Title

Performing Text Mining on the Series of Rememberance of Earth's Past by Liu Cixin

Author

Niu, Naiyu

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Performing Text Mining on the Series of Rememberance of Earth's Past by Liu Cixin

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Naiyu Niu

2024

ABSTRACT OF THE THESIS

Performing Text Mining
on the Series of Rememberance of Earth's Past
by Liu Cixin

by

Naiyu Niu
Master of Applied Statistics and Data Science
University of California, Los Angeles, 2024
Professor Frederic R. Paik Schoenberg, Chair

This thesis primarily employs text mining techniques to analyze the novel series "Remembrance of Earth's Past" by Liu Cixin, with the aim of extracting information and developing a comprehensive understanding through this process. It involves extracting the most frequently occurring words, calculating tf-idf (term frequency-inverse document frequency) scores, validating Zipf's Law, generating bi-grams, conducting sentiment analysis and applying topic modeling. The findings reveal significant keywords from the novel series, allowing readers to infer the main elements. While this analysis can help the audience achieve an overall understanding of the plots, reading the original work is encouraged to have a detailed sense of the entire series.

The thesis of Naiyu Niu is approved.

Yingnian Wu

Hongquan Xu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2024

To my parents who have provided me with the utmost support,
professors who have illuminated my path through academia,
and my friends who have shared the journey in joy and tears.
And pets that I will have in times to come.

TABLE OF CONTENTS

1 Introduction **1**

2 Background **3**

2.1 Plot Summaries 3

2.1.1 The Three-Body Problem 3

2.1.2 The Dark Forest 3

2.1.3 Death's End 4

2.2 Data 5

3 Methodology **6**

3.1 Word cloud 6

3.2 Data Cleaning 8

3.3 Most Frequent Words 9

3.4 Term Frequency-Inverse Document Frequency (tf-idf) 13

3.5 Zipf's Law 16

3.6 Bi-gram 17

3.7 Sentiment Analysis 21

3.7.1 Bing and AFINN Lexicons 21

3.7.2 NRC Sentiments 24

3.8 Topic Modeling 26

4 Conclusion and Discussion **30**

References **33**

LIST OF FIGURES

3.1 Original Word Cloud of the Series Remembrance of Earth's Past 7

3.2 Most Frequent Words of The Three-Body Problem Connected at Threshold Correlation = 0.9 11

3.3 Most Frequent Words of The Dark Forest Connected at Threshold Correlation = 0.9 11

3.4 Most Frequent Words of Death's End Connected at Threshold Correlation = 0.9 12

3.5 10 Most Important Terms in Each Book 14

3.6 10 Most Important Terms for (a) TBP (b) DF (c) DE 15

3.7 Log Term Frequency-Log Rank Order Plots for (a) The whole series (b) TBP (c) DF (d) DE 17

3.8 Original Bi-grams for (a) TBP (b) DF (c) DE 18

3.9 Cleaned Bi-grams for (a) TBP (b) DF (c) DE 19

3.10 Proportion of AFINN Lexicons for (a) TBP (b) DF (c) DE 22

3.11 Sentiments with Top AFINN Scores for (a) TBP (b) DF (c) DE 23

3.12 Set of Categories of NRC Lexicons 25

3.13 Proportion of Sentiments in NRC Lexicons 25

3.14 Proportion of Emotions in NRC Lexicons 26

3.15 LDA with 3 Models 29

3.16 Confusion Matrix of LDA 29

LIST OF TABLES

3.1 Number of Characters Before and After Cleaning 9

3.2 10 Most Occurrences in Each Book 10

3.3 Counts of Bing Lexicon in Each Book 21

3.4 Definitions of Variables in LDA 27

ACKNOWLEDGMENTS

I would like to thank Dr. Yingnian Wu, Dr. Frederic R. Paik Schoenberg, and Dr. Hongquan Xu for their support and guidance in writing my thesis. I would also like to thank professor Akram M Almohalwas for his instruction during the course.

My parents have been a huge power pushing me forward. I want to thank them for offering me a life where I can do whatever comes up in my mind. I have enjoyed the maximized freedom with their love and I am incredibly grateful for everything.

Pursuing a degree at UCLA will never be such a fun journey without my friends. I appreciate everyone, especially Yuxin Chen and Edward Yen, in my cohorts. I was lost and stressed in the beginning of my second year in this program and decided to take a leave of absence. Yuxin and Edward provided such generous help that I was well updated with the news of the department and program. Everyone from my last and current cohort is welcoming and easygoing and we shared a good time in and after class. I hope we'll all step into a brighter future.

CHAPTER 1

Introduction

Liu Cixin, a Chinese computer engineer and science fiction writer and a member of China Science Writers Association, has won China's Galaxy Award nine times and has also received the 2015 Hugo Award for his novel The Three-Body Problem as well as the 2017 Locus Award for Death's End. [Wik24b]

The Three-Body Problem, Death's End, together with The Dark Forest as a transition in between form the novel series Remembrance of Earth's Past. The series were first written in Chinese subsequently in 2006, 2008, 2010 and the English translations by Ken Liu were published in 2014, 2015, 2016, and the techniques of text mining are conducted on the English version of the books.

Publications like journal articles usually have abstracts in the beginning to convey the main takeaways of the article for higher efficiency. Unlike them, before readers delve into the content of a novel, they often encounter comments and reviews about it first. These reviews normally rate novels without providing a detailed overview of the actual content. For example, according to Volodymyr Osmak, this story was more than just a thrilling adventure — it was a chance to explore big ideas... "Remembrance of Earth's Past" is a must-read for fans of hard science fiction who love stories about space, time, and what it means to be human. [Osm23] Osmak's comment can only spark the reader's curiosity, but failed to unveil any details of the story itself.

By employing text mining techniques, we can generate word clouds to visualize the content using single words. Additionally, we can analyze the most frequently occurring words to gain a general understanding of the central themes. Calculating tf-idf (term frequency-inverse document frequency) allows us to identify significant terms within each book of the

series and validate their importance using Zipf's Law. Furthermore, generating bi-grams expands our knowledge of the books beyond single keywords. We can also conduct sentiment analysis and topic modeling to deepen our understanding of the narrative.

CHAPTER 2

Background

2.1 Plot Summaries

The Plot summaries are provided to verify the results in the following chapters.

2.1.1 The Three-Body Problem

A series of physicist suicides lead police officer Shi Qiang to investigate into a group named as the Frontiers of Science. Applied physicist Wang Miao assists the investigation and delves into a virtual reality game called Three-Body simulating survival on a planet in a three-sun system. The planet is Trisolaris, four light-years away from Earth. The trisolarans seek to escape their dying planet and colonize Earth using their exceedingly advanced technology.

The origins of this conflict trace decades back to Ye Wenjie, who had suffered from inhuman tortures during the Cultural Revolution period. When she worked at Red Coast military base, she sent to the space a message that was captured by the listener 1379 on Trisolaris. The princeps of trisolarans launched four super-intelligent protons, known as sophons, to Earth, ahead of the whole population, to inhibit the advancement of human technology. Because of Ye's message, she was entitled the leader of Earth Trisolaris Organization (ETO) which was established and expanded by Mike Evans, the son of an American oil baron.

2.1.2 The Dark Forest

Trisolarans' plan of invasion is exposed to the public and therefore human nations have to unite to fight against the trisolaran fleet. The UN comes up with the project Wallfacer,

entitling four wallfacers to form their own plan of combat with limitless grant of resources, since the only capability of human that is not grasped by trisolaran is misdirection. The four wallfacers are respectively Frederick Tyler, former US Secretary of Defense, Manuel Rey Diaz, former President of Venezuela, Bill Hines, an English neuroscientist and former president of the EU, and Luo Ji, an astronomer and sociologist. Trisolarans assign three human traitors to be wallbreakers and surprisingly the wallbreaker of Luo Ji is himself. Luo Ji broadcasted the location of a star throughout the galaxy and entered hibernation.

Luo's Behavior followed the dark forest hypothesis that has two axioms: First, survival is the primary need of civilization, and second, civilization continuously grows and expands whereas the total matter in the universe remains constant.[Cix15] This hypothesis has two underlying concepts: the chain of suspicion and technological explosion. This hypothesis was suggested by Ye Wenjie when she advised Luo Ji to combine his majors and invent Cosmic Sociology. Luo Ji proved the hypothesis by the fact that the exposed star was eliminated by a hyper-advanced civilization and therefore he was able to force the trisolarans into a truce.

2.1.3 Death's End

The timeline of this book consists of six eras, respectively Common Era, Crisis Era, Deterrence Era, Broadcast Era, Bunker Era, Galaxy Era.

In Crisis Era, aeronautical engineer Cheng Xin participated in the Staircase Project to launch a probe to the trisolaran fleet and Yun Tianming who anonymously gifted Cheng Xin the distant star DX3906 donated his brain to this project due to his body illness.

In Deterrence Era, Luo Ji was designated to be the Swordholder as a result of the truce, which was passed to Cheng Xin when Luo Ji retired. Trisolarans attacked Earth at the moment of Swordholder handoff. The starship previously launched to the space managed to sent the muturally assured destruction broadcast to the space, which exposed both Earth and Trisolaris and led the trisolarans to give up conquering Earth.

In Broadcast Era, trisolarans reconstituted Yun Tianming and permitted him to meet with Cheng online, during which he told three stories to imply how to build light-speed ships

and publish the safety notice after the exposure.

In Bunker Era, most people could not survive the alien weapon approaching the solar system that collapses three-dimension into two-dimension, but Cheng Xin and AA, her assistant, rode the light-speed ship and escaped from the solar system to DX3906.

In Galaxy Era, Cheng Xin and AA met Guan Yifan on Planet Blue. Cheng and Guan got trapped in a black domain when Yun arrived at Blue. When they returned to Blue, millions of years had been gone. Yun and AA left them a portal to a micro-universe to survive. Cheng, Guan and a sophon lived in the micro-universe therefater until they received the message saying the micro-universes deprives the main universe of mass, disrupting its cycle of expansion, collapse and rebirth.[Wik24a]

2.2 Data

The research is conducted on three documents in portable document format. The Three-Body Problem has three parts: Silent Spring, Three Body, Sunset for Humanity. The Dark Forest has four parts: Prologue, The Wallfacers, The Spell, The Dark Forest. Death's End has six parts.

CHAPTER 3

Methodology

3.1 Word cloud

A word cloud, also referred to as a tag cloud or weighted list in visual design, serves as a graphical portrayal of text data. It is commonly used to illustrate keyword metadata on websites or to visualize free-form texts. Word clouds normally consist of single words whose font size and color is decided by their significance in the text piece. An early printed example of a weighted list of English keywords was the "subconscious files" in Douglas Coupland's Microserfs (1995). [Wik24d]

There are three types of word clouds. They have different meanings while their appearances stay the same. In the first type, larger sizes represent higher frequencies of texts. The second type displays a word larger than the others when it is more significant. Here, the significance is determined by the term frequency-inverse document frequency (tf-idf) score. The third type uses categorization and a larger size imply a larger quantity of items in that category.

In this section, we will employ the "worldcloud" package [Fel18] and the "tidytext" package in R [SR16] and apply the first type of word cloud on the three books. In principle, the display font size s_i can be calculated as the following formula: let f_{max} be the maximum font size, t_i the count, t_{min} the minimum count and t_{max} the maximum count, then

$$s_i = \left\lceil \frac{f_{max} \cdot (t_i - t_{min})}{t_{max} - t_{min}} \right\rceil$$

for $t_i > t_{min}$; otherwise, $s_i = 1$.

From Figure 3.1, we can observe that the words are roughly same sized and in the same

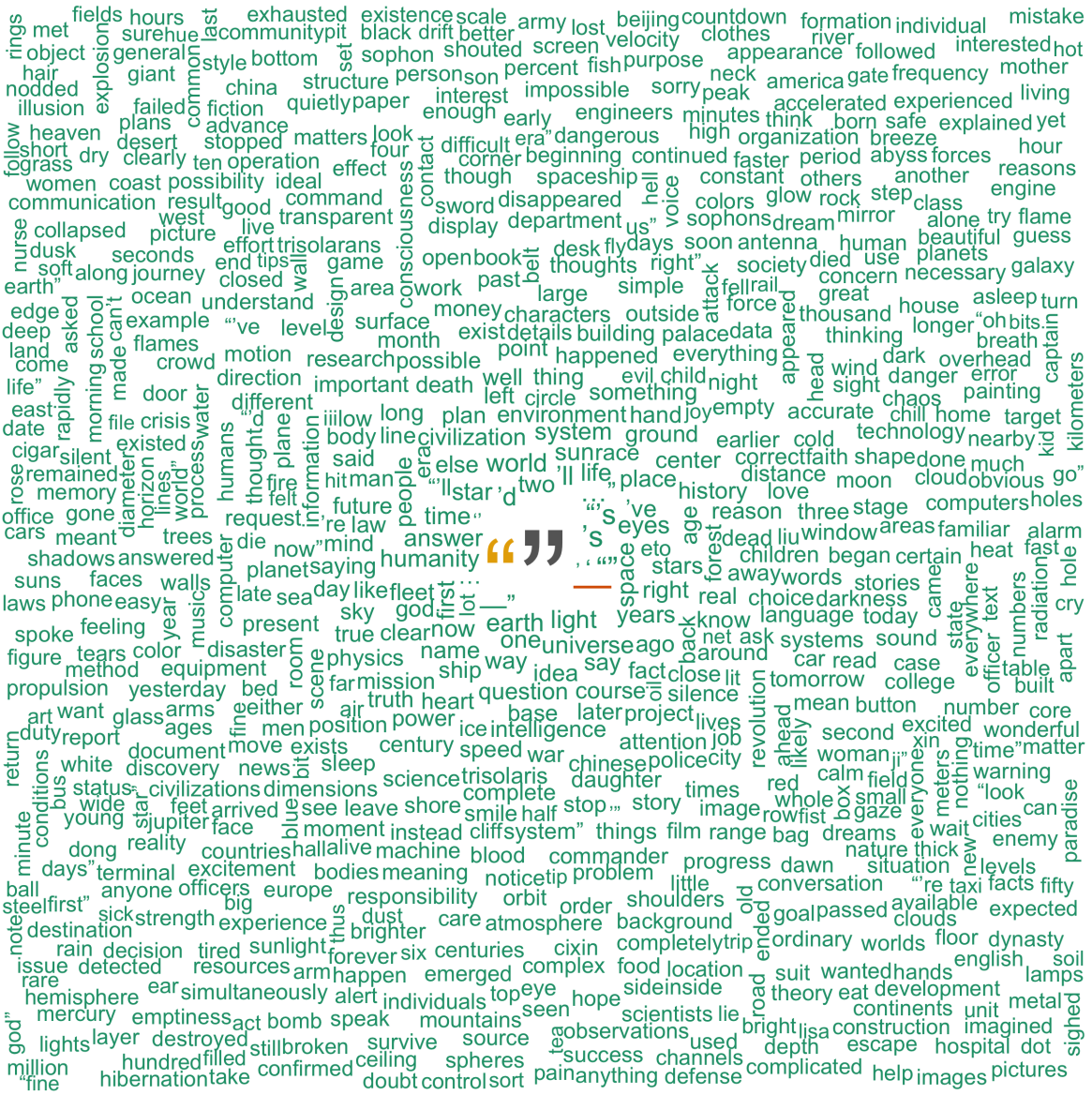


Figure 3.1: Original Word Cloud of the Series Remembrance of Earth's Past

color and the elements that have different colors than the words are quotation marks and dashes. Apart from the punctuation, words of common usage appear frequently in the word cloud that may not convey particular information about the novels.

3.2 Data Cleaning

Data cleaning is applied for the purpose of obtaining words that have a more specific description of our original texts. Before cleaning, we need to prepare corpora. A text corpus is a vast and unorganized collection of texts utilized for statistical analysis and hypothesis testing purposes. With respect to a corpus, a term-document matrix (TDM) can be extracted with each row having one term and each column one document. The entries in a TDM is the number of occurrences for the term in that document.

The cleaning process utilizes the "tm" package in R [FH24] and has eight steps:

1. Simple Transformation. In this step, we will map the symbols in the corpus to a white space with the function "tm_map."

2. Conversion to Lower Case. The conversion avoids the repetition of the same word in the word cloud. For example, appearances of either "Three" or "three" account for the frequency of "three."

3. Remove Punctuation. This step removes the punctuation recognized by the program, such as colons and periods.

4. Specific Transformation. After the removal of normal punctuation, curly quotation marks are still observed in the corpus. Therefore, the unrecognized punctuation in the previous step is targeted in this step.

5. Remove English Stop Words. Stop words are the commonly used vocabularies and usually filtered out during natural language processing. In English, stop words are composed of articles such as "a" and "an," prepositions such as "in" and "on", pronouns

	Before Cleaning	After Cleaning
The Three-Body Problem	697506	388290
The Dark Forest	1115749	592003
Death's End	1329479	737817

Table 3.1: Number of Characters Before and After Cleaning

such as "this" and "these," etc. The "tm" package [FH24] includes 175 English stop words in total.

6. Remove Own Stop Words. This step is to remove the words that convey little information but have high frequencies tailored to the content. For example, in the corpus, "can," "said," "like," and many other consistently occurring words are listed as my own stop words to be removed.

7. Strip White Space. In the documents, there are empty pages serving for separation of sections. The step of stripping white space will collapse multiple whitespace characters into a single blank and reduce the frequency of whitespace.

8. Stemming. Stemming uses an algorithm to remove the common word endings such as "ed" and "s" and return verbs to infinitive forms. This step is to eliminate the difference between words such as "physicist" and "physicists," "bring" and "brought".

Table 3.1 compares the number of characters of each book before and after cleaning. The number of characters goes down by 44.33% in The Three-Body Problem, 46.94% in The Dark Forest, and 44.50% in Death's End.

3.3 Most Frequent Words

Table 3.2 lists side by side the terms extracted from the results of inspection of the term-document matrices with respect to each book separately. It indicates an overview of the words after data cleaning. It is easily observed that in all three books the common words including conjunctions and pronouns are erased and the terms left in the texts have a potential

The Three-Body Problem		The Dark Forest		Death's End	
Before	**After**	**Before**	**After**	**Before**	**After**
and	civil	and	fleet	and	cheng
but	one	but	human	but	xin
for	red	for	look	cheng	human
had	shi	had	luo	for	one
she	sun	his	one	from	ship
that	three	that	shi	had	space
the	time	the	space	she	system
this	two	was	time	that	time
was	wang	with	two	the	two
you	will	you	will	was	world

Table 3.2: 10 Most Occurrences in Each Book

direction. For example, it exhibits the names of the main characters in each book.

In figure 3.2, every node represents one of the most frequent words and every connection in between indicates high correlation of the two words. We can tell from the figure that every most frequent word is highly correlated with others and it means that all characters are closely related to the main plot and there is only one main story line. Apart from the two main characters Shi Qiang and Wang Miao showing up in the figure, one can observe other useful information. For example, the node "red" is related to the "Red Coast military base" where the first message was sent to the outer space and the nodes "three" and "sun" support a reasonable deduction of the three-sun stellar system. The nodes "human," "civil," "world," "light" and "time" lead the audience to the direction of light-speed travels, birth and death of a civilization, incidents that matters to every human species in the world, a story with a long time span.

Similarly in figure 3.3, the most frequent words appear in the same cluster with multiple connections in between. It again suggests one main story line that involves every main character. From the figure we can see three most important characters: Shi Qiang, Luo Ji,

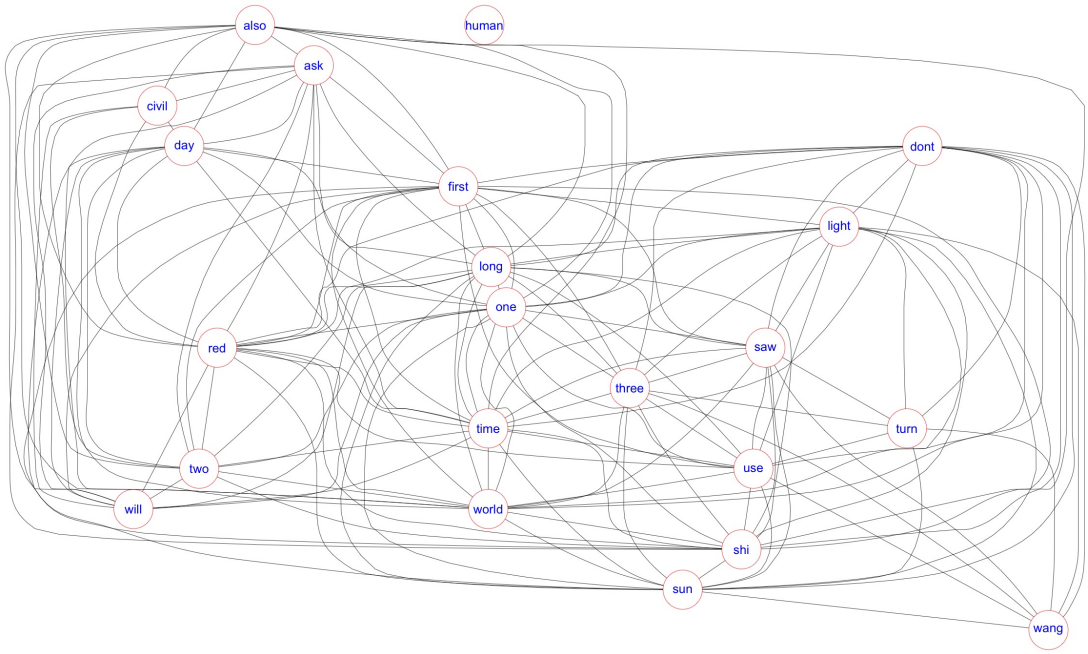


Figure 3.2: Most Frequent Words of The Three-Body Problem Connected at Threshold Correlation = 0.9

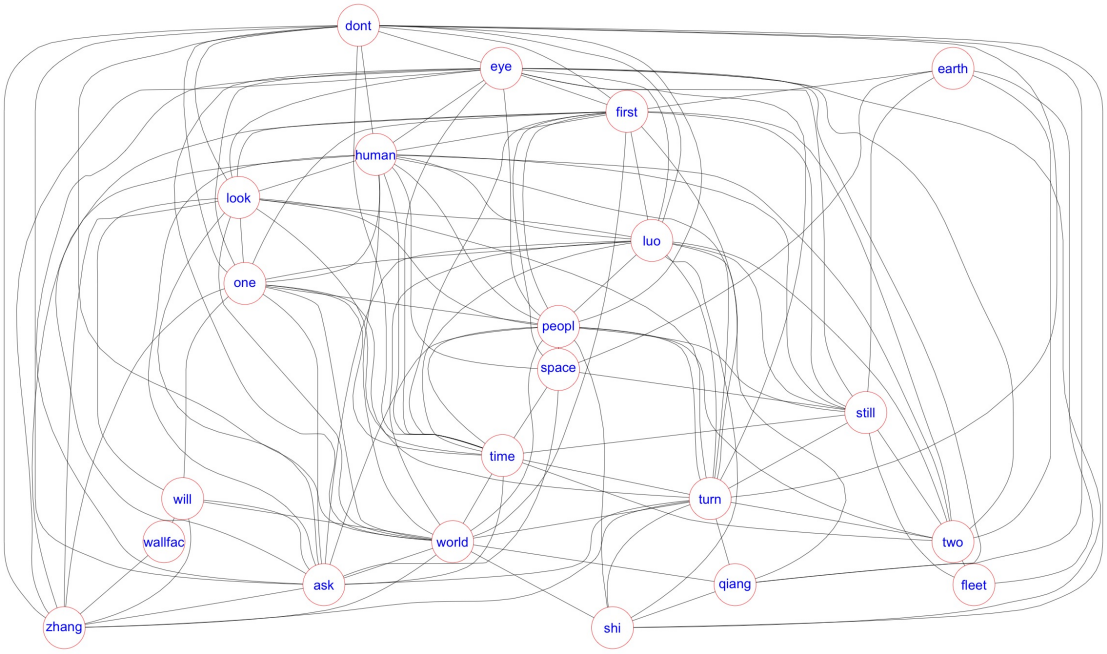


Figure 3.3: Most Frequent Words of The Dark Forest Connected at Threshold Correlation = 0.9

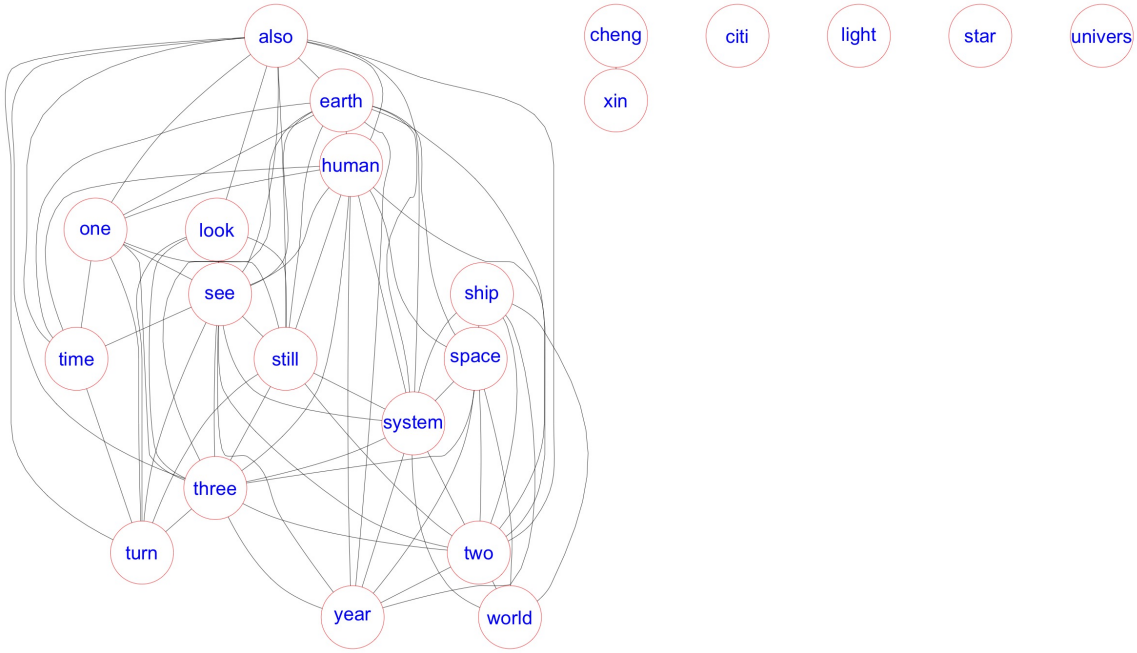


Figure 3.4: Most Frequent Words of Death's End Connected at Threshold Correlation = 0.9

Zhang Beihai. What's more, the "Wallfacer" project is noted in the plot, which provides a crucial message and a precise angle for the audience to acquire the main content. "Fleet" and "earth" in the figure can leave questions such as "Is it an earth fleet or alien fleet?" or "Is it heading to the earth or the outer space?" One can also learn that the main story line is unfolded around the two concepts "space" and "time." The nodes "human," "people," "world" remind the audience of the previous book as they also appear in figure 3.2 and suggests that the story in the second book is possibly developed from a context similar to the first one.

Unlike the previous two figures, the most frequent words of Death's End are separated into a larger group and several nodes aside. The major group obviously points out the three-sum stellar system with the nodes "three" and "system" implying that the Trisolaris plays an important role in this book. One can also discover the combination "space ship" and speculate space travels of human. The name "Cheng Xin" stands out, indicating that she is the main character throughout this book. The nodes aside the main cluster imply that the story line around the human species and the earth may end at some point and the main

character Cheng Xin continues her experience in the universe.

3.4 Term Frequency-Inverse Document Frequency (tf-idf)

The term frequency-inverse document frequency (tf-idf) is the product of two statistics, term frequency (tf) and inverse document frequency (idf). Here, the tf is the raw count of a term in a document. The inverse document frequency (idf) is a measure of how common or rare a term is across all documents and is calculated as the following:

$$\text{idf}(t, D) = log\left(\frac{N}{|\{d : d \in Dandt \in d\}|}\right)$$

where N is the total number of documents in the corpus with $N = |D|$ and $|\{d : d \in Dandt \in d\}|$ is the number of documents where the term t appears. It is only reasonable to calculate the statistics when considering an existing term t in the document d. The idf decreases when the term is common among the documents and accordingly the product tf-idf decreases, which is a sign for less important words.

To prepare the data for the calculation, we need to use the "unnest_token()" function in the "tidytext" package.[SR16] This function performs tokenization on the texts. A token is a meaningful unit of text, i.e., a word, and tokenization splits the texts into tokens. By appling the function, we end up with a table with one-token-per-row. The tf-idf score is computed using the "bind_tf_idf()" function in the "tidytext" package.[SR16]

Figure 3.5 lists the most important terms in each book according to descending tf-idf scores. The titles are the capitals of the book names. First of all, the rank of importance discovers the main characters Wang Miao and Shi Qiang in the first, Luo Ji and Shi Qiang in the second, and Cheng Xin in the third, which is consistent with the findings from the most occurrences and in addition emphasizes their unique importance in the corresponding book.

In particular, in the list for the first book, "princeps" is the leader of trisolarans, "lei" is the last name of Lei Zhicheng who was the superior of Ye Wenjie back at Red Coast, "shen" and "pan" lead to two important members of the group Frontiers of Science, Shen Yufei and

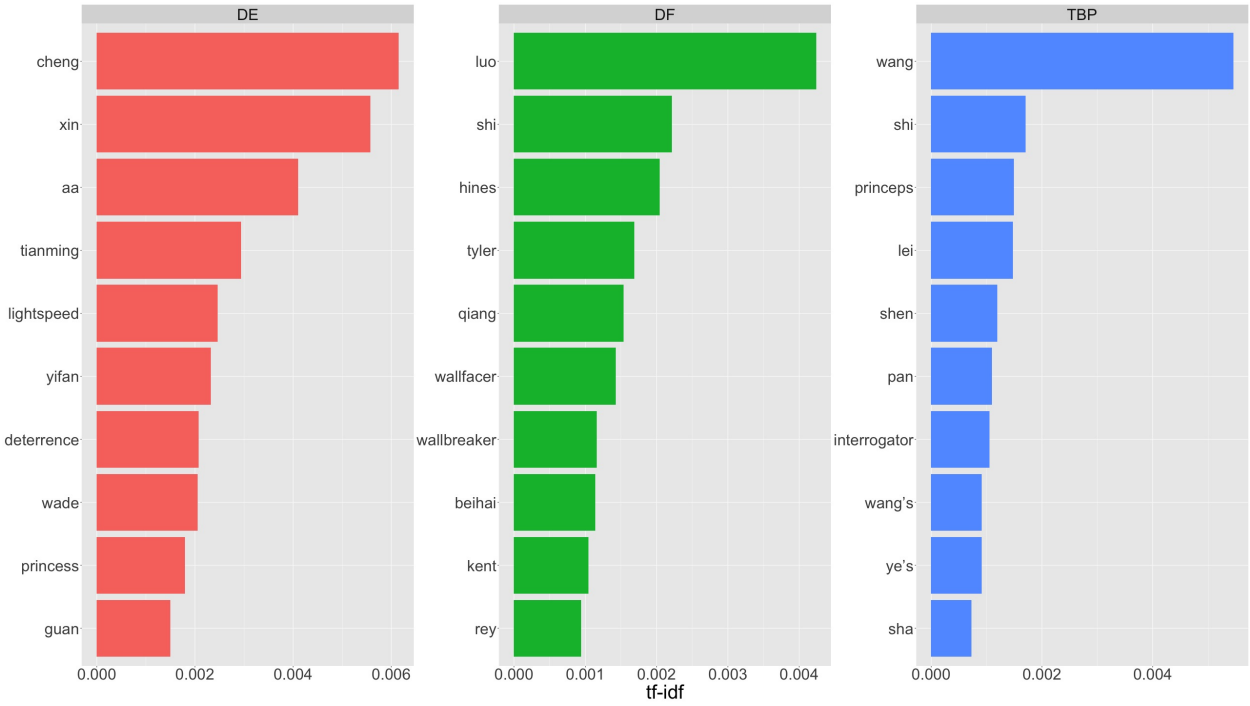


Figure 3.5: 10 Most Important Terms in Each Book

Pan Han, and "interrogator" implies the interrogation contains exclusively important information about the first book. In the list for the second book, "wallfacer" and "wallbreaker" signify the human stratagem against the alien invasion and the response of trisolarans. The names of the four designated wallfacers all show up. Zhang Beihai is an important character in the second book because he is the captain of Natural Selection Ship, which sent the broadcast that terminates the attack and invasion of trisolarans. In the list for the third book, lightspeed is an important concept and it can be related with the development of light-speed travels. The deterrence era is a time period when Luo Ji successfully forced the trisolarans to a truce and human enjoyed the temporary peace and earned the opportunity to develop technology with the guidance of trisolarans. Princess is a key to the stories Yun told Cheng, which enlightened people of the safety notice and the construction of light-speed star ships.

One can explore more detailed information by plotting terms with the highest tf-idf scores within each book. Figure 3.6 exhibits the study of the most important terms of every part within each book.

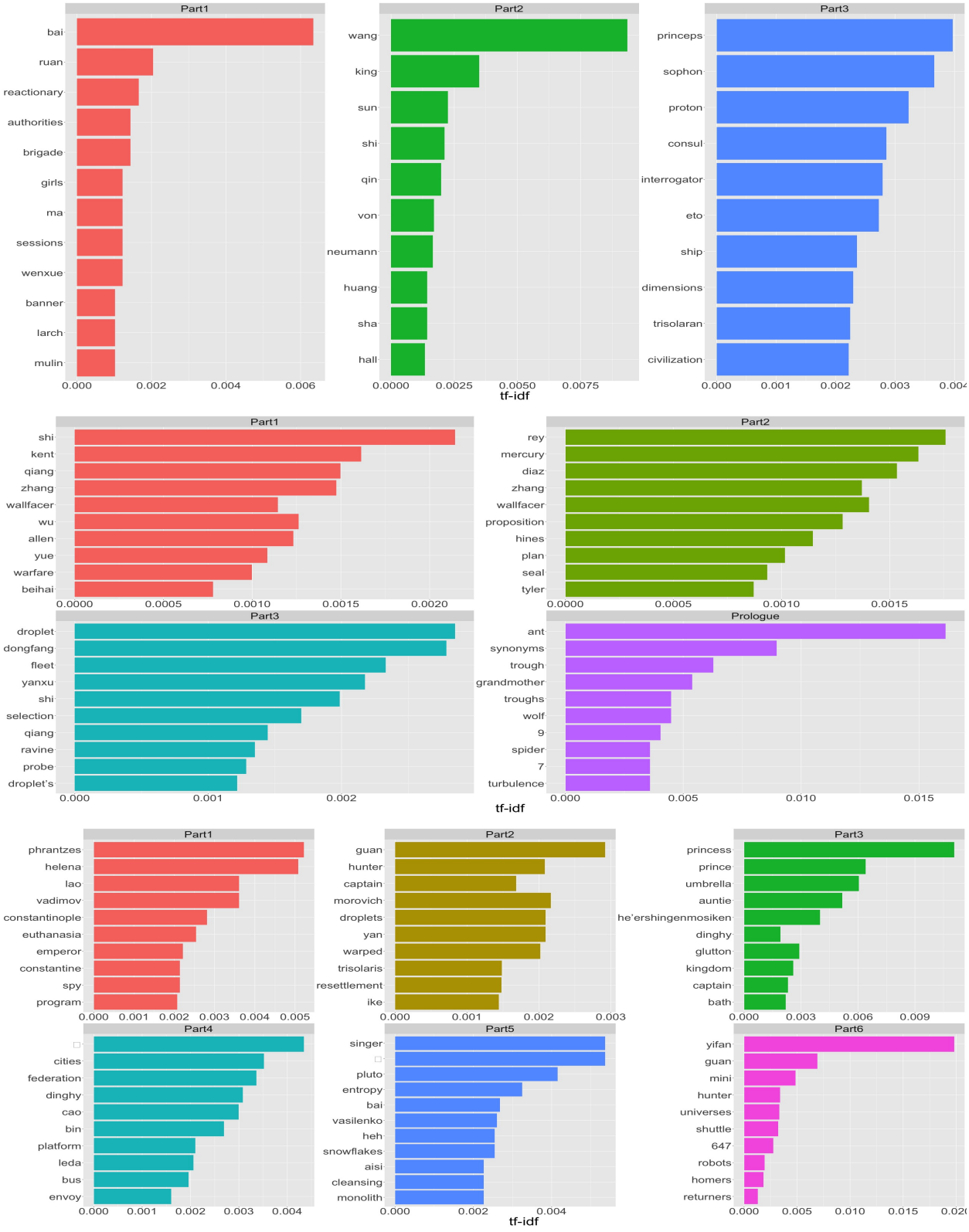


Figure 3.6: 10 Most Important Terms for (a) TBP (b) DF (c) DE

3.5 Zipf's Law

The Zipf distribution on N elements assigns to the element of rank $k \geq 1$ the probability

$$f(k; N) = \frac{1}{H_N} \frac{1}{k}$$

where H_N is a normalization constant with $H_N = \sum_{k=1}^{N} \frac{1}{k}$. The equation can be generalized to $f(k; N, s) = \frac{1}{H_{N,s}} \frac{1}{k^s}$, for $s < 1$, where $H_{N,s} = \sum_{k=1}^{N} \frac{1}{k^s}$. When $s > 1$, the generalized Zipf distribution can be extended to infinitely many as N approaches infinity, where the generalized harmonic number $H_{N,s}$ becomes Riemann's Zeta function:

$$\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s} < \infty$$

Zipf's law is an empirical law and states that the states that the value of the n-th entry is inversely proportional to n.[Wik24e]

$$\text{word frequency} \propto \frac{1}{\text{word rank}}$$

Zipf's law can be visualized by plotting a log-log graph with logarithm of term frequency versus logarithm of rank order. The data with exponent s follow Zipf's law when the plot approximates a linear function with slope $-s$.

Figure 3.7 shows that in either case, whether the whole series or each book is examined, Zipf's law holds true for the target data as the lines grows from top left to bottom right and demonstrating the negative correlation of the logarithm of term frequency with the logarithm of rank order.

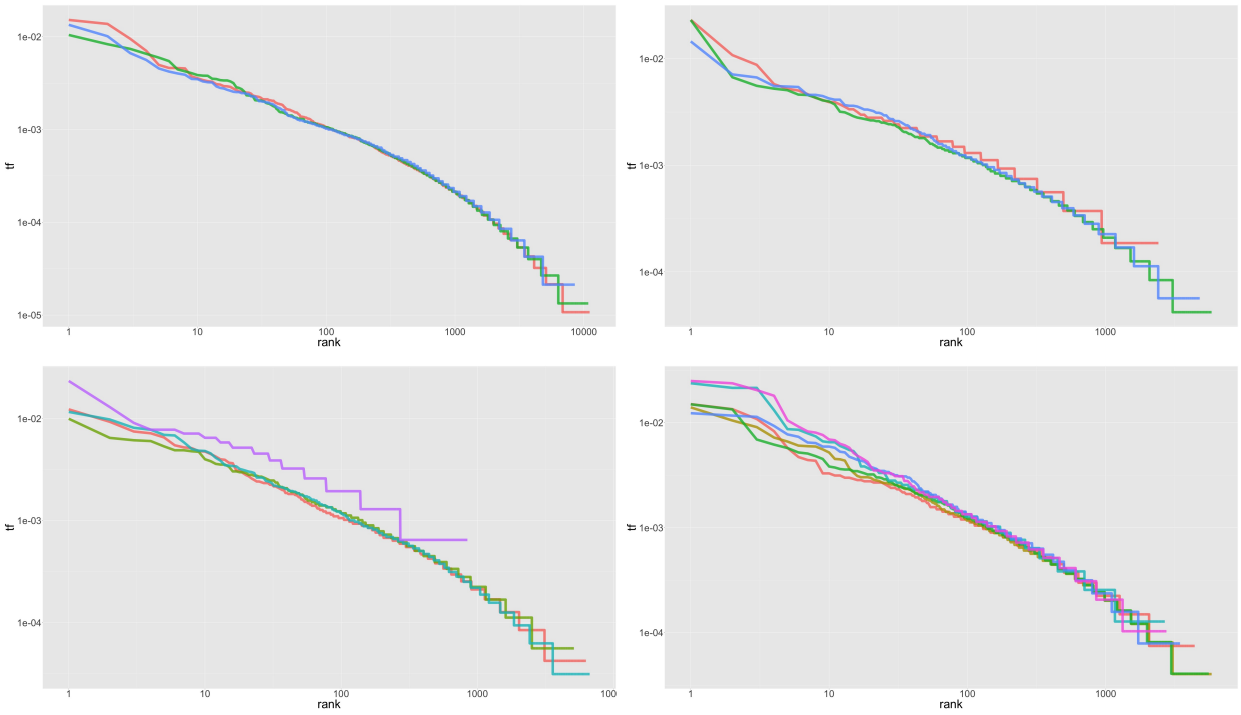


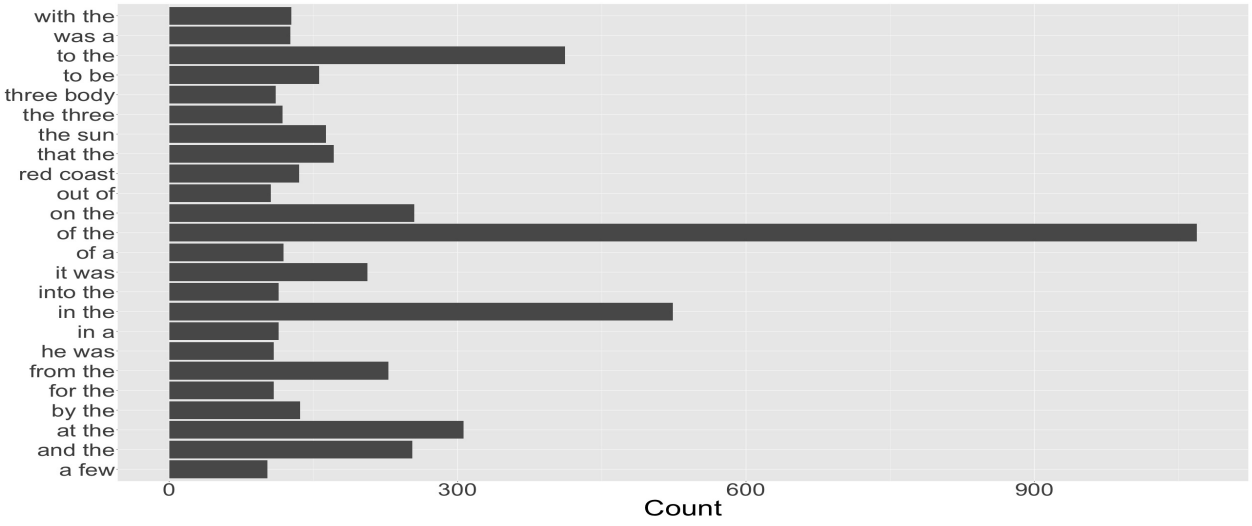
Figure 3.7: Log Term Frequency-Log Rank Order Plots for (a) The whole series (b) TBP (c) DF (d) DE

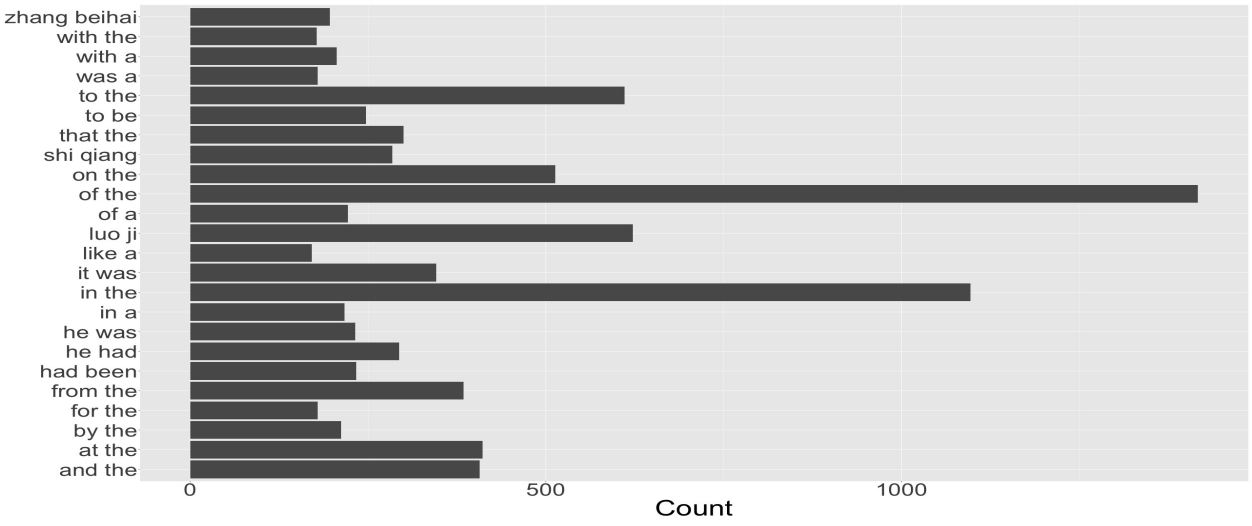
3.6 Bi-gram

An n-gram in the context of text analysis is a sequence of n adjacent terms. For example, there are 13 tri-grams in the previous sentence. N-grams are useful in the process of analysis when sequences of words are relevant. Here, we study bi-grams within each book of the series.

Figure 3.8 reveals the most frequent bi-grams in each book, where the useless bi-grams take a large proportion, such as "to the," "it was" and "of a." In this case, a similar data cleaning process is required to exclude the commonly used words. Figure 3.9 plots the most frequent bi-grams after cleaning.

In the first book, von Neumann, Qin Shi Huang, King Wen of Zhou are simulated game characters in the virtual reality game Three Body. The stable era is the simulation of a stable circulation of sunrise and sunset when people have the opportunities to develop their capabilities, whereas people cannot predict sunrise and sunset in the chaotic era nor even





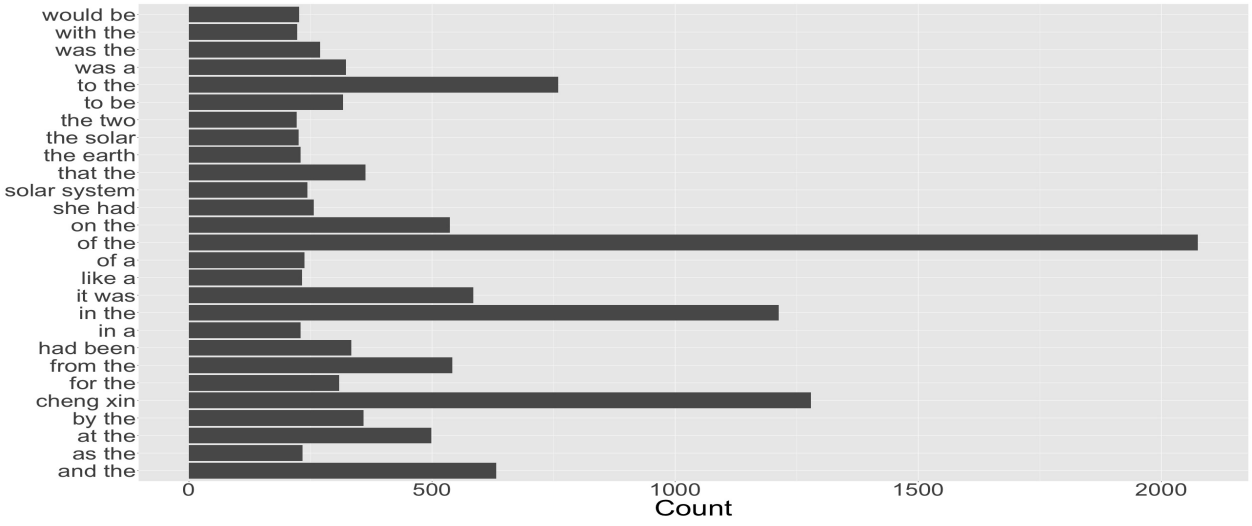


Figure 3.8: Original Bi-grams for (a) TBP (b) DF (c) DE

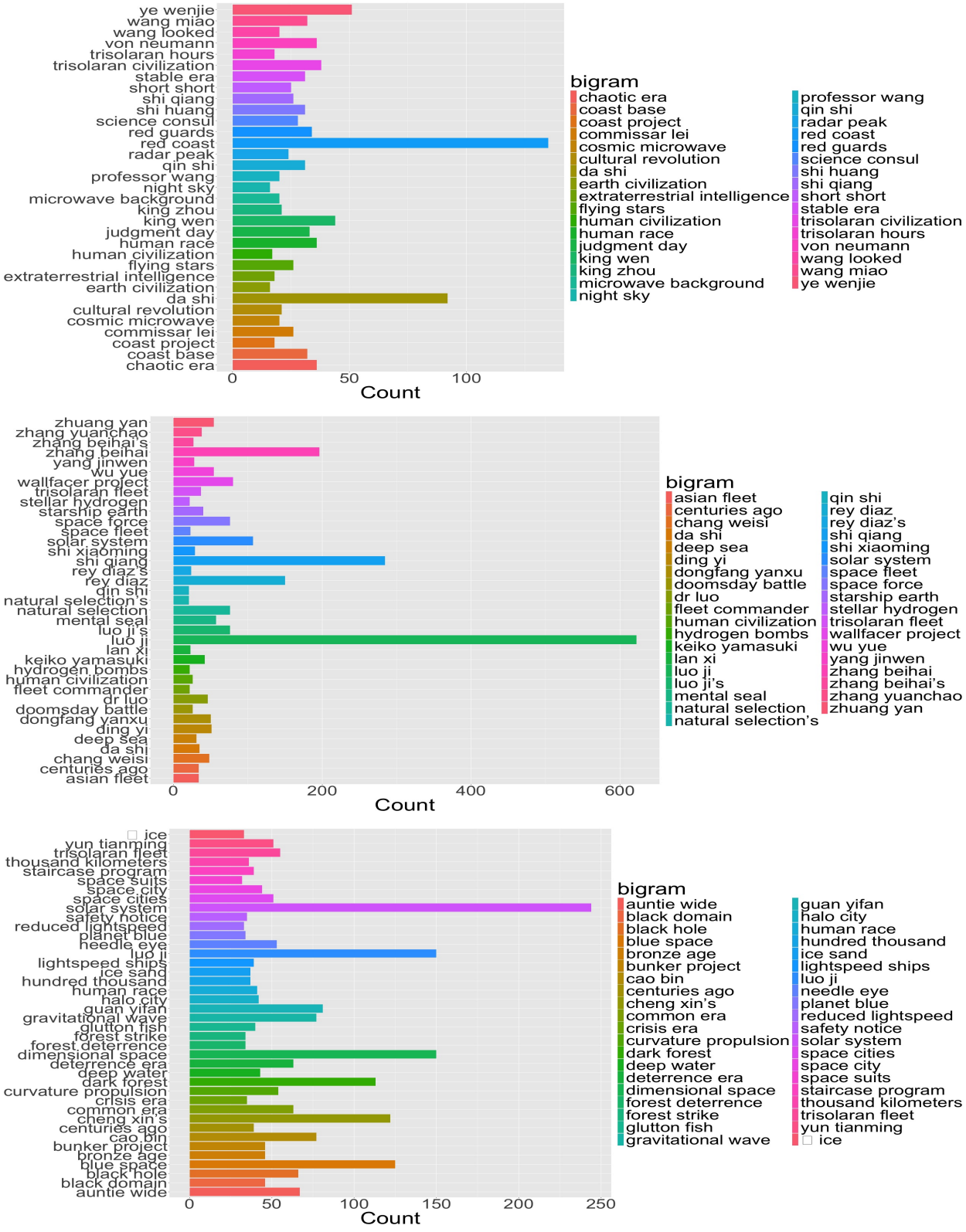


Figure 3.9: Cleaned Bi-grams for (a) TBP (b) DF (c) DE

the number of suns in the sky. Radar peak refers to Red Coast military base by the ordinary people when it was the top secret. Cultural revolution is when Ye Wenjie experienced inhuman tortures both physically and mentally; the experience results in Ye's hope that an advanced alien civilization could conquer Earth and discipline the human race. Cosmic microwave background is related to the principle Ye utilized to send her broadcast. Trisolaran civilization can be related with extraterrestrial intelligence, which helps the audience to define trisolaran. Science consul is a position among trisolarans. Judgment day is the name of a ship where Mike Evans established and kept Earth Trisolaris Organization (ETO) and where they contact trisolarans to assist with the conquering.

In the second book, Chang Weisi is one of the commanders on Earth. Keiko Yamasuki is the wife and wallbreaker of Bill Hines, an neuroscientist and former President of EU, who invented mental seals to make sealed person believe that the human race will eventually win. Doomsday battle is easily deciphered to be a battle between trisolarans and human. Space fleet and starship earth reasonably construct a speculation of the earth fleet. Following the speculation, one of the starships is named as Natural Selection. This is consistent with previous analysis and it stands out because the captain sent a broadcast that stopped trisolarans' attack.

In the third book, the main plots may be concentrated in the chapters about Common Era, Deterrence Era, Crisis Era as they are marked out. Staircase program is proposed by Cheng Xin and executed by Yun Tianming to absorb information from trisolaran fleet. Bunker project is the solution of human when the location of Trisolaris is broadcast to the universe to hide away from the obliteration of a hyper-advanced civilization. Auntie wide and needle eye are characters of the stories that Yun Tianming told Cheng Xin, which also mention glutton fish. Based on the previously mentioned stories, light-speed ships are finally invented. The notion of dimensional space emphasizes the concept of dimension and one may infer experience in a space of different dimension or the change of dimension of space from the phrase. The curvature propulsion is related to the black domain, which Cheng Xin encountered when she inspected alien traces on a planet with Guan Yifan.

	Positive	Negative
The Three-Body Problem	417	750
The Dark Forest	593	1061
Death's End	801	1330

Table 3.3: Counts of Bing Lexicon in Each Book

3.7 Sentiment Analysis

Sentiment analysis is a useful technique of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.[Wik24c] The basic analysis regarding a text file is to break it down to two polar sentiments, good or bad, positive or negative; everything that cannot be categorized into the polarities is labeled as neutral. A more complex analysis on the text evolves in more emotion categories, such as anger, disgust, sadness, enjoyment, and surprise.

This study explores the three most commonly used lexicons: AFINN, Bing, and NRC. We will conduct AFINN and Bing sentiment analysis on each book to compare their results and conduct NRC sentiment analysis on the whole series.

3.7.1 Bing and AFINN Lexicons

Bing lexicons[Liu12] present the most basic categorization of dividing words into positive and negative groups. AFINN lexicons[Nie11] consist of a list of English terms that are manually rated with an integer between -5 and 5 by Finn Årup Nielsen where negative scores label negative sentiments and positive scores indicate positive sentiments.

According to table 3.3, in the Bing sentiment analysis, 64.27% of the texts in the first book is classified as the negative sentiment versus only 35.73% is classified as the positive sentiment, 64.15% of the texts in the second book is negative versus 35.85% is positive, and 62.41% of the texts in the third book is negative versus 37.59% is positive.

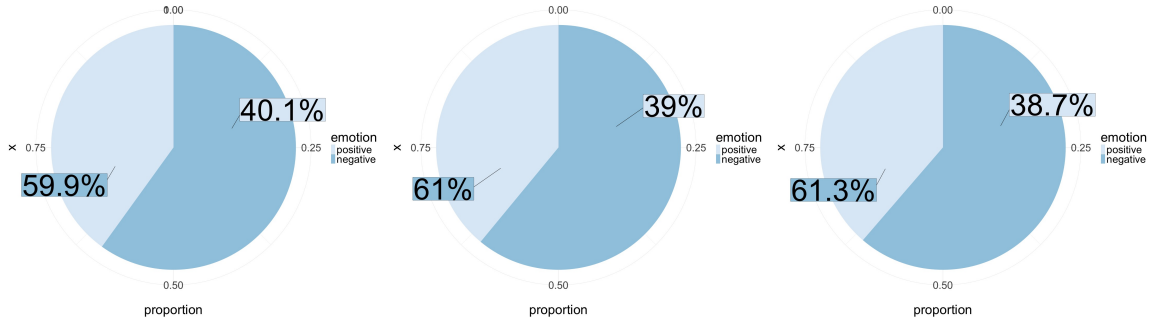


Figure 3.10: Proportion of AFINN Lexicons for (a) TBP (b) DF (c) DE

Figure 3.10 unveils the proportion of the positive and negative sentiments in each book without consideration of their manually rated AFINN scores. The first book has 40.1% of texts being counted for the positive sentiments and 59.9% for the negative sentiments; the second book has 39% of texts being counted for the positive sentiments and 61% for the negative sentiments; the third book has 38.7% of texts being counted for the positive sentiments and 61.3% for the negative sentiments.

Comparing the proportions from Bing and AFINN analyses, we can see that with both lexicons, the proportion of positiveness and negativeness stays relatively stable between the first book and the second book. However, Bing lexicons find a notably larger proportion of positive sentiments in the third book than the first two, while AFINN lexicons demonstrates an insignificant decrease of the proportion of positive sentiments in the third book in contrast to the first two. In the eye of Bing lexicons, the story in the third book grows to a brighter end and the audience would be able to sense the latent positiveness. On the other hand, AFINN lexicons suggests that the story throughout the whole series stays roughly stable on positiveness and negativeness.

Figure 3.11 takes into consideration the manually rated AFINN Scores when examining the texts of the three books and specifies the most occurrences with their contribution to the positive and negative sentiments. It can be observed that the proportion of positiveness and negativeness in every plot is close to each other, which is consistent with the stable numbers in figure 3.10. However, among the most occurrences, more positive sentiments get recorded

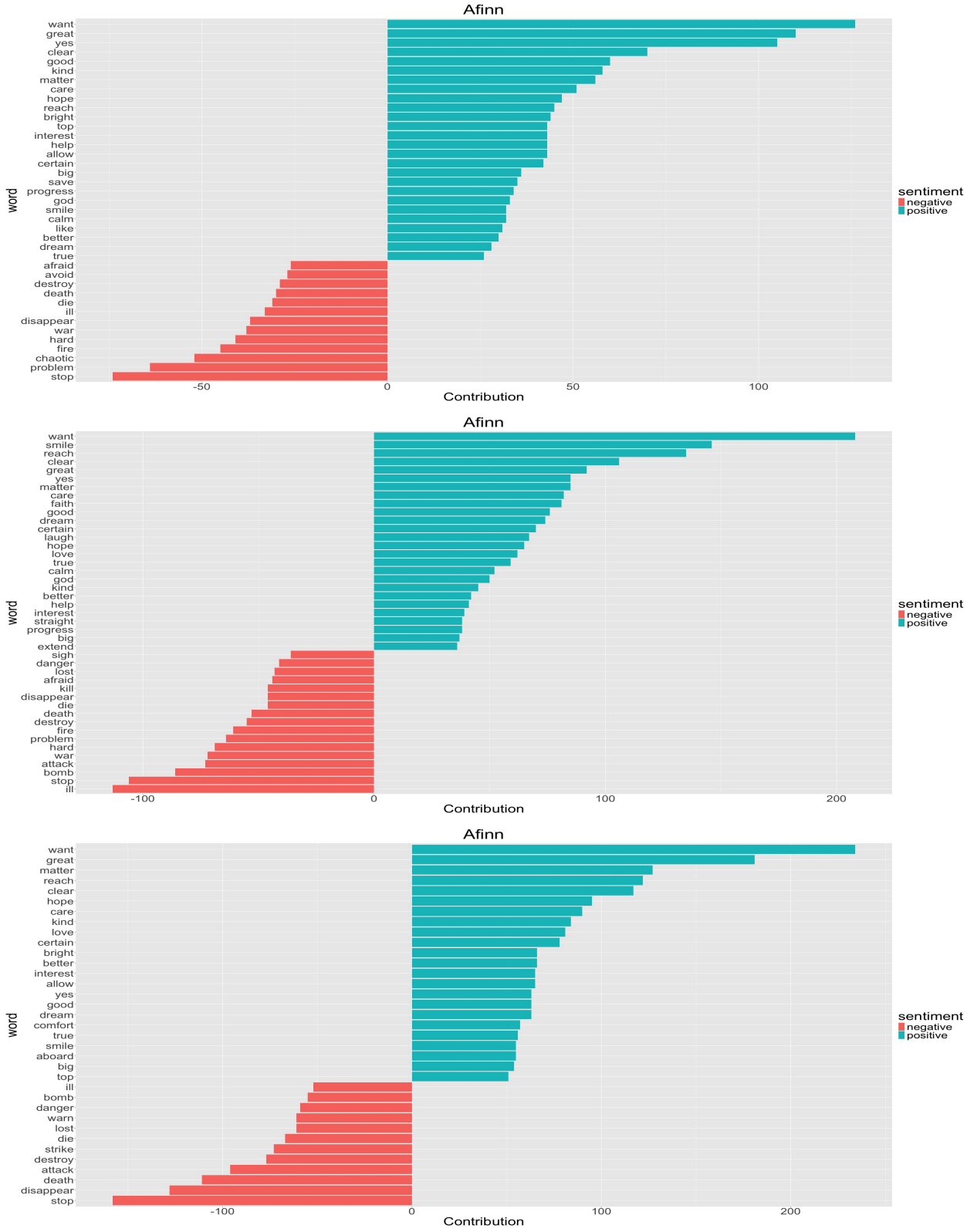


Figure 3.11: Sentiments with Top AFINN Scores for (a) TBP (b) DF (c) DE

than negative sentiments, which implies that the majority of negative sentiments are less frequent.

The positive sentiments with the most contribution to the first book include hope, bright, save, progress, better, and dream, which are hints for the audience to imagine within the scope of the first book human make good progress in their way of countering trisolarans and some character may save the world for the human race entering a brighter tomorrow. The negative sentiments with the most contribution include destroy, death, war, and ill. They point to a guess of war that causes illness and death and destroys people's normal life. The positive sentiments in the second and third book resemble the first one and similarly do the negative sentiments. The audience may use the top rated terms to expand their assumption of the development of plots.

3.7.2 NRC Sentiments

NRC Word-Emotion Association Lexicon[MT10], also called EmoLex, has 10 categories in total where 2 categories are sentiments and 8 categories are emotions. The 2 sentiments are positive and negative and the 8 emotions are respectively anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Figure 3.12[Moh22] is a treemap with sets of categories.

Figure 3.13 plots the proportion of positive and negative sentiments with respect to the whole series. 54% of the texts is considered conveying negative sentiments while 46% is considered positive. This is consistent with the previous individual sentiment analyses using Bing and AFINN lexicons that confirm a larger proportion of negativeness than positiveness.

Figure 3.14 reveals the distribution of the emotions with respect to the whole series. We can observe from the plot that the most proportion 17.2% of the text content is categorized as fear, followed by 16.3% of the content being trust, 13.3% anger, 13.2% anticipation, 13% sadness, 9.7% joy, 9.3% disgust, and 8.1% surprise. The emotions provide a supplementary angle of comprehension of the content as they can well support the deduction of a inevitable battle with aliens and the great unity of human through the emotions fear and trust. Along with the other emotions, anger and sadness leave hints about the conflicts within the human

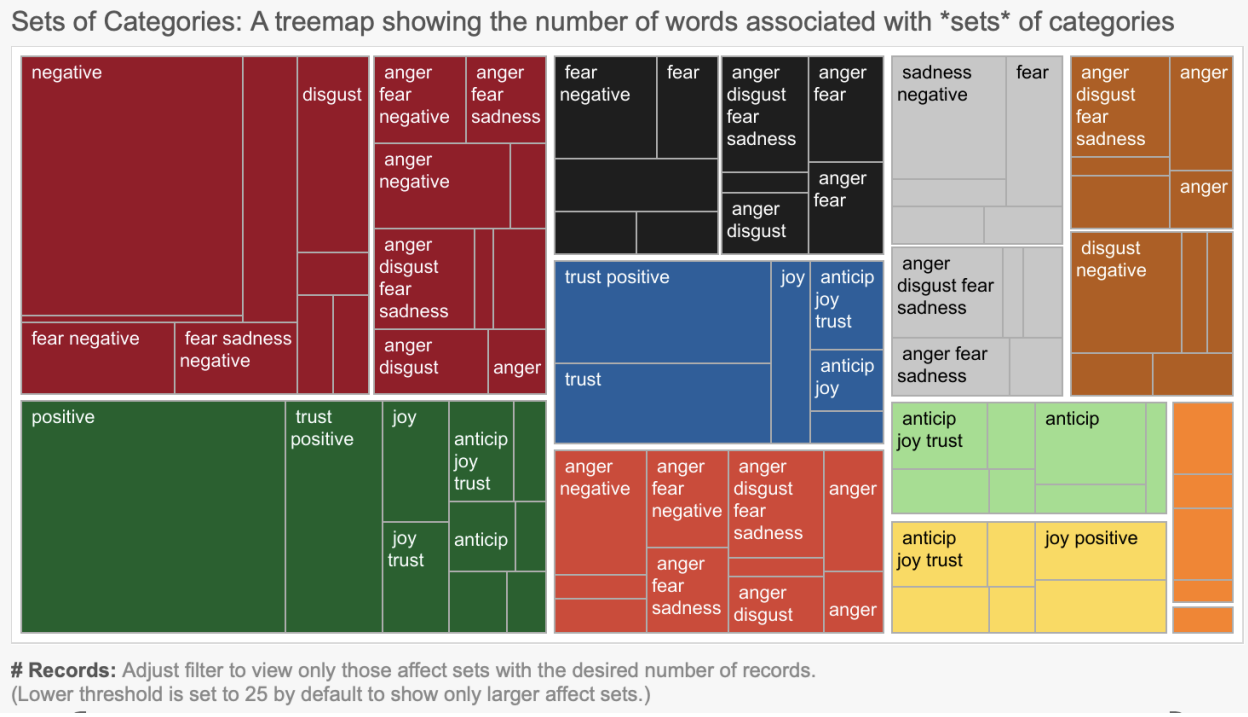


Figure 3.12: Set of Categories of NRC Lexicons

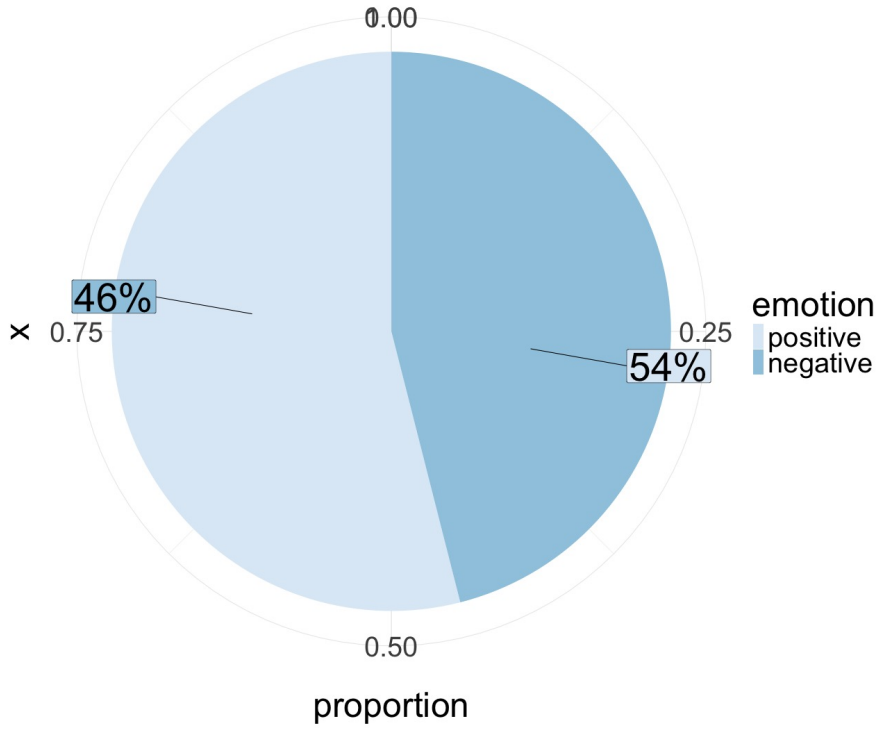


Figure 3.13: Proportion of Sentiments in NRC Lexicons

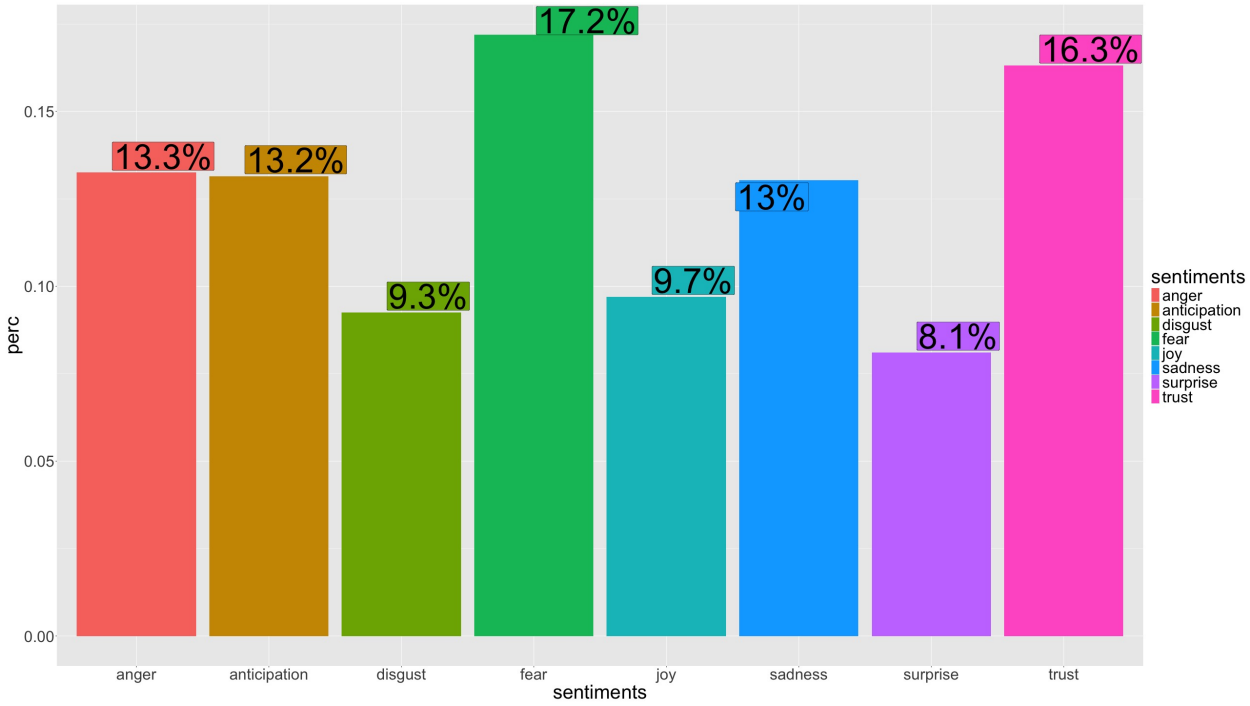


Figure 3.14: Proportion of Emotions in NRC Lexicons

race, such as Earth Trisolaris Organization gathering the betrayers of the earth and the arguments and fights between different stances such as hawkish and dovish factions.

3.8 Topic Modeling

A topic model is a type of statiscal model that uses unsupervised classification to detect the abstract topics in a group of documents. Topic modeling is frequently used in the field of text mining and in this study we will apply Latent Dirichlet Allocation (LDA) to analyze the topics in the novel series Remembrance of Earth's Past.

Latent Dirichlet Allocation is a Bayesian network for automatic modeling in text corpora. It allows the content of the texts to overlap while it identifies the topic from each piece. LDA mainly has four assumptions:

1. The semantic content of a document is composed by combining one or more terms from one or more topics.

Variable	Type	Meaning
K	integer	number of topics
V	integer	number of documents
M	integer	number of words in document d
$N_{d=1...M}$	integer	total number of words with $N = \sum_{d=1}^{M} N_d$
$\alpha_{k=1...K}$	$\alpha_{k=1...K} \in R^+$	weight of topic k in a document
α	vector of K	collection of all α_k
$\beta_{w=1...V}$	$\beta_{w=1...V} \in R^+$	weight of word w in a topic
β	vector of V	collection of all β_w
$\psi_{k,w}$	$\psi_{k,w} \in (0, 1)$	probability of word w in topic k
ψ_k	vector of V	distribution of words in topic k
$\theta_{d,k}$	$\theta_{d,k} \in (0, 1)$	probability of topic k in document d
θ_d	vector of K	distribution of topics in document d
$z_{d,w}$	integer between 1 and K	identity of topic of word w in document d
Z	vector of N	identity of topic of all words in all documents
$w_{d,w}$	integer between 1 and V	identity of word w in document d
W	vector of N	identity of all words in all documents

Table 3.4: Definitions of Variables in LDA

2. In a document, the accompanying presence of specific neighboring terms belonging to only one topic can disambiguate the terms that could not be classified by themselves.

3. Most documents will contain only a relatively small quantity of topics. Also, Topics have a probability distribution over the given documents.

4. Terms of a topic have their probability of distribution.

A formal definition of LDA needs to be set up with the table 3.4.

Then mathematically, the random variables can be described as

$$\psi_k \sim \text{Dirichlet}_V(\beta) \tag{3.1}$$

$$\theta_d \sim \text{Dirichlet}_K(\alpha) \tag{3.2}$$

$$z_{d,w} \sim \text{Categorical}_K(\theta_d) \tag{3.3}$$

$$w_{d,w} \sim \text{Categorical}_V(\psi_{z_{d,w}}) \tag{3.4}$$

To apply topic modeling in R, we will employ the "topicmodels" package in R.[GH24] Here, since there are three books in the novel series Remembrance of Earth's Past, an LDA model with 3 topics is attempted and figure 3.15 plots the highest frequencies.

The blue chart explains the topic in the first book. The most occurrences in the topic, struggling, crude, betrayed, swinging, civilization, finished, faded, indicate that the topic of this book may be folded around the survival of human civilization from compatriots' betrayal. The term amplification is a concept in the theory that Ye Wenjie relied on to send her broadcast, which replenishes the main idea of the topic. The green chart explains the topic in the second book with the most occurrences. One may deduce that the topic of the second book is about swirls of situations Earth is in. The terms unwilling and scared adds information to the understanding of the topic and render the negative sentiments. The red chart explains the topic in the third book. The top terms, stake, operator, invented, mechanical, witnesses, assault, may lead the audience to a topic of witnessing the development of technology and the following fights and wrecks. In fact, Cheng Xin is the character who has witnessed the short peace between Earth and Trisolaris and the attack of dual-vector foil that collapses three dimensions into two and who survived from the black domain and lived until the end.

Figure 3.16 is a confusion matrix of the topic, indicating where LDA assigns the words from each book. It can be seen from the figure that the topics are well differed from each other in each book.

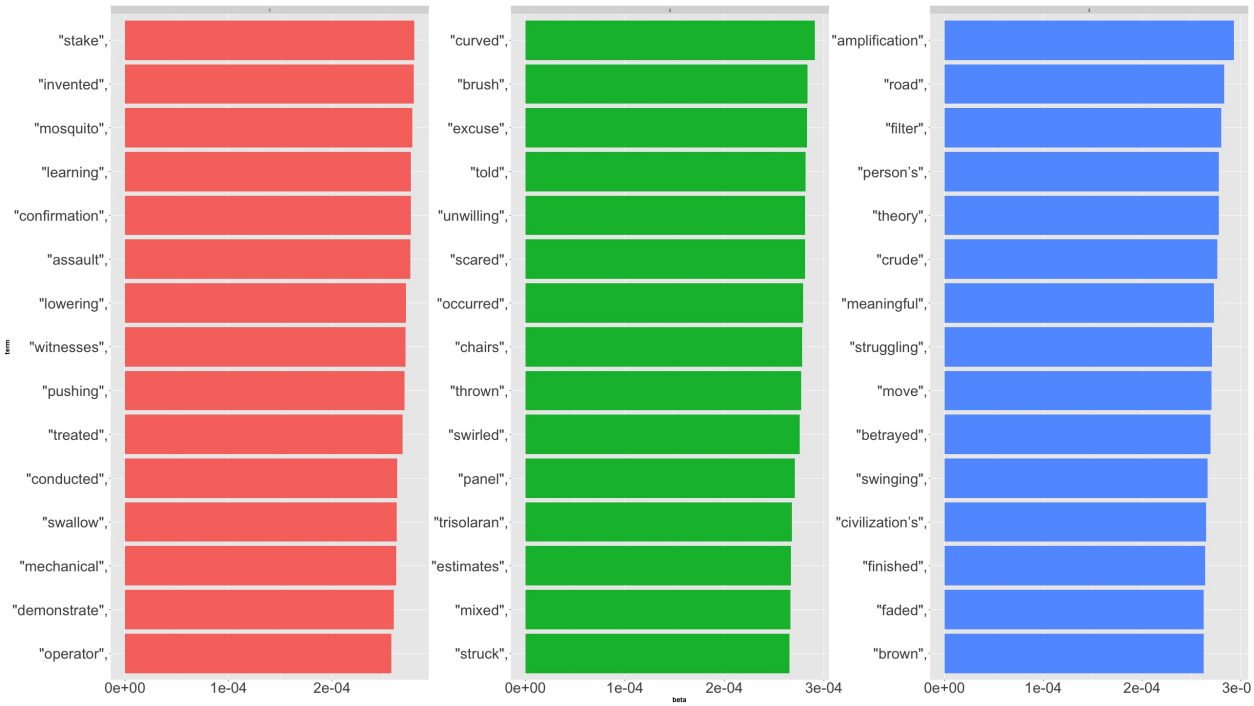


Figure 3.15: LDA with 3 Models

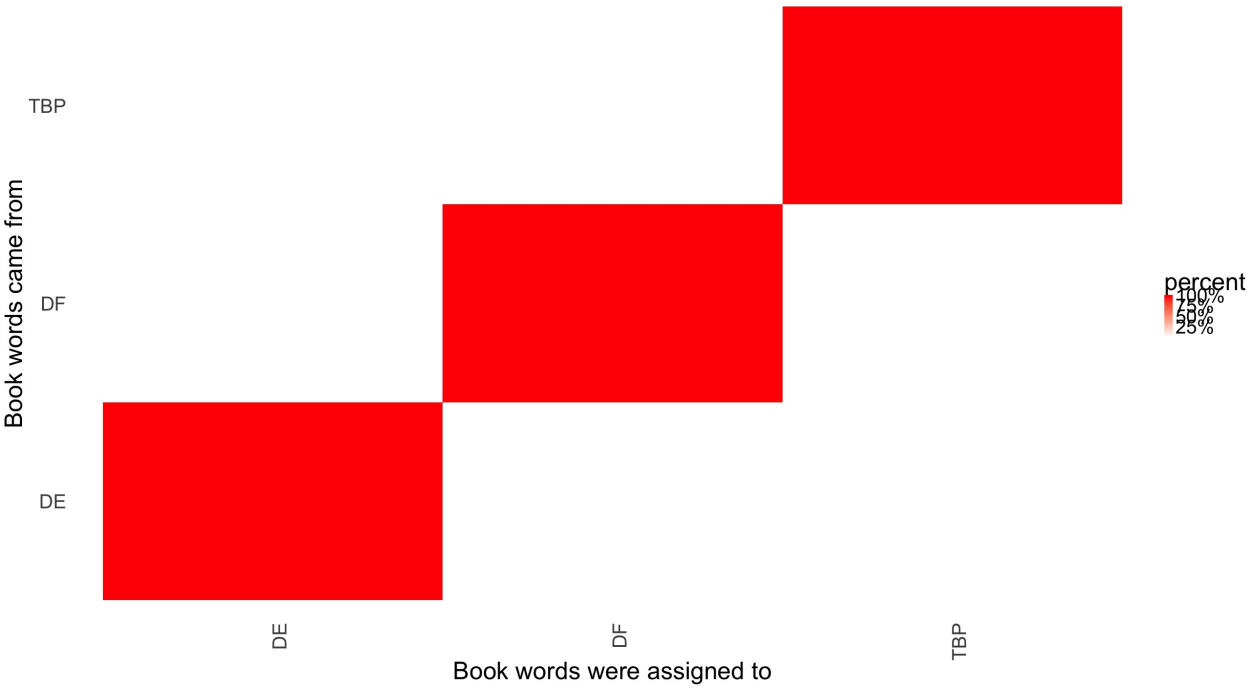


Figure 3.16: Confusion Matrix of LDA

CHAPTER 4

Conclusion and Discussion

In this study, the text content of the novel series Remembrance of Earth's Past by Liu Cixin is cleaned at the beginning, with simple transfirmation, conversion to lower case, punctuation removal, specific transformation, removal of English stop words and self-defined stop words, stripping white spaces, and stemming, for a more accurate text analysis using the text mining techniques. The data cleaning deprives respectively 44.33%, 46.94%, 44.50% of the content in the consequent books.

By exploring the most frequent words in the cleaned texts and establish connected nodes with respect to high correlation, the first book presents one story line that tightly connects every main character, similarly with the exploration of the second book. The third book shows multiple story lines, which might result from a large span in time in the context of this novel. The main characters are all precisely pointed out in the most frequency exploration of each book.

The term frequency-inverse document frequency (tf-idf) scores is computed by the product of two statistics term frequency and inverse document frequency and it rates the importance of a word in a document. Through the research, one can find the lists of the most important terms contain the names of the main characters, which emphasizes their irreplaceable roles in their dominating book. Additionally, one of the most important terms for the first book it is interrogator, signifying a crucial interrogation; for the second book it is wallfacer, leaving us a hint of the important project of human named Wallfacer and the corresponding project of trisolarans named Wallbreaker; for the third book the words are princess and lightspeed, with princess being the key word for the stories that Yun Tianming told Cheng Xin to inspire the construction of lightspeed star ships on Earth. The texts

are plotted to verify that they follow the empirical Zipf's law. Bi-grams show consistent knowledge with the previous research on single words and enlarge the understanding of the text content.

Sentiment analysis mainly utilizes three most common lexicons: Bing, AFINN, and NRC. Within the scope of each book, Bing lexicons recognize more negative sentiments than positive sentiments, while the proportion of positiveness in the third book increases by about 1.7% from the proportion in the first book. In contrast, AFINN lexicons discovers a slight decrease of positiveness from the first to the third book; however, the decrease may be considered insignificant. Despite the larger proportion of negative sentiments in every book, more contribution of positive sentiments are found in the most counted occurrences when the mannually rated scores are taken into consideration. The NRC lexicons are used on the scope of the whole series and it shows a similar determination of positiveness versus negativeness: 46% VS 54%. What's more, it analyzes the text content with respect to the 8 emotions in its lexicons and the results replenish the understanding with the colors of emotions.

An LDA model with 3 topics is fitted to the series to find the three topics, since it contains three books. As a consequence, the highest occurrences in the first topic are "amplification," "theory," "struggling," "betrayed," "swinging," "civilization," "finished," "faded," implying a topic of uncertainty and struggling. The highest occurrences in the second topic are "curved," "excuse," "unwilling," "scared," "swirled," "trisolaran," "estimates," "struck," suggesting an unclear situation with the extraterrestrial civilization Trisolaris. The highest occurrences in the third topic are "stake," "invented," "assault," "witnesses," "conducted," "mechanical," "operator," indicating a lucky witness of the conflicts between the human race and the hyper-advanced civilizations.

In conclusion, the analysis can point to the right direction of understanding the main themes of the content.

A further analysis could be conducted in finding the slope of the regression line. A research in larger n-grams could also help with the comprehension of the novels. Also, topic

modeling can employ different models for the topic detection and it can be applied within the scope of each book to discover more details of the story.

REFERENCES

[Cix15] Liu Cixin. *The Dark Forest*. Tom Doherty Associates, LLC, 2015.

[Fel18] Ian Fellows. *Word Clouds*, 2018. R package version 2.6.

[FH24] Ingo Feinerer and Kurt Hornik. *tm: Text Mining Package*, 2024. R package version 0.7-13.

[GH24] Bettina Grün and Kurt Hornik. *topicmodels: Topic Models*, 2024. R package version 0.2-16.

[Liu12] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[Moh22] Saif M. Mohammad. "NRC Word-Emotion Association Lexicon.", 2022.

[MT10] Saif Mohammad and Peter Turney. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon." In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 26–34, Los Angeles, CA, June 2010. Association for Computational Linguistics.

[Nie11] Finn Årup Nielsen. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.", 2011.

[Osm23] Volodymyr Osmak. *The Three-Body Problem: A Review of Liu Cixin's "Remembrance of Earth's Past" Trilogy*, 2023.

[SR16] Julia Silge and David Robinson. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, **1**(3), 2016.

[Wik24a] Wikipedia contributors. "Death's End — Wikipedia, The Free Encyclopedia.", 2024.

[Wik24b] Wikipedia contributors. "Liu Cixin — Wikipedia, The Free Encyclopedia.", 2024.

[Wik24c] Wikipedia contributors. "Sentiment analysis — Wikipedia, The Free Encyclopedia.", 2024.

[Wik24d] Wikipedia contributors. "Tag cloud — Wikipedia, The Free Encyclopedia.", 2024.

[Wik24e] Wikipedia contributors. "Zipf's law — Wikipedia, The Free Encyclopedia.", 2024.