SI 618 WN 2020 Final Project Report
Marilu Duque
4/20/2020

Harry Potter Series Analysis: Analyzing Key Characteristics, Sentiment and Popularity Amongst Houses.

Motivation

The Harry Potter 7 part book series has taken the world by storm since its release in the late 1990s. The 7 books are

1. Harry Potter and the Philosopher's Stone (1997)
2. Harry Potter and the Chamber of Secrets (1998)
3. Harry Potter and the Prisoner of Azkaban (1999)
4. Harry Potter and the Goblet of Fire (2000)
5. Harry Potter and the Order of the Phoenix (2003)
6. Harry Potter and the Half-Blood Prince (2005)
7. Harry Potter and the Deathly Hallows (2007)

Since then, there have been 8 movie adaptations, 1 official broadway show and millions of contributions posted online in the form of fan art. This series chronicles the life of Harry Potter and his battles in the wizarding world while being a student at the Hogwarts School of Witchcraft and Wizardry.

This generationally transcendent series has been a favorite of mine ever since I was little. I chose to focus my final project on this because I wanted to know more about the data behind the magic. This project aims to explore the differences in books, sentiment, and dynamics between my favorite characters and the Hogwarts houses. Through analyzing the books, I intend to answer the following questions:

1. How do the characteristics (i.e. Word count, popular words, popular characters) of all 7 Harry Potter books compared to each other?

2. How have the Hogwarts houses (Gryffindor, Slytherin, Ravenclaw and Hufflepuff) popularities changed throughout the 7 books?

3. How can we classify the sentiment of the Harry Potter books?

4. How do Harry, Dubmbledore and Malfoy compare on a good vs evil scale?

SI 618 WN 2020 Final Project Report
Marilu Duque
4/20/2020

Data Source

URL: https://github.com/formcept/whiteboard/tree/master/nbviewer/notebooks/data/harrypotter

The data for this project is from a GitHub repository created by Formcept, a data analysis company in India. The repository possesses the entire series (7 books/files) of the Harry Potter Books in .txt format. This dataset was edited by one of their contributors, Prakhar Mishra in 2017. This data was collected straight from the harry potter books and sorted through the python glob tool. The type of data and variables are mostly all text, spaces and commas/periods. In total, there are about 6,554,138 words when the books are compiled together. For processing and stop word deletion purposes, there are ~62,000 words in total to train and test my NLP model.

Questions

1. How do the characteristics (i.e. Word count, popular adjectives & popular characters) of the 7 Harry Potter books compared to each other?

Method:

For my first question, I wanted to get an overall analysis of the books to include total word count, unique word count, most popular character and adjectives. I wanted to get a bare bones understanding of the word counts so after importing the libraries and .txt files, I just counted the length of each book. After that, I joined the total counts into a dataframe and proceeded to count the total unique words per book and placed that into a dataframe. To get an overall look at both totals, I merged the data and plotted them in a bar graph with Matplotlib. To find the top 10 adjectives, I used the Text Blob, an API for understanding natural language processing (NLP) that tags parts of speech such as nouns, and adjectives . I used this to find the 'type' of word and then dropped any types that were not adjectives. I did face some issues with noise data as some of the word types were not all typed correctly so to solve this I went through and dropped the irrelevant rows, analyzed by using .value_counts and .nlargest(), and then plotted my findings. I went through the same process to find the most popular characters by dropping rows whose type were not proper nouns.

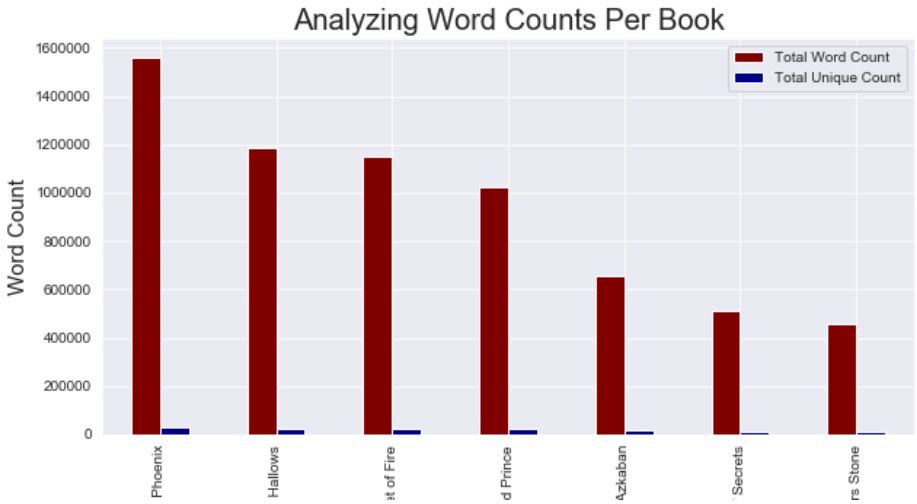
Analysis & Results:

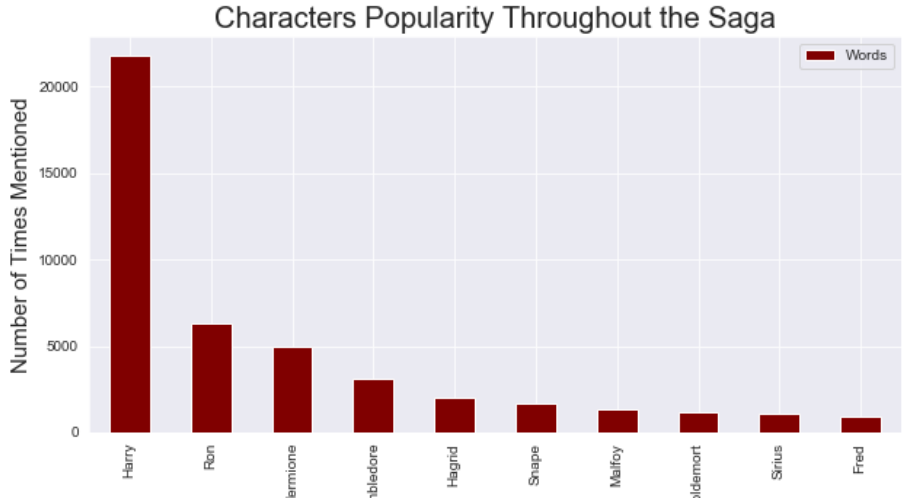
From the total and unique word count, we can see that the overall word count per book, greatly outweighs the unique word count. The total word count for the entire series is 6,554,139 while the total unique word count is 135,786. From this, we can assume that a vast majority of words are filler or stop words used to keep sentence cohesiveness. In merging both the total and unique word counts, we can analyze the average word counts across the series. The average Total word count is 936,305 while the total unique is 19,398. The longest book was the 5th book, Order of the Phoenix (2003), the shortest was the 1st book, Philosopher's Stone (1997). The first book being the shortest makes sense because J.K. Rowling was just getting the story of Harry Potter started and almost testing

the market. Since then, J.K. Rowling had been publishing a Harry Potter book every year since 1997 so it makes sense that Order of the Phoenix was the longest as she spent 3 years writing and publishing it.

Additionally, I calculated the top 10 most popular adjectives which were 'other', 'last', 'little', 'good', 'first', 'few', 'old', 'long', 'large', 'own'. I was surprised at how basic these adjectives were with the dark content in the book. I expected words like 'insane', 'terrible','black' or 'dead'.

As for the most popular characters throughout the series, obviously, Harry Potter would be the most popular. Continuing the next 2 were his best friends, Ron and Hermione. Then his mentors Dumbledore and Hagrid, then his nemesis Snape, Malfoy, Voldemort, and secondary characters Sirius and Fred.





SI 618 WN 2020 Final Project Report
Marilu Duque
4/20/2020

2. How have the Hogwarts houses (Gryffindor, Slytherin, Ravenclaw and Hufflepuff) popularities changed throughout the 7 books?

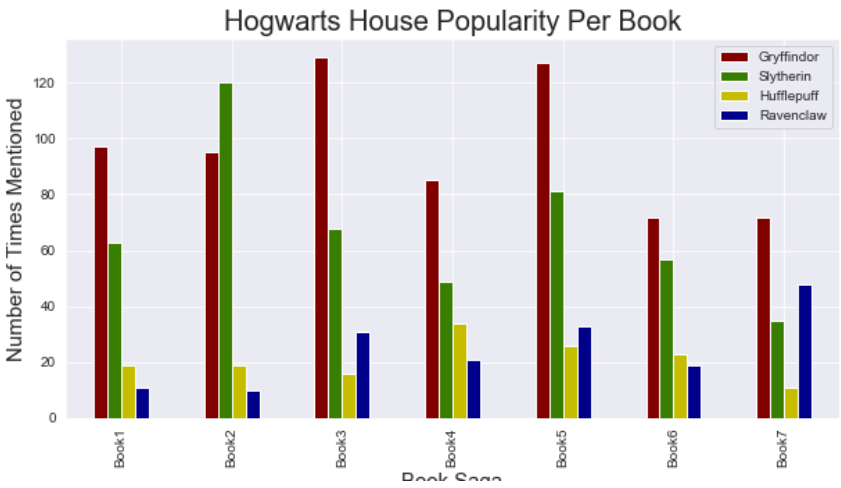
Method:

For my second question, I wanted to get an overall analysis of the popularity of each house per book. To do this, I parsed through every book separately and added the total count of mentions per house in a dataframe. I then merged the dataframes to better show the house mentions per book and then transposed the frame for better visibility and plotting ease. I did have a lot of trouble with getting and putting all the data together as it was from a large data set and took a while. I worked through this by doing each one separately rather than than all at once. I then found the total mentions for each house and plotted that total as well with Matplotlib. Lastly, included the correlation between the most popular house to the other 3.

Analysis & Results:

After getting the count for each Hogwarts house mentioned, we can see with the bar plot that across all but one book, Gryffindor is the most popular with 677 mentions. This is not surprising as the 3 main characters are in Gryffindor. Leading behind them is Slytherin with 473 mentions which is not surprising as well since the book's antagonist, Voldemort, has a history in the Slytherin house. It was surprising that Slytherin was the most popular house in the 2nd book, Chamber of Secrets, since the main characters were still leading the story in this one. I believe it has to do with the fact that the story took place in mainly Slytherin dominated parts of Hogwarts such as the actual Chamber of Secrets. The 3rd most popular is Hufflepuff with 148 mentions and the 4th Ravenclaw with 173 mentions. These last two rankings are not surprising as they did not have a giant storyline in the books.

Additionally, I found the correlation between the most popular house and the other three. Gryffindor and Slytherin had the strongest correlation with .986, then it was Gryffindor and Hufflepuff with a .983 correlation and lastly Gryffindor and Ravenclaw with a .968 correlation. The strongest correlation is expected as the houses are continued nemesis throughout the series.



SI 618 WN 2020 Final Project Report
Marilu Duque
4/20/2020

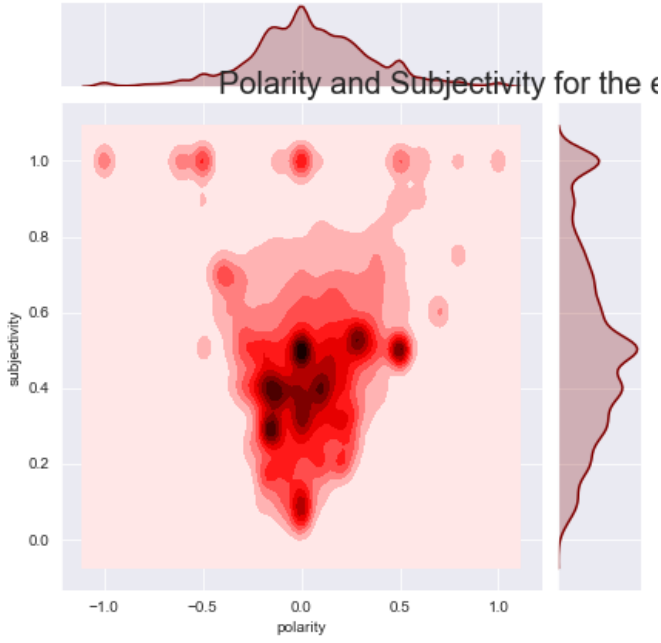
3. How can we classify the sentiment of the Harry Potter books?

Method:

For my third question, I wanted to conduct a sentiment analysis of the Harry Potter books as a whole. I had a lot of issues with doing sentiment analysis with this specific text. Due to the size of the data set, my program kept crashing and at one point all of the work I had done had been corrupted. In order for this not to happen again and to still be able to find sentiment in the book series, I found the TextBlob API that would help with that. I started by merging all the .txt files into a single .txt file and then ran the book through the TextBlob API to utilize their sentiment analysis text classification tool. I did not remove any text or characters in order to keep the true nature of the sentiment intact. In doing so, I was able to calculate the series sentiment polarity and subjectivity and plotted my findings using Seaborn.

Analysis & Results:

After running the entire series through TextBlob, we can get our sentiment analysis data. In doing so, it was able to analyze 63,914 sentences to achieve the current sentiment analysis. We can see that subjectivity, feelings expressed in the text, is .483 which means that it is slightly above neutral subjectivity due to its close nature to +1. The polarity, emotions in the text, is much more neutral at .044. The subjectivity is expected, but I thought the polarity might be a bit higher as the life and death nature of the books can stir up many emotions throughout its sentences. This should be understood though because the subjectivity and polarity are directly correlated. In looking at the joint plot, we can see the distribution in the sentiment analysis. Through the peaks in the graph, we can see that polarity is much more stable and neutral while subjectivity is more fluid.



SI 618 WN 2020 Final Project Report
Marilu Duque
4/20/2020

4. How do Harry, Dubmbledore and Malfoy compare on a good vs evil scale?

Method:

For my fourth question, I wanted to analyze my favorite characters, Harry Dumbeoldore, and Malfoy to see how they rate on a good vs. bad scale. To do this, I imported all the books separately using the sorted .glob, combined the books and stripped them off extra characters and digits. From there I tokenized by importing the Gensim tool and ran a for loop that deleted the stop words. To build and train my NLP model I used the Word2Vec model and trained it to the full corpus of the Harry Potter books. To test the model, I checked for words that I knew would be in one of the books and one that would not. With that, tested the semantic relationships between various characters such as Harry, Ron, Malfoy and Hermione, the model thus performed adequately. I also tested what vectors were related to various words such as 'wizard' and 'hagrid'. With this, I was able to run the model in comparison to the characters and reveal their scale data. I plotted this information with Seaborn scatter plots and bar graphs.

Analysis & Results:

In creating my NLP model, I trained it with ~62,000 words and had success when testing the accuracy. Using Word2Vec's get_unmatching_word function, I create a list of Slytherin students and one non-Slytherin. The model was able to accurately detect which was not the Slytherin student. Additionally, I used the most_similar function to test the vectors semantic relations for various characters such as Hagrid. I also tested the positive and negative correlations with James Potter, the word 'Father' and Snape. With this, the model detected the correlation between the two to be Lily Potter, 'mother', 'parents' and others. As an avid Harry Potter fan, I can see how the model was correct with those correlations. To assess Harry, Dumbledore, and Malfoy on a good vs. bad scale, I used the scipy cosine_similarity function to give each character a score. These scores and through the Seaborn scatter plot, we can see that Dumbledor is the 'goodest' of the three in which he is more on the good side than the bad side of the plot. Following second was Harry Potter falling in both the good and bad sides. Third came Malfoy falling completely on the bad side of the plot. I believe this output is accurate due to the nature of the character's roles. Dumbledor is always a mentor and protector, while Harry Potter is a good person but does mischievous and dangerous tasks. Malfoy is just a villainous character through most of the entire series.