Project 2 Proposal

Crystal Sawtelle

STA_486C

Title:

Sentiment and Word Frequency Analysis of The Chronicles of Amber

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 follows 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, word frequency analysis, N-grams 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. Word frequency analysis is used to identify the most common used words in a text document. This will be used to compare the most used words in each book in contrast to words across all the books combined. Additionally, word frequency will be compared between the first five books and the last five books. N-gram analysis connects a string of N words together and is combined with sentiment analysis and word frequency to better analyze multiple words and even sentences. 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:

The data collected is a PDF from the website OceanofPDF of *The Great Book of Amber* published in 1999 which is comprised of 10 books written and published by Roger Zelazny between 1970 and 1991. When I decided my project was on text data I wanted books that meant something to me. I read these books when I was a teenager at which time I fell in love with the fantasy of the stories. Additionally, my nephew was named after the main character of the first five books, Corwin.

Table 1: Displays the first line of text of each book, the book name, line number, and chapter.

text	book	linenumber	chapter
nine princes in amber	Nine Princes in Amber	1	0
the guns of avalon	The Guns of Avalon	1	0
sign of the unicorn	Sign of the Unicorn	1	0
the hand of oberon	The Hand of Oberon	1	0
the courts of chaos	The Courts of Chaos	1	0
the trumps of doom	The Trumps of Doom	1	0
blood of amber	Blood of Amber	1	0

text	book	linenumber	chapter
sign of chaos	Sign of Chaos	1	0
knight of shadows	Knight of Shadows	1	0
prince of chaos	Prince of Chaos	1	0

Methods:

In order to extract the text data from a PDF the package “pdftools” needs to be installed. This package includes a function called pdf_text(), this function extracts the words from the document and returns them as a text document. Surprisingly, after figuring out the right way to go about this, these steps are pretty easy. The harder steps are cleaning the data to make it usable for analysis. I want to separate out each individual book from the larger text and clean them which will allow me to view and analyze each book separately. I will be using read_lines() function from readr package to specify which lines to skip to reach the beginning of each book and how many lines to include. After that is completed I want to left justify all the text and make it lowercase. Using grepl() function I will remove all empty lines. At the beginning and end of each chapter there is a reference to OceanofPDF website that will need to be removed and the punctuation, such as quotes and apostrophes are not recognized by R, they will need to be replaced with either a recognizable punctuation or a blank. For example, the text has the apostrophe as “’”, which R only recognizes “’”. After all the cleaning is complete these will be saved as a .rda file and then combined back into one overall function chronicles_of_amber(). The next step in putting the data into a usable format is to add the line number and identify which line belongs to which chapter in which book as displayed in Table 1.

Amazingly, there is a lot of analysis that can be performed on text data. The first I would like to attempt is a 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 I want to see how positive or negative Roger Zelazny’s writing is. Additionally, I would like to see if there is a pattern to his books. For example, if he starts out positive and there are notable dips and/or uptick in sentiment throughout the series or vice versa. The next analysis I would like to perform is a word frequency analysis. Figure 1 displays the top 10 words found in the Chronicles of Amber. However, this is not very informative so I would specifically like to focus on term frequency - inverse document frequency (tf-idf). This 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 for one book, or more common for one book than another in the series.

N-gram analysis connects a string of two or more words together to identify words that are commonly next to each other. N-gram analysis can also be used in conjunction with count and correlation of words within sections of a text. Using the pairwise_cor() function from the widyr package we can identify how often specific words appear together and relative to how often they appear separately. I will use the phi coefficient to focus on how more likely both word A and B appear together, or where neither appear together, or they appear without each other.

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}}$$

The pairwise_cor() function finds the phi coefficients between words based on how often they appear in the same section.

n_{11} = Has word X and word Y, n_{10} = Has word X but not word Y, n_{01} = Not word X but has word Y, n_{00} = Not word X or word Y, $n_{1.}$ = Row total of Has word X, $n_{0.}$ = Row total of no word X, $n_{.1}$ = Column total has word Y, $n_{.0}$ = Column total no word Y, n = total

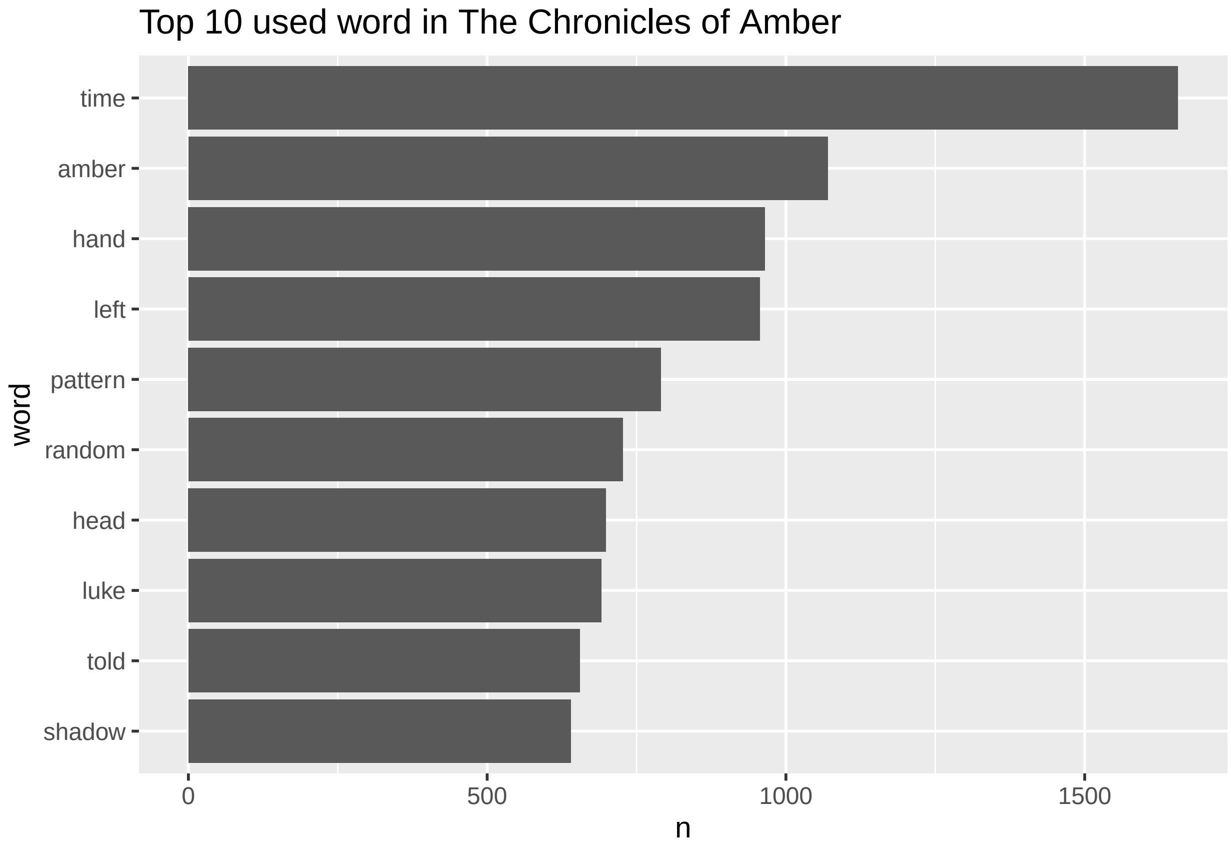

Figure 1: *Top 10 used words in The Chronicles of Amber*

The last process I would like to attempt is Topic Modeling. Topic Modeling is a method for unsupervised classification of documents, this is similar to clustering of numeric data, which finds natural groups/topics of items with in text. I will be using the Latent Dirichlet Allocation (LDA) approach based on a Bayesian framework which identifies the themes or hidden topics structures in text documents though conditional probabilities. I would like to see how well the model can learn to tell the difference between the 10 books based on the text content. I would also like to see how well the model learns if I separate the books into the Corwin series and the Merlin series.

Expected Outcomes:

If all goes correctly, I expect to find significant patterns to Roger Zelazny's books to identify his specific writing techniques. Since text analysis and modeling is not taught in the Data Science curriculum at Northern Arizona University the journey and learning I receive from this analysis is especially important to me. This is doubly important for Topic Modeling as it is a machine learning technique for text data.

Appendix:

Github page: https://github.com/dutchess3030/Amber