Selecting Subsets Of Amazon Reviews

Y. E Goeke
gokcey@uncw.edu

Douglas Kline
klined@uncw.edu

Jeffrey Cummings
cummingsj@uncw.edu

Ronald Vetter
vetterr@uncw.edu

University of North Carolina Wilmington
Wilmington NC

Abstract

Advancing technology enabled people to more easily create, store, and exchange information. While the amount of information definitely increased, the same cannot be said about the quality. As information kept flooding, finding and selecting quality information in a massive pile became more difficult than ever. As a result, the whole process of finding and reading the right information costs more and more every day. Like every problem, this one also created the need for a solution, and searching for ways of reading more efficiently became a hot topic. We have also become interested in contributing to the solution and conducted this study. Selecting a subset of amazon reviews aims to solve the problem stated above, specifically for a large number of relatively shorter documents (reviews).

Keywords: Subset selection, Amazon reviews, document subsets, subset optimization, text analytics, text summarization.

1. INTRODUCTION

The time it takes to read multiple documents on the same subject is sometimes not feasible — especially when the documents are vast with high redundancy or bias in their content. Summarization and subset selection methods can help to ease this burden.

Text summarization has two main approaches: abstractive and extractive. The abstractive approach attempts to understand the content of the corpus (all of the reviews), then build a summary. In contrast, the extractive approach constructs the summary from representative snippets of the corpus. Both of these approaches have their advantages, but they can also produce disjointed summaries.

This paper explored the subset selection methods defined by Kline, D.M. (1998) and applied them to Amazon reviews. Using the subset selection methods did not require understanding the corpus's content or construct cohesive sentences from scratch. Instead, it kept documents in their entirety; by doing so, it preserved the author's original intent.

A good subset of reviews would help consumers who want to know all the concepts represented in the corpus without reading it all. In his paper, Kline, D.M. (1998) presented ways to optimize and select the best subset based on the length

and distance. While the length of the subset meant the number of words in it, distance implied content similarity between the corpus and the subset.

Essentially, this study aims to help answer a question such as "Of the n number of reviews, which k should I read?" Amazon already attempts to help consumers with this by allowing other customers to rate reviews and present reviews in various ways: by recency, positive, negative, etc. Although, there is no guarantee that the presented reviews will cover all aspects of the product. In contrast, the subset selection methods we use will account for that.

2. DATA

Prior to choosing a product, we suspected that certain properties of the product or the reviews could lead to different behaviors and produce different results. Count of reviews and type of product were two main properties we mainly paid attention to. While other large Amazon review datasets already exist, in order to observe differences in behavior and result, we picked three products that met the following criteria:

- Chosen products should belong to different categories (clothing, technology, literature, etc.) so that they could have a different vocabulary.
- Each product should have at least one thousand and at most four thousand reviews. Hypothetically, there was no need to put an upper limit to the count of reviews. However, our resource limitations forced us to do so.
- The count of reviews for each product should vary.

Following the criteria above, the data was collected using a free Chrome plug-in called WebScraper. This product facilitates the creation of a web scraping task and scrapes the web pages, producing output in a variety of formats without the need to write custom code. The Webscraper defines the scrape job in a JSON format and allows for delays between page requests to prevent bot detection.

The reviews were collected in February of 2020. Products chosen consisted of

- A woman's dress: Sylvestidoso Women's A-Line Pleated Sleeveless Little Cocktail Party dress
- An SSD hard drive: Crucial MX300 525GB 3D NAND SATA 2.5 Inch Internal SSD
- A fiction novel: Less (Winner of the Pulitzer Prize): A Novel Hardcover

In addition to the raw text of the reviews, we also collected the title, the date submitted, the username of the reviewer, and other items related to the review. The actual scraping job was performed on a standard laptop and took less than an hour per product.

The scraping job produced one comma-separated value (CSV) format file per product. After the collection of datasets, Microsoft PowerBI was used to pre-process the datasets. Pre-processing included cleaning special characters, converting data types, and maintaining the CSV format. Table 1 shows the resulting data set.

Product	Price	# reviews	Avg word count	Max word count	Min word count	Dictionary Length
Dress		3797	26.9(28.2)	196	0	4085
SSD		1040	65.6(74.8)	995	1	4762
Book		1801	54.3(91.2)	1359	1	7392

Table 1. Dataset Statistics

3. IMPLEMENTATION

The experiment was programmed in the Python language with notable usage of the following libraries:

- Pandas, Numpy
- Sklearn for text analytics
- Matplotlib

Evaluating the subsets of reviews from entire reviews is a combinatoric problem, represented by n-choose-k. A brute force approach to such a problem could be time costly. Take the women's dress product for an example: evaluating all subsets of three from 3797 reviews would create more than 9 billion combinations. Unfortunately, our computational resources were not sufficient enough to perform a brute force approach in a timely manner. To address this limitation, we had to reduce the dimensions of our data. We performed the following data reduction steps and made the following judgment calls in that order:

1. First, we eliminated the standard sklearn stopwords from the dataset. Our main

concern with choosing the stopword lists was their extend. It is a known issue for stopword lists to have certain words that could have a meaning in the context. In that sense, skylearn's list was not too inclusive and deemed reasonable.

2. Second, we eliminated terms with low standard deviation. The low standard deviation would indicate highly frequent or infrequent terms. Those terms would not be helpful in differentiating reviews from each other.

3. Finally, we eliminated low word count reviews because we could not justify their chances of being selected for the optimal subset, especially when n is larger.

Product	# reviews	Dictionary Length
Dress	314	32
SSD	448	70
Book	286	39

Table 2. New dimensions after data reduction

This reduced the number of combinations for the dress product to approximately 5 million.

	Constraints
	Number of Words
Minimize Distance	$\boldsymbol{min\, dist}\left(\frac{\sum_{j=1}^{N} x_i a_j}{\sum_{j=1}^{N} x_i}, H\right)$ **Subject to:** $\sum_{j=1}^{N} x_i < p_w$

p_n = threshold for the number of documents
p_w = threshold for the number of words (terms)

After data reduction, reviews were represented as term frequency (TF) vectors produced with standard sklearn functions. The corpus vector was created by summing all of the TF vectors. After generating it, we also normalized each vector, including the corpus vector. At this point, reviews were ready to be used. We calculated the Euclidean distance between the corpus and each subset of reviews to represent the distance (similarity) between the two. We recorded data for all possible subsets of size 1, 2, and 3 reviews. Furthermore, the study was performed using 1-grams, 2-grams, and 3-grams. For each subset and n-gram combination, the word count and distance were saved. After that, we were able to sort by distance and/or word count to find the best subset. The entire study written in python was run on a standard business laptop. Evaluating all subsets of size 3 of the SSD reviews took less than two hours.

We evaluated two optimization models. The first minimized the distance subject to a maximum word count constraint.

The second model we evaluated minimized the number of words subject to a maximum distance constraint.

		Constraint
		Maximum Distance
Minimize	**Number of terms (words)**	$\boldsymbol{min} \sum_{j=1}^{N} x_j w_j$ **Subject to:** $\boldsymbol{dist}\left(\frac{\sum_{j=1}^{N} x_i a_j}{\sum_{j=1}^{N} x_i}, H\right) \leq p_d$

N = total number of terms
a_j = vector of the relative term frequency
x_j = binary-valued or zero-one variables
p_d = threshold chosen for distance
H = normalized vector

4. RESULTS

Variables for the optimization would normally be chosen by the user. While some users can say that they do not wish to read more than 400 words, others could say something else. To show some examples, we arbitrarily picked those numbers.

The results for the two models above were not much different to us. The real evaluation was left to the reader to be made. To do your own evaluation, you can refer to the results below.

In addition to findings for each model, we also found that 1-gram, 2-gram, and 3-gram models were nearly identical. Thus, the choice of n-gram did not appear to be meaningful in Amazon reviews.

Below are the results for the first model (distance minimization).

Dress
The optimal solution for the dress corpus was a subset of 3 reviews with a total of 169 words. The calculation was based on 1-gram terms. The word count constraint was 200 words.

Review 1: *"I absolutely love this dress. For the price, it is a steal. I am 5'5, 120 pounds and usually wear a dress size 2. The small fit perfectly. The fabric is also thicker than you would expect. I will be purchasing this in another color!"*

Review 2: *"Super cute dress. Fits as expected. I ordered the small and I'm 5"0 and 110 lbs with broad shoulders. The quality of the material is great and not see through like most white dresses I've seen/tried on. I wore nude panties and a nude strapless bra with it just in case though."*

Review 3: *"Bought this dress when my boyfriend said he was taking me out and to "dress up really nice". I didn't want to spend a ton of money and found this simple black dress! It was perfect and fit great!! Come to find out my night out was a surprise 50th bday party for me! I got tons of compliments All night on this dress! It was the best purchase!!! I'm 5'3 and 135 lbs, the medium fit perfect and was just above my knee. Hid my little belly too!"*

SSD
The optimal solution for the SSD corpus was a subset of 3 reviews with a total of 171 words. The calculation was based on 2-gram terms. The word count constraint was 200 words.

Review 1: *"The drive is good and that is the reason for the 5 stars. Windows 10 boot time between 15 and 20 seconds. Programs like Photoshop load in a few seconds.*
The Acronis software would not clone my original 1TB hard drive to the new SSD. Spent too much time trying to make it work and ended up using a stand alone duplicator."

Review 3562: *"worked well for upgrading a late 2011 Macbook pro to an ssd and El Capitan. The screw-in posts from the old hdd need to be used here to better secure this in the macbook rather than trying to use the including plastic spacer thing, which will still allow movement. Computer is booting the OS and opening apps much faster."*

Review 3965: *"Just installed Crucial MX300 525GB SATA 2.5 Inch internal SSD drive, in my ACER Apire 7741Z-4433 laptop. The original hard drive had disk error problems. It gave new life to the laptop. It is fast and I have not had any problems at all. I recommend this drive. Problems FIXED."*

Book
The optimal solution for the Book corpus was a subset of 3 reviews with a total of 241 words. The calculation was based on 2-gram terms. The word count constraint was 400 words.

Review 1: *"2018 Pulitzer Prize for Fiction. This novel about a gay man (a struggling novelist) who decides to travel around the world when his lover of 10 years marries someone else is very funny overall and even touching in places. The writing is generally good and lyrical, although a bit over-done here and there (i.e., a few too many metaphors and similes). The novel is extremely funny, although the humor sometimes seems like an inside joke that only "the right people" will get. The characters are well developed although perhaps a bit stereotyped and predictable. Still, the protagonist (Arthur Less) is likeable and very human. A fun read."*

Review 2: *"Less" is so well written with word phrasing that would make me read and re-read the passage just for the enjoyment. By the end of the book, I felt like I knew Arthur Less and all of his complexities. The book turns and regroups and then turns again. I really enjoyed this book on many levels. Congratulations to the author...and thank you."*

Review 3: *"Unlike some reviewers, I couldn't wait to finish this book. I did not find it humorous, endearing, or an interesting travelogue. It is the story of a sap who has the misfortune of having a poor love life. I couldn't muster up any feelings toward the character, or any other in the book. I have no idea why it won an award, except if it's because it's a gay love story. But I don't think I would've liked it if it were heterosexual. Just a dumb book IMO."*

We experimented with visualizing the subset in relation to the corpus. We used two main methods. First, we showed the word cloud representing the subset next to the word cloud representing the corpus for visual comparison.
Second, we showed the histogram of normalized term frequencies. The graph charted the subset's histogram and the corpus's histogram as lines on the same graph for comparison.

See Appendix A, B, and C for visualization of each corpus.

6. CONCLUSIONS

Based on the results, we deemed two methods/models from Kline, D.M. (1998) to be useful in creating a subset of Amazon reviews from a large review set. We claim that this approach is less complicated, more straightforward, and promising. Visualization of the selected solutions supported this claim.

There are many more questions to be asked and paths to be followed. We hope this study will help other studies, open up new questions, and encourage future studies.

7. ACKNOWLEDGEMENTS

We thank Congdon School of Supply Chain, Business Analytics, and Information Systems for the support they provided throughout this study.

8. REFERENCES

Kline, D.M. (1998) Optimization Models for Creating Reduced Reading Lists. Working paper # 98-02MIS, Center for Business and Economic Development, Sam Houston State University, https://www.shsu.edu/centers/cbed/documents/working-papers/No.98-02MIS.pdf.

APPENDIX A: DRESS CORPUS VISUALS

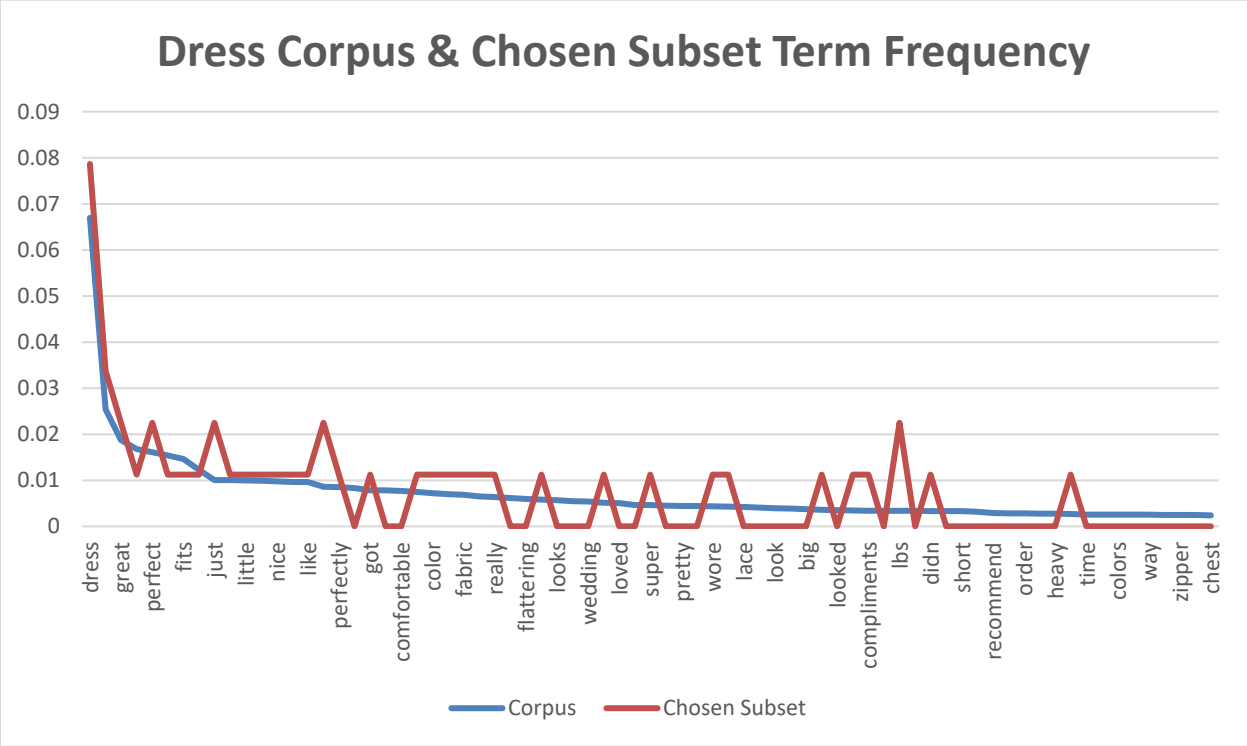

Figure 1. Dress Corpus & Chosen Subset Term Frequency

Figure 2. Dress Corpus Word Cloud

Figure 3. Dress Chosen Set of Reviews Word Cloud

APPENDIX B: SSD CORPUS VISUALS

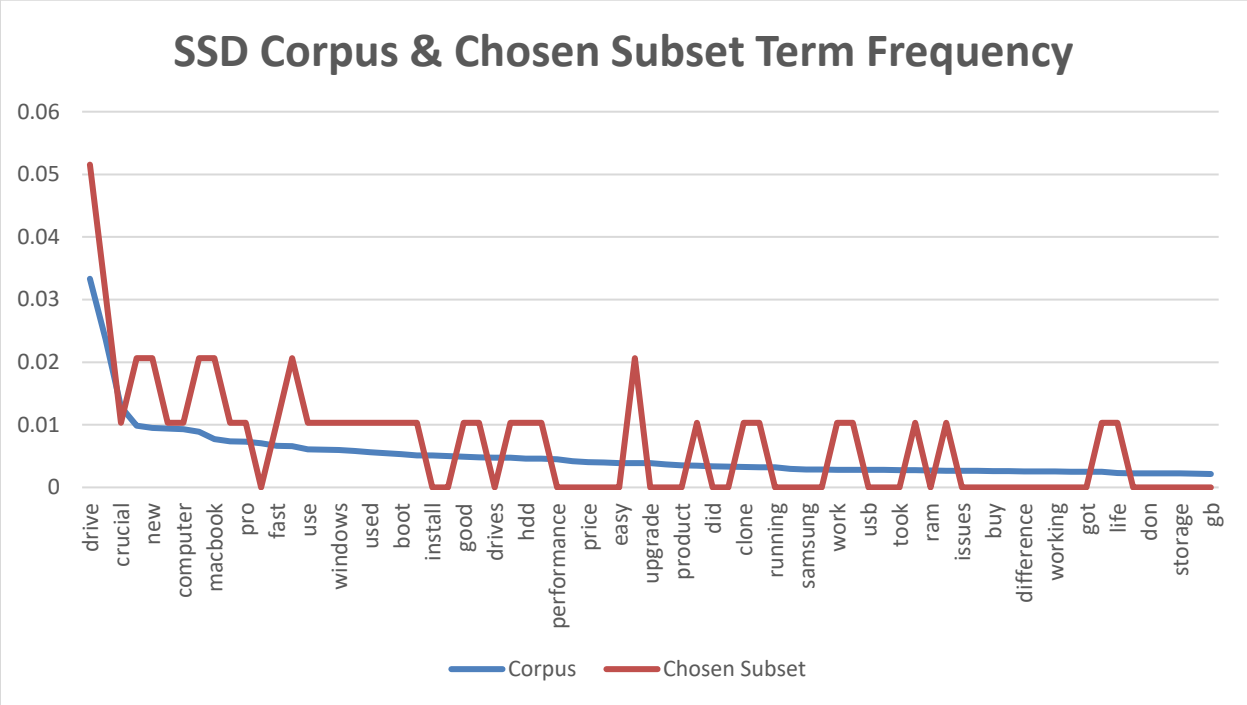

Figure 4. SSD Corpus & Chosen Subset Term Frequency

Figure 5. SSD Corpus Word Cloud

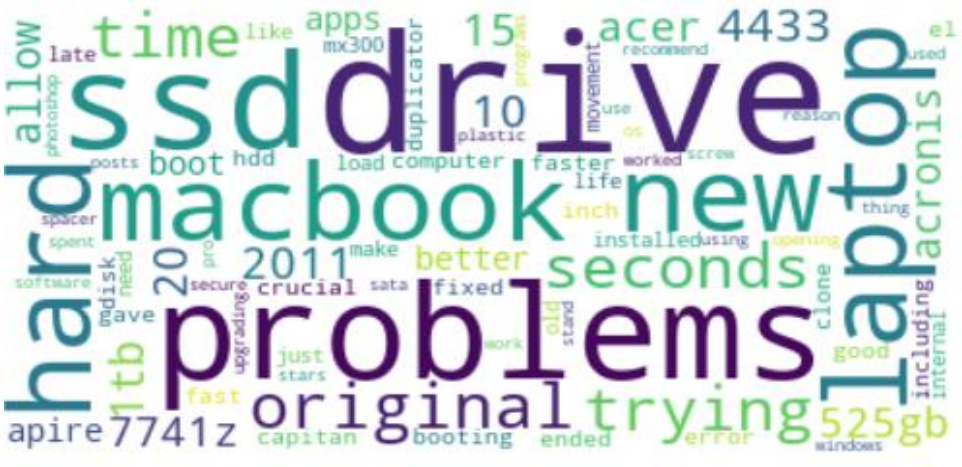

Figure 6. Chosen Set of Reviews Word Cloud

APPENDIX C: BOOK CORPUS VISUALS

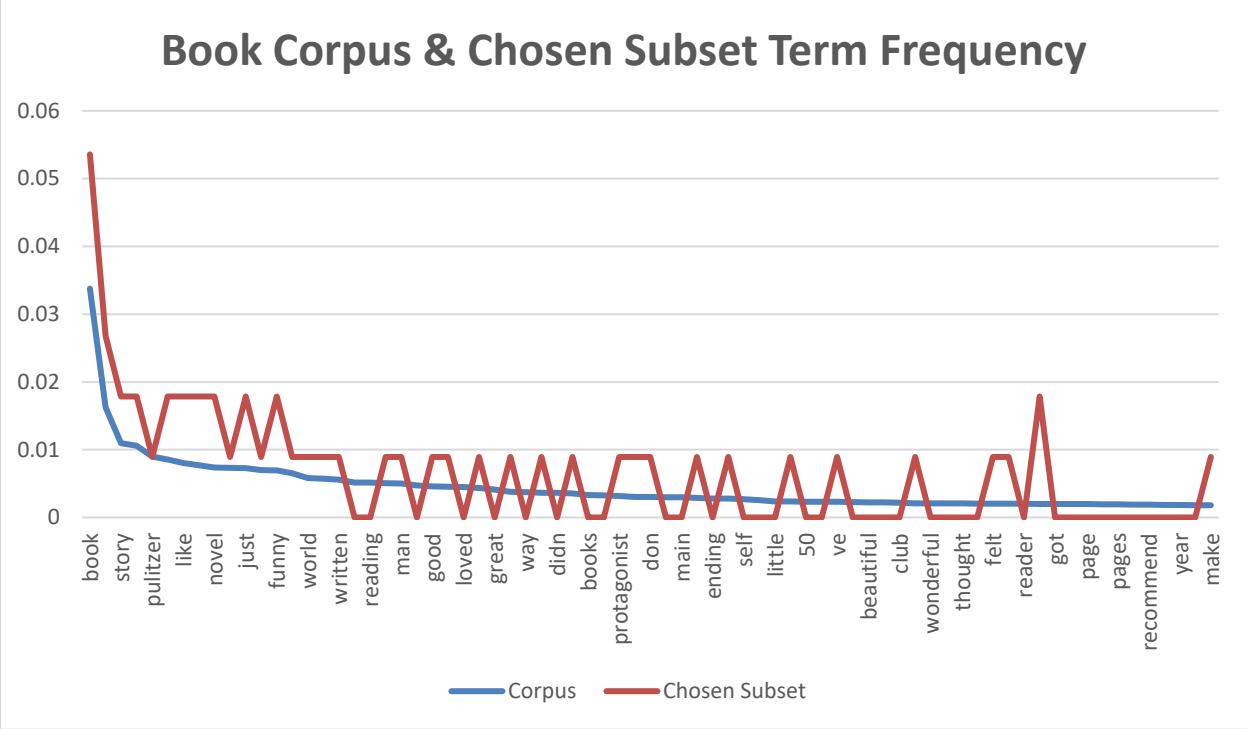

Figure 7. Book Corpus & Chosen Subset Term Frequency

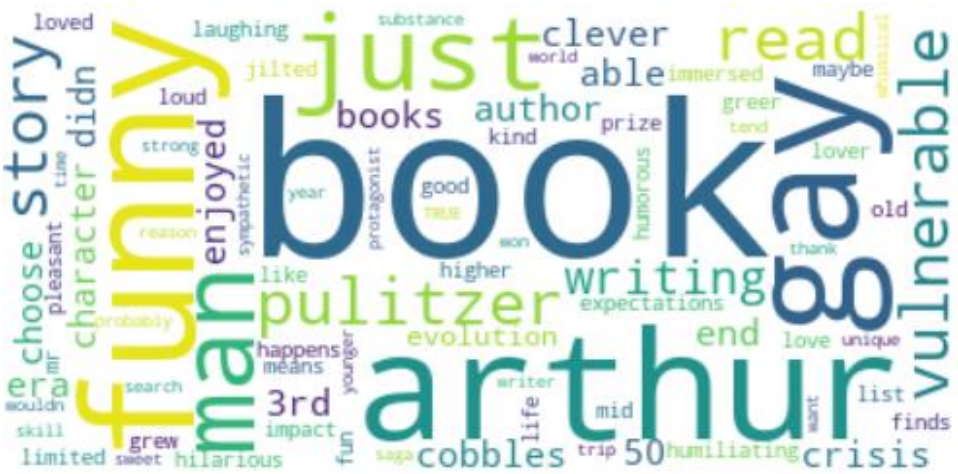

Figure 8. Book Corpus Word Cloud

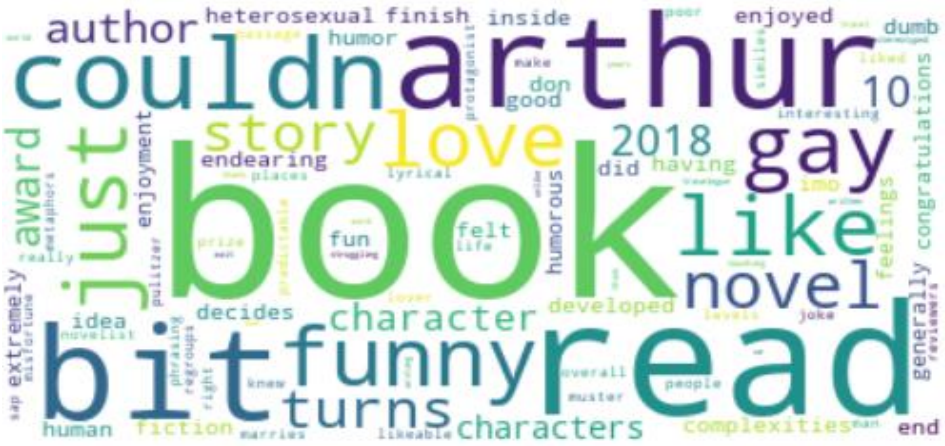

Figure 9. Book Chosen Set of Reviews Word Cloud