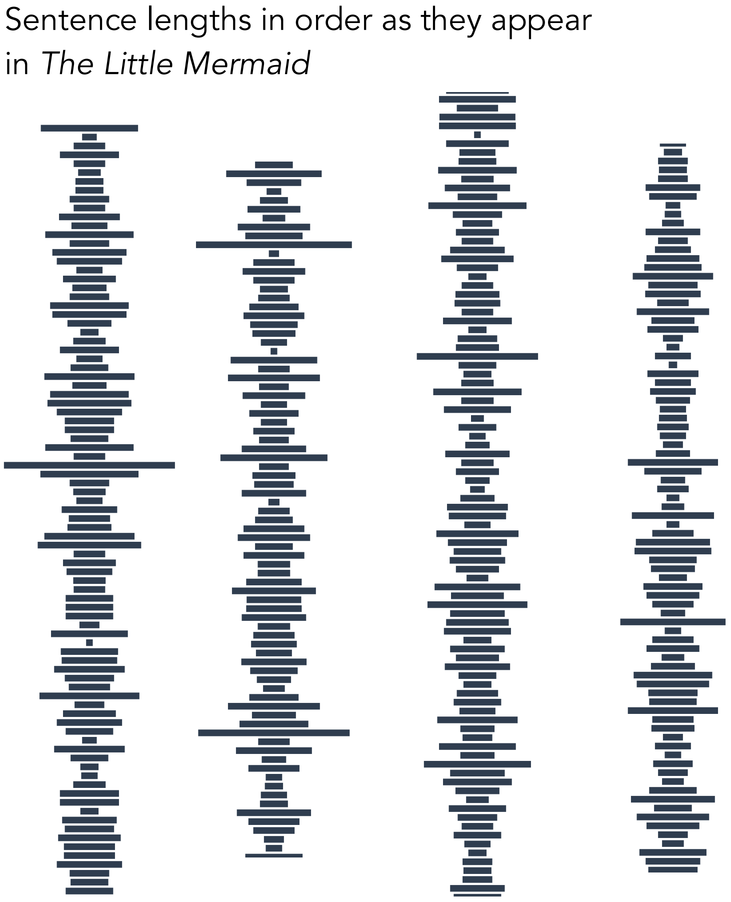

Figure 2. Sentence lengths in *The Little Mermaid* in order as they appear in the story

As a short story, *The Little Mermaid* totals 21 pages. Even so, there are 337 sentences. As *Figure 2* demonstrates, it is difficult to show sentence counts as one long image, depending the presentation format. This would be a difficult visual aid to use for longer texts or novels. Alternatively, the average length of sentences could be a useful metric in determining story cadence. Still, the ability to see the different sentence lengths is what makes this metric interesting and useful.

Punctuation Count

Similar to word count, punctuation count can help readers identify the tone and purpose of a document. For instance, a document that has more question marks than periods would suggest the article acts more as a prompt or thought-provoking guide. A document that has equal parts question marks and periods may be a joke book. A document that predominantly leaves it open for interpretation, as *Figure 3* shows, although the reader would form a stronger case for not categorizing *The Little Mermaid* as a joke book.

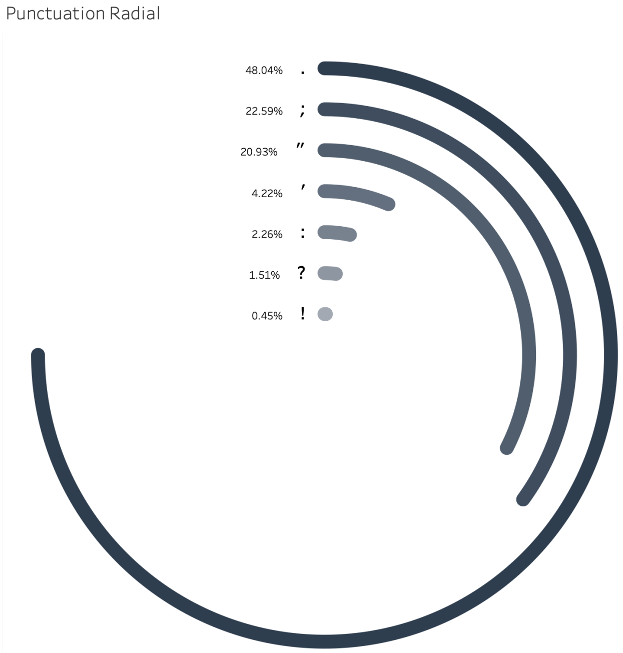

Figure 3. Punctuation count in *The Little Mermaid*

Results – Sentiment Analysis

Sentiment analysis has extensive business applications, ranging from product research, consumer insights, and language filtering. Understanding the sentiments of a company’s users and target audience provides them with the knowledge on how to improve or maintain the relationship. Another application could be in language filters for social platforms to prevent harmful sentiments being shared.

Overall Sentiment

The *bing* lexicon [see Appendix A] was used to assign sentiment to the words within the text. *Figure 4* shows that *The Little Mermaid* is generally a positive story. However, with such close counts, it could indicate to the reader that there are some darker themes within the story. Another personal setting sentiment analysis could help may be for parents trying to find appropriate material for their child to read. Without having the time to read the book itself, or reviews of the book, running a simple sentiment analysis could help them determine if the book is sentimentally appropriate to read.

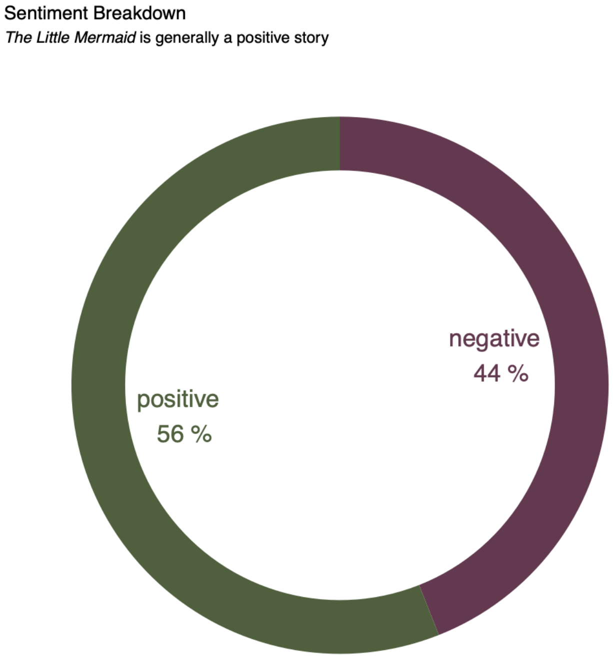

Figure 4. Sentiment ratio of words in *The Little Mermaid*

Sentiment Contribution

Overall sentiment can be broken down into more specifics, for example, sentiment contribution. Sentiment contribution focuses on understanding which words are contributing the most towards the story. Acting as a hybrid of sorts between sentiment analysis and word frequency, sentiment contribution was calculated by assigning negative words with a value of *-1* and positive words with the value of *+1.* For this reason, words used more often have a higher contribution count. As one would predict, the words present in *Figure 1* would be seen in *Figure 5;* however, the *bing* lexicon is not complete. For this reason, not all words are included, and so not all the words in *Figure 1* had an assigned sentiment, and thus was considered to contribute zero value.

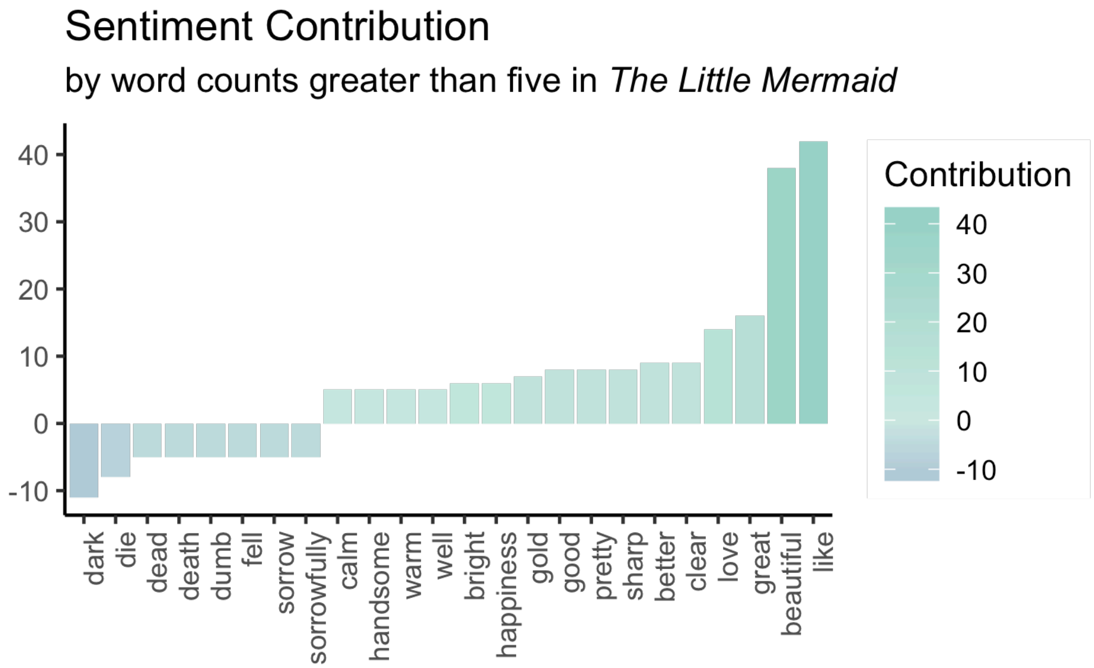

Figure 5. Sentiment contribution for words used more than five times in *The Little Mermaid*

Despite missing a few words from the lexicon, *Figure 5* further supports the insights uncovered in *Figure 4* that *The Little Mermaid* is predominantly a positive story. Bringing attention to the words themselves, one can see the author uses light and dark imagery a lot within the texts. In a literary application, understand which words are used to pull the story in a negative or positive way can help an author maintain the same themes throughout a series. Should the author hit writer's block, they may be able to use sentiment contributions to understand common words and motifs they used in past novels.

Results – Gender Analysis

Trailing Words in Bigrams

With a constant struggle in gender equality, it is useful to understand how men and women are portrayed differently within literature or articles. Within the entertainment industry, this is especially important as society starts to demand equal attention to female characters as their male counterparts are given in important and screen time. Gender analysis was conducted by extracting the bigrams [see Appendix B] within the text, and then filtering by those that started with *she* or *he*. The frequency of the trailing words was then used to calculate the ratio between assigned gender. For instance, the word *could* was found behind both *he* and *she,* but twice as much behind *she*, as *Figure 6* demonstrates.

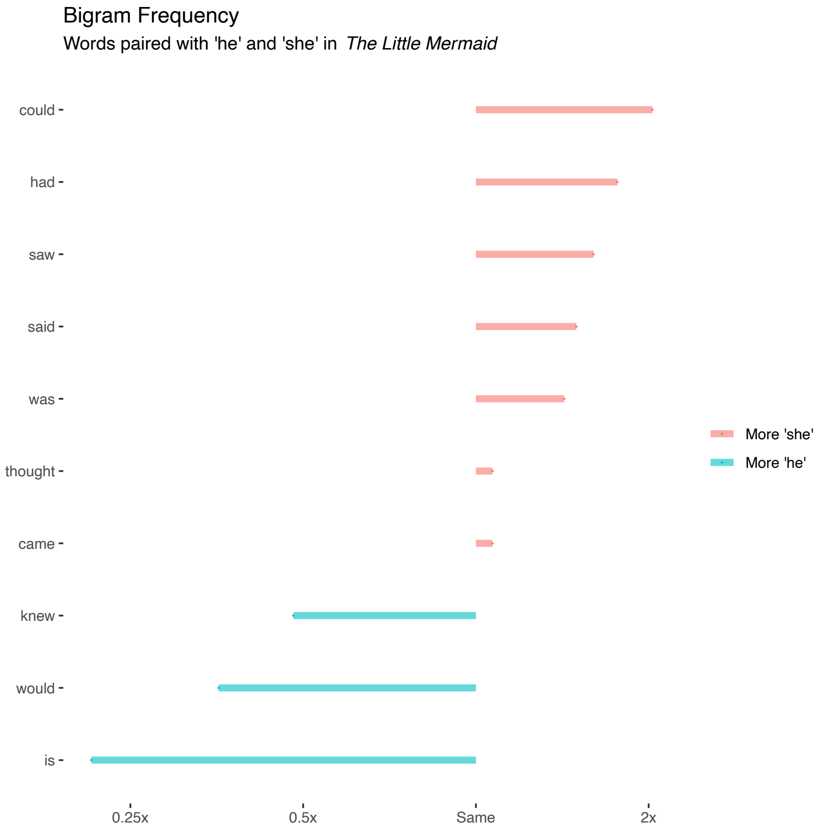

Figure 6. Bigram frequency for trailing words behind *he* and *she* for words used more than three times

Knowing the words associated with each gender can being insight to gender imbalances within a text. *Figure 6* shows that a female character is referred to far more than a male character, which suggests the main character is female. Additionally, the trailing words that lead both genders are fairly similar in their sentiment and purpose. *Could, had,* and *saw* are not strikingly different from *knew, would,* and *is.* In a more dynamic story, trailing gender bigram analysis could be used to identify villains and victims.

Benefits

The visualizations provided within this report help demonstrate how otherwise ignored metrics of a story can be brought to life through graphs. While reading a story, the reader is unlikely to keep a running tally of all the words used. However, by visualizing them all at once helps reinforce the themes and motifs of a story. Additionally, uncovering how males and females are portrayed within literary analysis can showcase how equal or different each character's status is.Text mining is the steppingstone to using machine learning and automation to help categorize and analyze unstructured literary data even more.

Future Work

As this project starts to uncover patterns with *The Little Mermaid,* taking into consideration additional texts could provide deeper insight. *The Little Mermaid* is only one of many of Andersen’s fairy tales. By creating a corpus of Andersen’s fairy tales, similar analysis could be conducted to understand how each story differs. Understanding the chronological changes in sentiment from story to story could potentially uncover patterns in his mental well-being. Additionally, running a bigram gender analysis on most popular works per century could provide insight into the changes between the male and female dynamic, if there is one. Within sentiment analysis, additional lexicons could be used to see how the sentiment analysis results differ.

Conclusion

Uncovering patterns within the texts by running simple calculations and sums is a huge aid in understanding the tone and intention of a piece. Within literary analysis, uncovering such insights can be used as supplemental guides for a reader or to understand patterns between works. In the business world, text analysis can be used to understand user experience and categorize documents appropriately. Text mining techniques has uncovered that *The Little Mermaid* is a text heavy, structurally well balanced, positive, gender balanced piece of work. With so many directions text analysis can go, the approach is just as endless as the documents are unstructured.

References

"Lexicon", *Merriam-Webster.* Retrieved from https://www.merriam-webster.com/dictionary/lexicon

"Text Mining vs Text Analytics", *Educaba.* Retrieved from https://www.educba.com/text-mining-vs-text-analytics/

"What is a PDF file?", *GCF Global.* Retrieved from https://edu.gcfglobal.org/en/basic-computer-skills/what-is-a-pdf-file/1/

Andersen, H. C. (1870). *Andersen's Fairy Tales.* Global Grey Publishing. Retrieved from https://www.globalgreyebooks.com/andersens-fairy-tales-ebook.html

Chang, Terry. 14 Mar, 2018. "Three Text Sentiment Lexicons in R's tidytext", *Data Critics.* Retrieved from https://datacritics.com/2018/03/14/three-text-sentiment-lexicons-in-r-tidytext/

D'Souza, Sean. n.d. "How To Retain 90% Of Everything You Learn", *Pyschotactics*. Retrieved from https://www.psychotactics.com/art-retain-learning/

Hope, Computer. 30 Jun, 2019. "Stop Words", *Computer Hope.* Retrieved from https://www.computerhope.com/jargon/s/stopword.htm

Jarema, Kerri. 19 Apr, 2018. "How Many Books Did The Average American Read In The Last Year? This New Study May Surprise You", *Bustle.* Retrieved from https://www.bustle.com/p/how-many-books-did-the-average-american-read-in-the-last-year-this-new-study-may-surprise-you-8837851

Kapadia, Shashank. 26 Mar, 2019. "Language Models: N-Gram", *Towards Data Science.* Retrieved from https://towardsdatascience.com/introduction-to-language-models-n-gram-e323081503d9

Appendix A

Lexicons

Lexicons in the general sense is simply the "vocabulary of a language" [Merriam-Webster], or a dictionary of terms. Within text analysis, in R, the *tidytext* package has three lexicons that can be used for sentiment analysis. There is the NRC, AFINN, and BING, each with their own advantages and disadvantages [Chang, *Data Critics*].

NRC

The NRC lexicon is a compilation of binary categorization of *positive* and *negative* in addition to emotions such as *joy* and *trust* to name a few. With the addition of emotion sentiment, this means each word can have multiple associated sentiment. For instance, the word *beautiful* could be categorized as *positive* and as *joy.* While the addition of emotion sentiment can be more insightful, it can lead to longer load and run time.

AFINN

The AFINN lexicon assigns words a score between -5 and 5. For instance, the word *trouble* is assigned -2 while the word *beautiful* is assigned 3. While the numeric value in sentiment for each word is incredibly useful for understanding sentiment contribution, the AFINN lexicon also has a longer load time.

BING

The BING lexicon is a compilation of binary categorization of *positive* and *negative*. As the simplest of the three lexicons, BING is useful for basic text analysis to get a general idea of a text. Additionally, it has the quickest load time of the three.

Appendix B

N-grams

N-gram by itself is a term used to refer to a sequence of *n*-words. For instance, a bigram is a sequence of two words, while a trigram is a sequence of three words [Kapadia, *Towards Data Science*]. For example, in the phrase “the little mermaid swam up”, the bigrams would be as follows:

- the little
- little mermaid
- mermaid swam
- swam up

Often, n-grams are used for statistical language models. One of the more common applications of such modeling is with Apple’s iPhone predictive texting.