NORTHERN ARIZONA UNIVERSITY
Department of Mathematics and Statistics

Sentiment and Word Frequency Analysis of The Chronicles of Amber

Crystal Sawtelle, Dr. Robert Buscaglia

Math and Statistics Department, Northern Arizona University, Flagstaff, AZ 86011

NAU NORTHERN ARIZONA UNIVERSITY
Office of Undergraduate Research and Creative Activity

Abstract

The Chronicles of Amber is a series of books written by Roger Zelazny between 1970 and 1991. The first five books in the series follow Corwin's story and the last five follow his son Merlin. The book series will be downloaded as a PDF document and converted into a text file. The text file will be converted into a data frame with two columns, 'text' and 'book'. Linguistic analysis of the book series will include using sentiment analysis, Term Frequency – Inverse Document Frequency (TF-IDF) analysis, word correlation analysis and topic modeling. Sentiment analysis is a process of categorizing opinions expressed in a text to determine whether the text is mostly positive or negative. TF-IDF analysis attempts to find words that are important to a text, but not necessarily common throughout the text. This allows us to find words that are characteristic of one book, or more common for one book than another in the series. Word correlation analysis allows us to find the correlation between specific words found in the same section of text. Finally, the last process will be Topic Modeling, a method for unsupervised classification of text data. The benefits of this analysis are to find patterns in Roger Zelazny writing and to see how these patterns changed over the course of writing the series.

Introduction

Text mining is the process used to identify meaningful patterns and insight from unstructured text. Text mining is an automatic process that uses Natural Language Processing (NLP), a subset of artificial intelligence that combines computation linguistics with statistical, machine learning, and deep learning models, to gain these meaningful patterns and insight. By transforming unstructured text into a structure machines can understand, we can automate the process of categorizing text by sentiment, topic, and intent.

Using text mining techniques, R studio, and *The Chronicles of Amber,* published in 1999 which is comprised of 10 books written and published by Roger Zelazny between 1970 and 1991, I will attempt to identify and analyze the sentiment, topics, and intent derived from this series of books.

Method

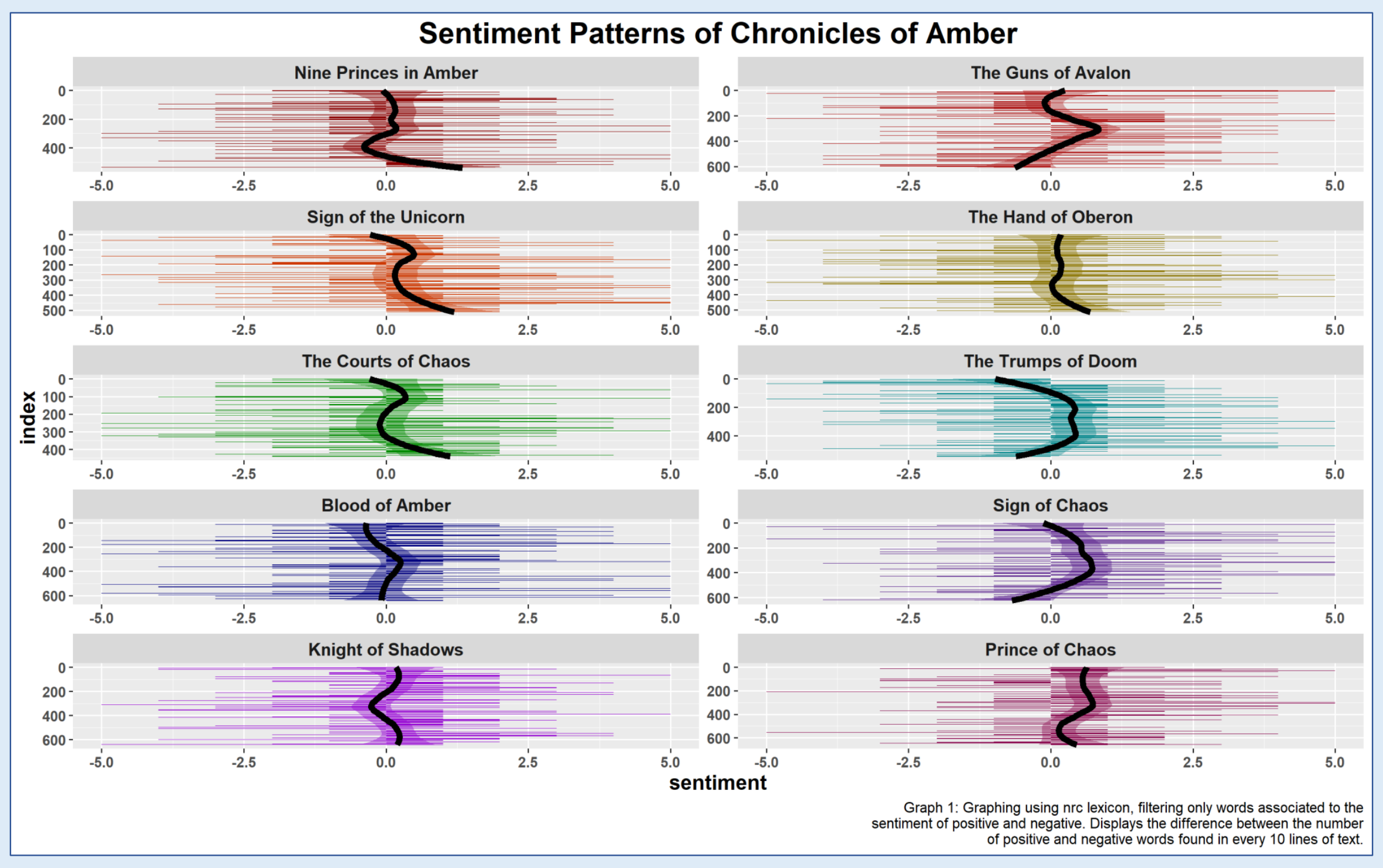
The PDF of the whole series was downloaded from oceanofpdf.com, a website consisting of free downloads of books, magazines, newspapers, etc. The PDF was then loaded into R and converted to text by using an R package called ***pdftools***. This package includes a function called *pdf_text()* which extracts the words from the document and returns them as a text document. Each book in the series was then broken out from the larger text, the punctuation was then removed, each line was left justified, and blank lines were removed. Additionally, at the beginning and the end of each chapter there is a reference to oceanofpdf.com which had to be removed. Each book is then saved as a *.rda* file and combined back into one overall function called *chronicles_of_amber()* that puts the data into a data frame with two columns, "text" and "book." The books are then further broken down to include a column for line numbers and chapters.

The books are then tokenized, the process of tokenization breaks up the text into units called tokens, or in this case the individual words in the book and puts them into a data frame. Stop words are then removed. Stop words are words that give little or no meaning to the sentence, words like "I", "the", and "and." There are many different lists of stop words, for this analysis we used the base *stop_words* provided by the R base package and the ***tidytext*** package "*snowball*" list.

Result – Sentiment Analysis

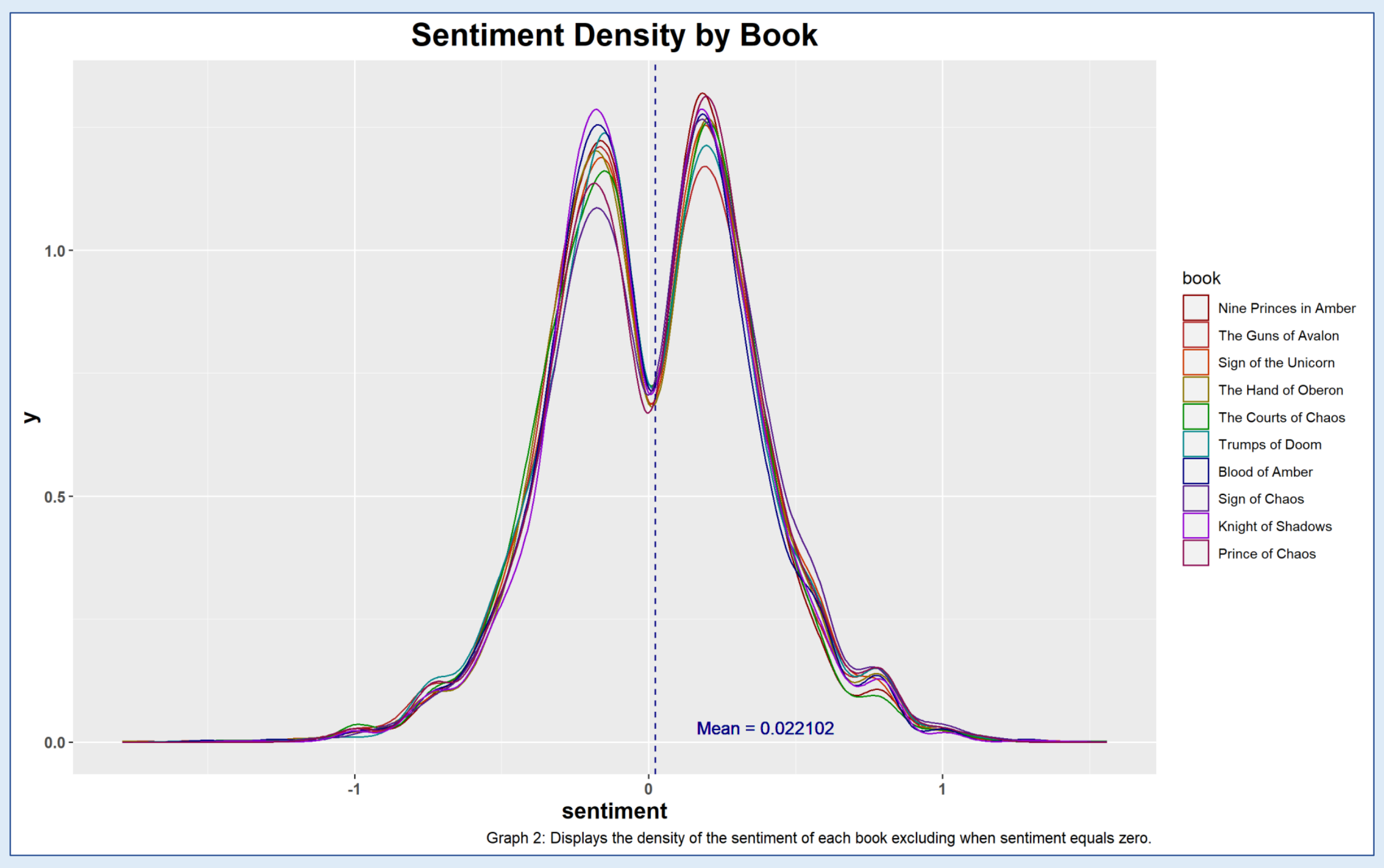
Sentiment analysis is a process of categorizing opinions expressed in a text to determine whether the text is mostly positive or negative. Typically, sentiment analysis is used for things like reviews, be they movies, restaurants, stores, etc., to measure how customers feel overall about their product. In this case we can measure how positive or negative Roger Zelazny's writing is and we can see patterns of sentiment to his books.

Calculating the sentiment of each book using the National Research Council Canada (NRC) lexicon, the NRC lexicon is a list of English words that assigns words to the sentiments of negative, positive, and to the emotions of fear, anger, trust, sadness, disgust, anticipation, surprise, and joy. After filtering only words associated to the sentiment of positive and negative, the difference was then taken between the number of positive and negative words found in every 10 lines of text. Similar patterns of sentiment can be seen between *Sign of the Unicorn* and *The Courts of Chaos,* starting negative increasing to positive, then decreasing towards negative and finishing the book positive. *The Trumps of Doom* and *Sign of Chaos* also have a similar pattern going from negative to positive ending negative.



Graph 1: Graphing using nrc lexicon, filtering only words associated to the sentiment of positive and negative. Displays the difference between the number of positive and negative words found in every 10 lines of text.

To find the density of the sentiment in each book, the code goes through each line of text and identifies whether a word has a positive or negative sentiment and assigns it a score, then calculates the total score for each line. The graph displays the density of each book's sentiment, excluding words that have a score equal to zero. Words with a zero sentiment cannot be categorized into a positive or negative sentiment. In fact, the seventh book in the series, *Sign of Chaos* has approximately 6.667 times more zero sentiment words than any other book in the series. The density shows that the sentiment of each line has a range of between approximately -1.8 and 1.6 and a mean of approximately 0.022102. This shows the books tend to be slightly more positive than negative.



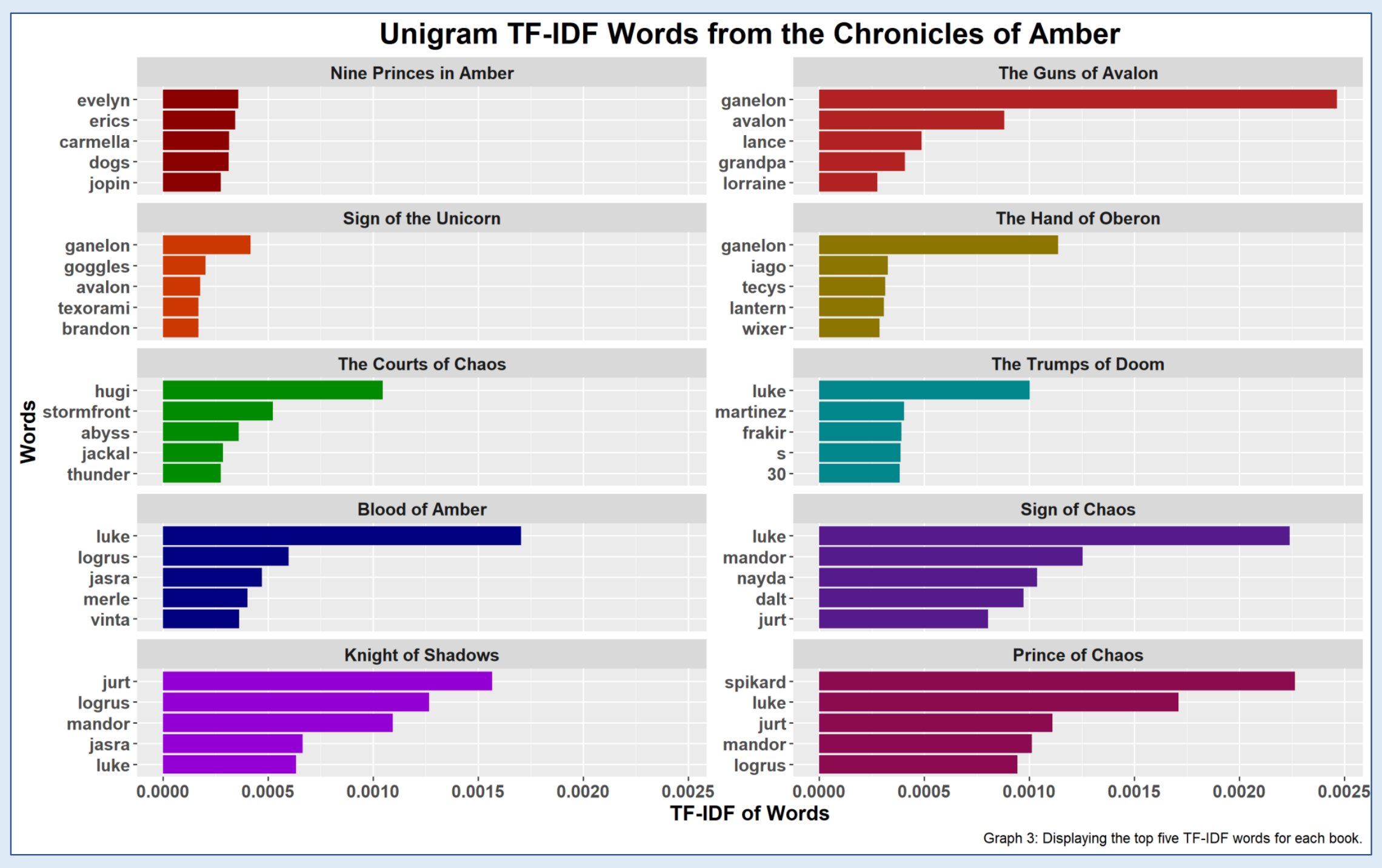
Graph 2: Displays the density of the sentiment of each book excluding when sentiment equals zero.

Result – TF-IDF

Term frequency is the number of times a specific word is found in the text. The inverse document frequency is how commonly a word is used. Calculated by IDF = log(N/DF$_t$)) where N is the number of documents (in this case 10 books) and DF$_t$ is the number of documents the term appears in. If a term appears in all the documents the IDF would equal zero, i.e., log(10/10) = 0. The Term Frequency - Inverse Document Frequency (TF-IDF) is the term frequency (TF) multiplied by the inverse document frequency (IDF). The TF-IDF attempts to find the words that are important/common in the text, but not too common. When IDF and TF-IDF are zero, these are extremely common words and thus are not as important. This approach decreases the weight for those common words and allows us to find words that are characteristic for one book within all the books, i.e., words that are more commonly found in one book than another.

Unigram TF-IDF

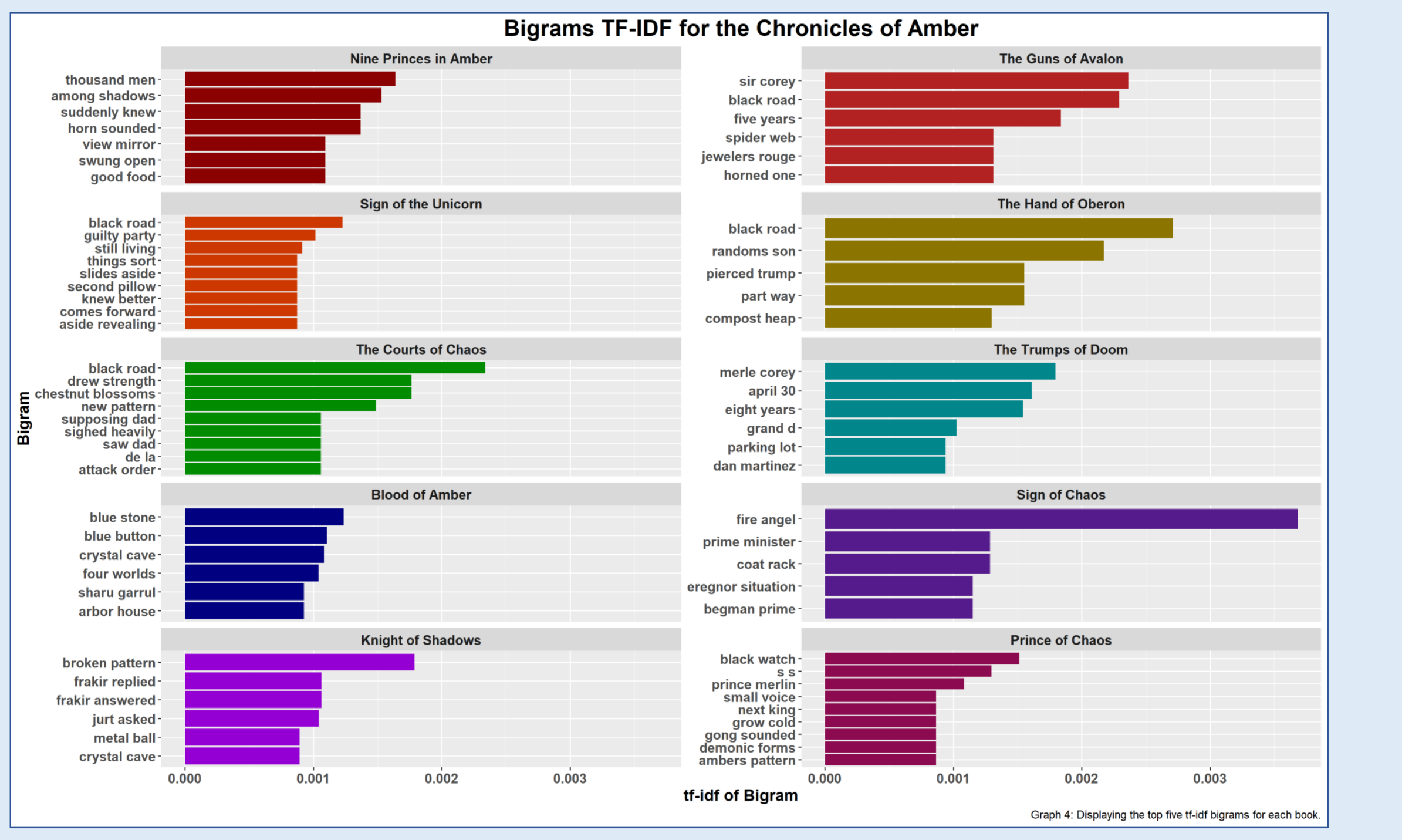
Unigram TF-IDF are single words identified as "important" within the text. For example, in the *Sign of the Unicorn* one of the important words was identified as "goggles." This was from a retelling of a rescue attempt that was important to the series. Another example is "spikard" in the *Prince of Chaos*, which was a ring meant to control the main character of the book. Even "Ganelon" in the *Guns of Avalon, Sign of the Unicorn,* and *The Hand of Oberon*, this character appears to be a side character throughout those three books, until it is revealed at the end of *The Hand of Oberon* who he really is.



Graph 3: Displaying the top five TF-IDF words for each book.

Bigram TF-IDF

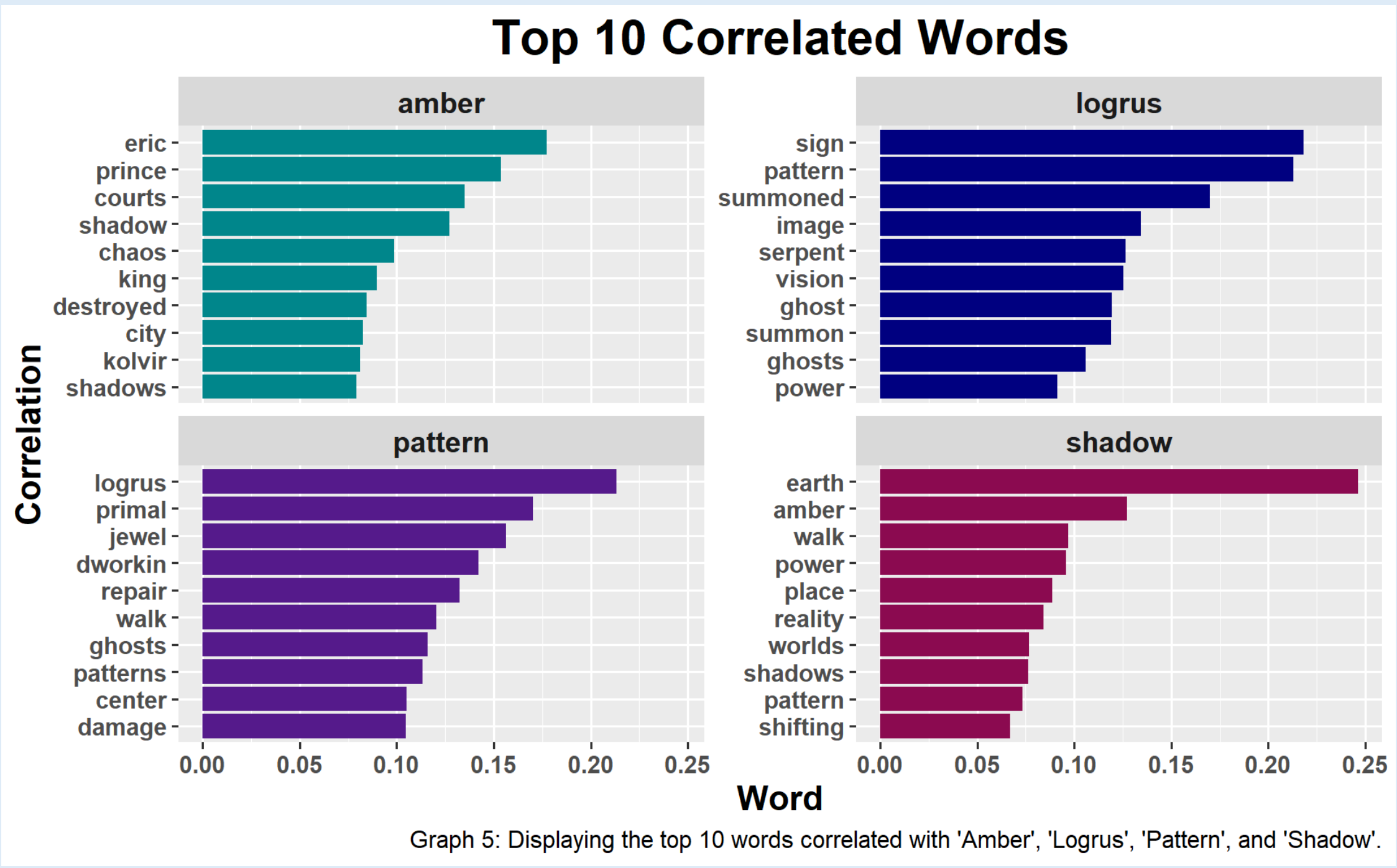
N-gram analysis connects a string of two or more words together to identify words that are commonly next to each other. Comparable to the Unigram TF-IDF, Bigram TF-IDF attempts to identify two consecutive "important" words. A good example comes from *The Trumps of Doom* where the bigram is April 30. April 30th was the date that someone tried to kill the main character for eight years straight. The date had more significance that stemmed from the first five books.



Graph 4: Displaying the top five tf-idf bigrams for each book

Result – Word Correlation

The *pairwise_cor()* function from the ***widyr*** package finds the phi coefficients for measuring between words based on how often they appear in the same section. Using this function, we can identify how often specific words appear together and how often they appear separately. The phi coefficient calculates the binary correlation based on how often a word appears in the same 10-line section. Using the words "Amber", "Logrus", "Pattern", and "Shadow" we can see the top 10 words that are correlated to each. I chose these words specifically because they play a large role in the series.



Graph 5: Displaying the top 10 words correlated with 'Amber', 'Logrus', 'Pattern', and 'Shadow'.

References

Silge, J., & Robinson, D. (2017). *Text mining with R : a tidy approach*. O'reilly.

RS, A. (2020, April 16). *Sentiment Analysis in R — Good vs Not Good — handling Negations*. Medium. https://towardsdatascience.com/sentiment-analysis-in-r-good-vs-not-good-handling-negations-2404ec9ff2ae

Acknowledgements

Northern Arizona University, Department of Math and Statistics

Northern Arizona University, Office of Undergraduate Research and Creative Activity

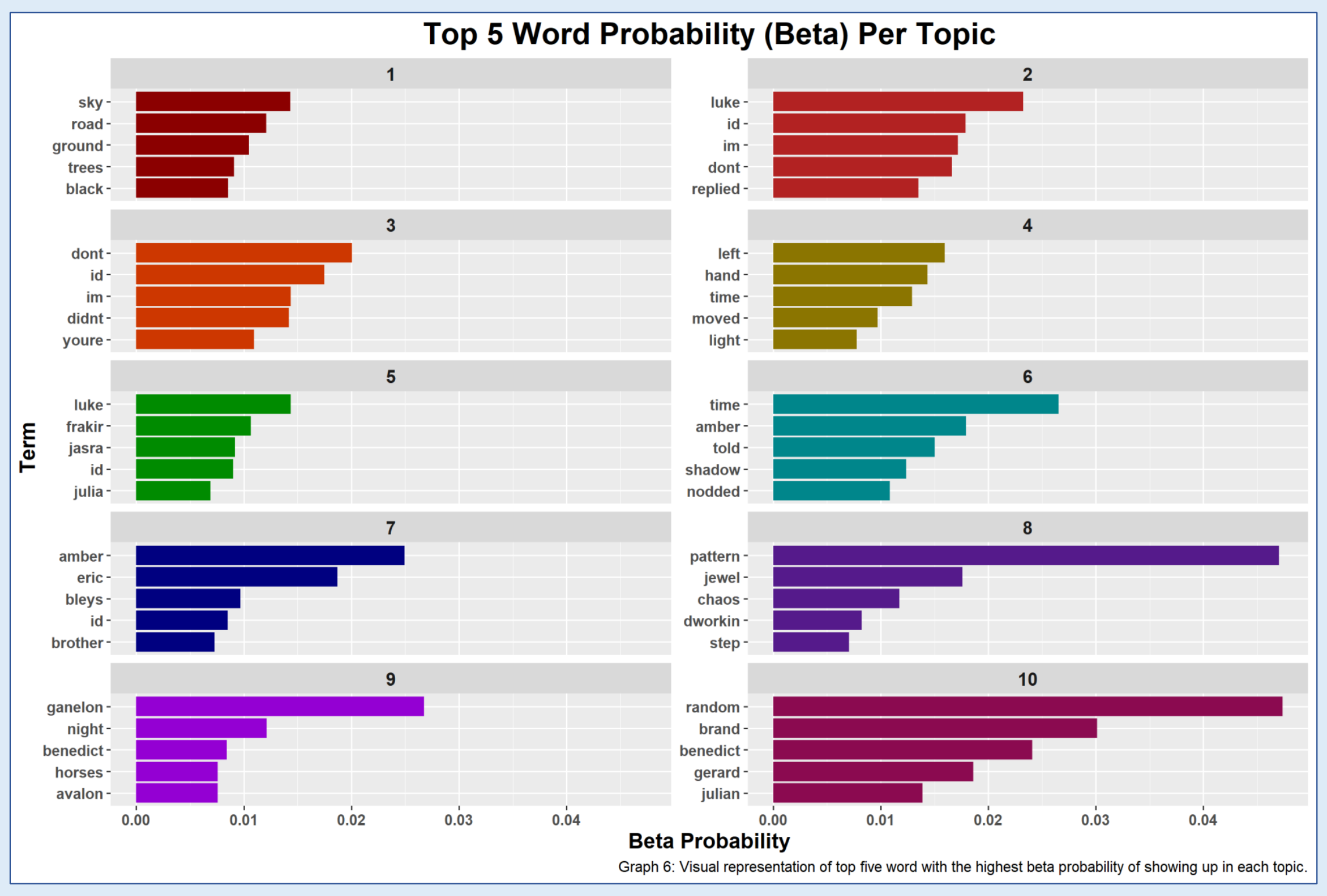
Result – Topic Modeling

Topic modeling is a method for unsupervised classification of a collection of documents, not unlike clustering numeric data, that attempts to learn to differentiate the documents based on the text content. Since this is an unsupervised classification, the topics are randomly chosen by identifying words that appear to have the same content.

The first thing that was done to make this data usable was to break the series down into several documents, in this case each document is the book name and the chapter number, i.e., Nine Princes of Amber_1. The documents are tokenized, and the tokens (words) are then cast into a document term matrix (DTM), the rows of this matrix are the document, each column represents one word, and the value is the number of times that word shows up in that document.

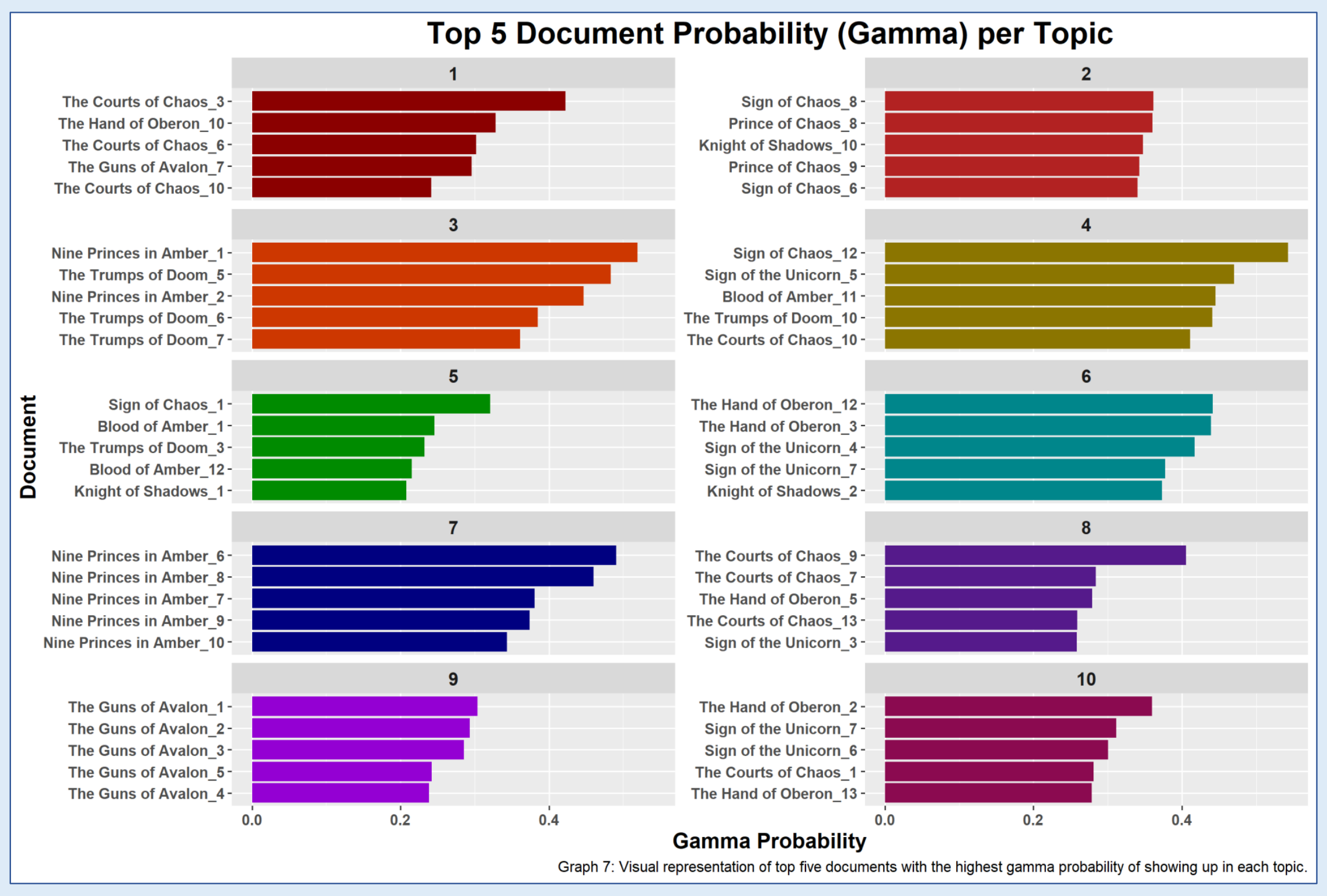
Using Latent Dirichlet Allocation (LDA) and incorporating Gibbs sampling, the DTM was broken down into 10 topics. LDA treats each document as a mixture of topics and each topic as a mixture of words. Gibbs sampling is a method of Markov Chain Monte Carlo that approximates difficult joint distributions by consecutively sampling from conditional distributions. The number 10 was chosen because there are 10 books in the series and the iteration of the Gibbs sampling was 500.

There are two probabilities generated from this model, the first is the beta probability or word-topic probability. The beta in the model is the calculated probability of a word being generated from a topic. For example, Ganelon has approximately 27% probability of appearing in topic 9.



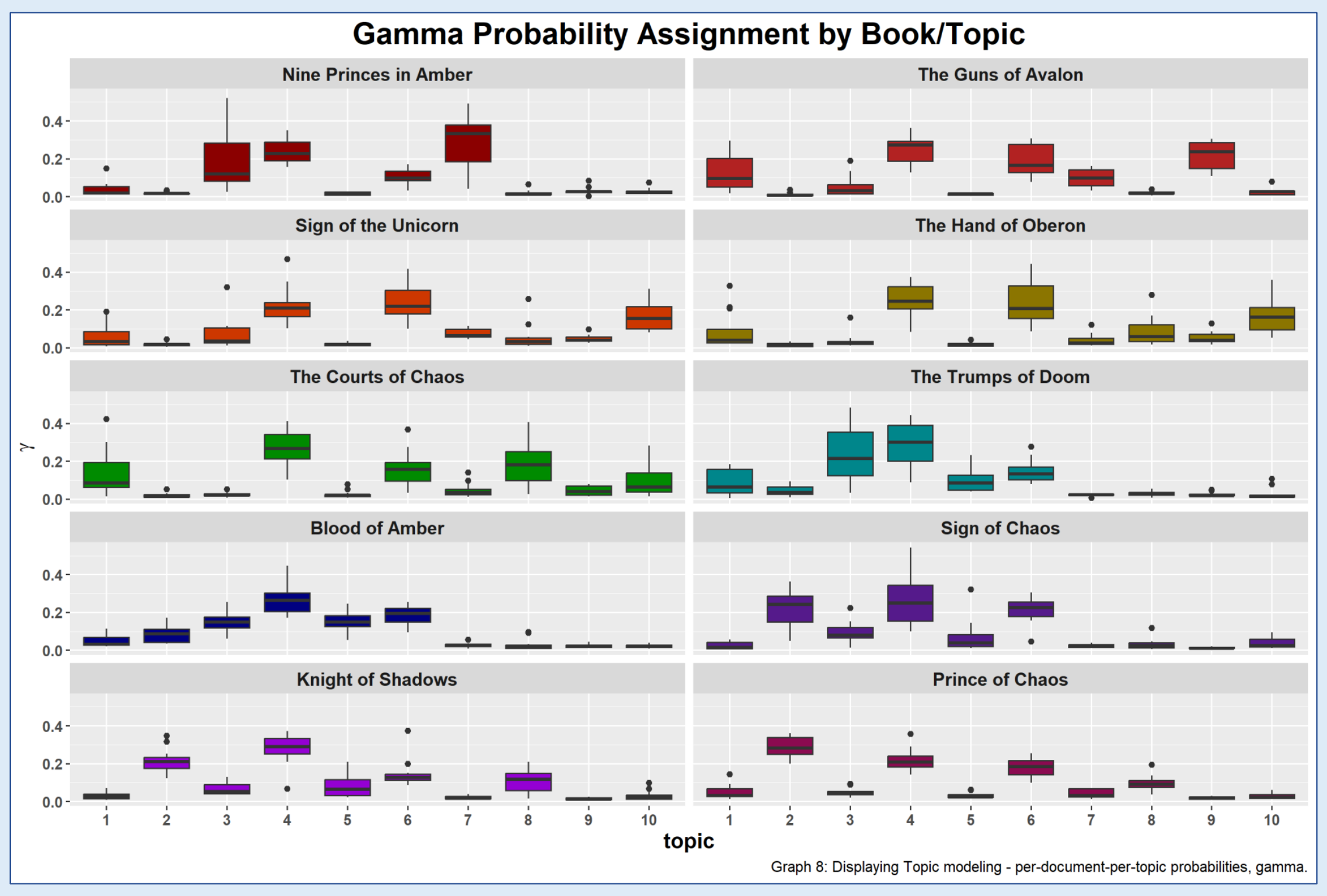
Graph 6: Visual representation of top five word with the highest beta probability of showing up in each topic.

The next is the gamma probability or document-topic probability. The gamma value is an estimated proportion of words from a document that are generated from a topic. For example, the model estimates that approximately 30% of the words in The Guns of Avalon chapter 1 are generated from topic 9.



Graph 7: Visual representation of top five documents with the highest gamma probability of showing up in each topic.

There are 115 chapters in *The Chronicles of Amber* series, excluding one prolog chapter in *The Trumps of Doom*. The boxplot graph displays the gamma probability for each chapter within each book. In other words, this shows the probability of each chapter within the series being assigned to a specific book. If we look at Topic 7, we can see that the words assigned to topic 7 have the highest mean probability of being assigned to *Nine Princes in Amber.*



Graph 8: Displaying Topic modeling - per-document-per-topic probabilities, gamma.

A perfect model would have assigned each book as a topic and would have been able to differentiate the words and documents between the books. Because this series employs similar locations, a great deal of the same characters, and primarily the same plot line, the topic model was not able to learn very well how to discern between the documents and classify them correctly. Topic Modeling works well when there are clear distinctions between the plots and characters in books, but not for this series.

Conclusion

Text mining and Natural Language Processing goes far beyond what we have seen here and is a fascinating subject. But from this analysis we have seen the sentiment of Roger Zelazny's books, what words were important and which words are correlated to each other. We can also see that a series of books, at least not this series, is not necessarily the best data to use machine learning techniques like topic modeling.

If I had more time, I would have liked to have scrapped reviews of the series from the internet and view how the sentiment of readers may have changed over time. Roger Zelazny died four years after publishing the last book in the series that I believe ultimately resulted in a few plot holes in the overall story line. There is another author, John Gregory Betancourt, a huge fan of Roger Zelazny's, who attempted to carry on with the world of Amber by writing a new series called *The Dawn of Amber*. It would be interesting to compare the writing techniques between the authors, writing patterns, and sentiment patterns.