Sentiment Analysis of the Harry Potter Book Series

Richard Raybon

Abstract

This project takes the raw text of the Harry Potter book series and performs various methods of analysis on it using R. Some methods include generating word clouds based on the most common words, calculating term frequency - inverse document frequency for each book to see the most important words per book, and performing sentiment analysis using the Bing, AFINN, and NRC lexicons.

1 Introduction

In this project I decided to perform sentiment analysis on the Harry Potter book series. I felt that this would be an interesting topic to choose based on the widespread popularity of the series in various forms of pop culture. I believe that this will be an interesting topic to those who are familiar with the series, so they are my target audience for this project.

2 Data

The data for this project comes from Dr. Rei Sanchez-Arias' Github. The data comes in the form of a csv file, with three columns: book, text, and chapter. The "book" column denotes which book the text comes from, the "text" column contains each chapter from the book series, and the "chapter" column denotes which chapter in its respective book the text comes from.

3 Methods

In this project I used various methods for analysis. The first thing I did after importing the data was to tokenize the data and remove the stop words. The first analysis I did was to create a basic word cloud (Figure 1) of the most common words in the series. The second analysis I did was to calculate the term frequency – inverse document frequency of the book series (Figure 2), with the "book" column being the documents I searched.

The rest of the project had to do with sentiment analysis. I began by performing sentiment analysis using the Bing lexicon. The sentiment score of each book was calculated by the total positive words subtracted by the total negative words. I also created a word cloud (Figure 3) that showed the most common positive and negative words and colored them accordingly.

I then used the AFINN lexicon to determine the sentiment of each book. The sentiment was calculated by simply summing together the score of each word, since AFINN scores each word numerically on a scale from -5 to 5, as opposed to Bing which categorically scores each word as either positive or negative. Given Bing scores and AFINN scores I was able to compare how each lexicon scored each book graphically (Figure 4). For example, Bing rated book 5 as being more negative than book 7 due to the ratio of negative words to positive words being higher, but AFINN rated book 7 as the most negative.

Lastly, I used the NRC lexicon to score each book based on several emotions, such as anger, anticipation, etc. I then made a graph (Figure 5) that portrayed how prevalent that emotion was by book for each emotion.

4 Results

This section contains all the figures that were referenced in the Methods section.

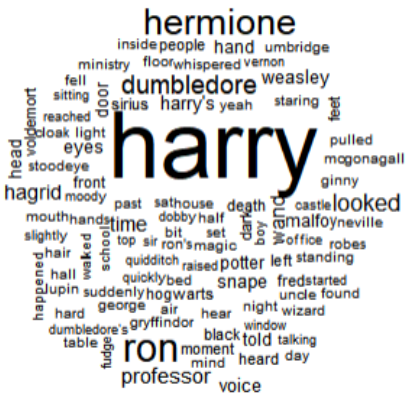

Figure 1: Basic Word Cloud

book	word	tf_idf
Half-Blood Prince	slughorn	2.450174e-03
Deathly Hallows	c	2.199099e-03
Order of the Phoenix	umbridge	1.624111e-03
Goblet of Fire	bagman	1.357994e-03
Chamber of Secrets	lockhart	1.290902e-03
Prisoner of Azkaban	lupin	1.179371e-03
Goblet of Fire	winky	9.466788e-04
Goblet of Fire	champions	8.518592e-04
Deathly Hallows	xenophilius	7.728621e-04
Half-Blood Prince	mclaggen	7.384470e-04

1-10 of 67,881 rows

Figure 2: Words with highest tf-idf

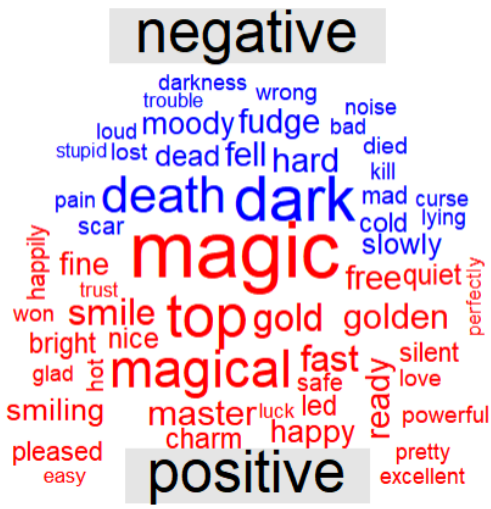

Figure 3: Bing word cloud

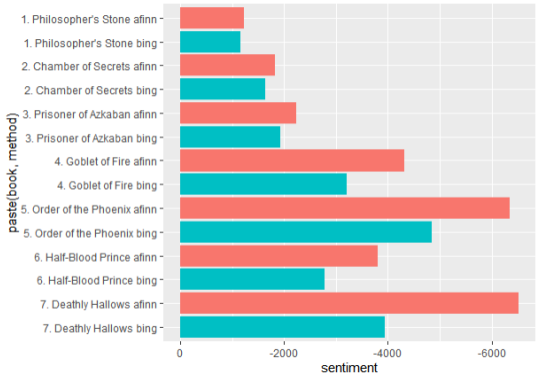

Figure 4: Bing v AFINN sentiment

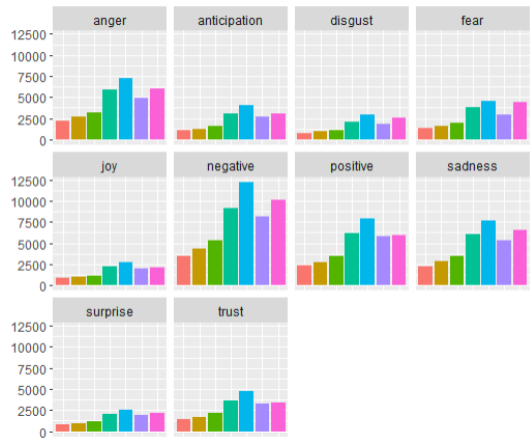

Figure 5: NRC visualization

References

https://github.com/reisanar/datasets/blob/master/hp.csv