**LING571 Final Paper**

**Corpus Analysis of the Harry Potter Series**

*By Sharanya Akkone and Shravani Hariprasad*

# Introduction:

In this project, we have tried to analyze a corpus of the ".txt" format from the wildly popular and adored Harry Potter series for patterns and interpretations. Using Python tools like NLTK and Pattern, we attempted to accomplish a textual analysis. for our final paper. As Potterheads, we have always been quite curious about and wanted to know the answers to questions like "which spells were the most used?". With this effort, we attempted to comprehend the corpus better through code rather than just continuing to live in the fictional world of Hogwarts.

To give a brief introduction about the series, Harry Potter was first featured as an orphan who was mistreated by his guardians, his aunt and uncle and their son, in the 1997 book Harry Potter and the Philosopher's Stone (also known as Harry Potter and the Sorcerer's Stone). Harry learns that his parents were witches and wizards on his eleventh birthday and that he has been invited to enroll in the Hogwarts School of Witchcraft and Wizardry as a wizard.

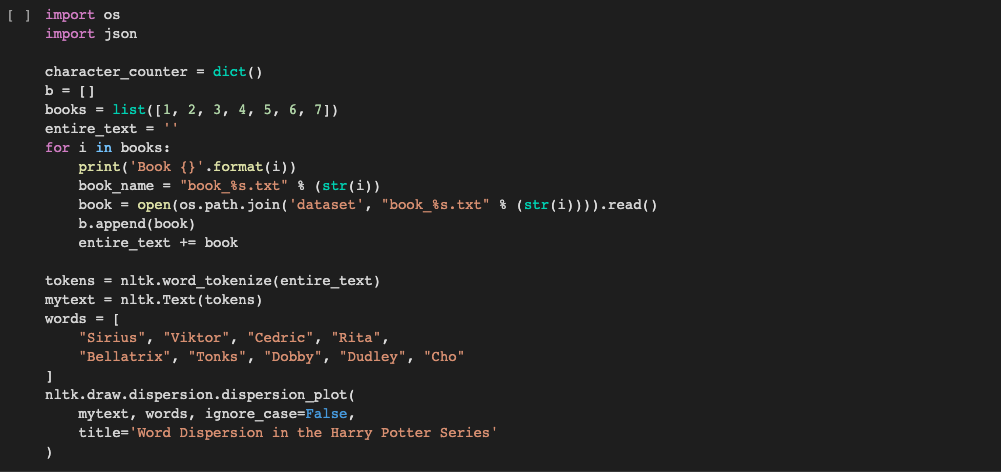
Worldwide, the Harry Potter series has proven to be very popular with both kids and adults. Each book was a bestseller, distributed in more than 200 nations and about 60 different languages.

# Distinctiveness of the Corpus

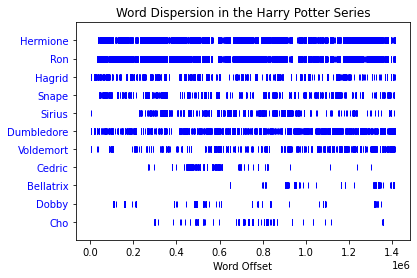
As previously established, the Harry Potter series centers on the fictional school of Hogwarts and its students who are learning witchcraft and wizardry, which inevitably involves spells, charms and potions that the author made up to fit the overall theme of the book. In academic writing, there is little to no personal pronoun use, no contractions, no slang terms and uses rigid print format, while in fictional writing, the author is given more liberty to express themselves in a more open setting where they are able to invent anything from names to locations to, in this context, even words!

Queries performed on Corpus:

### **Dispersion of words in each book**

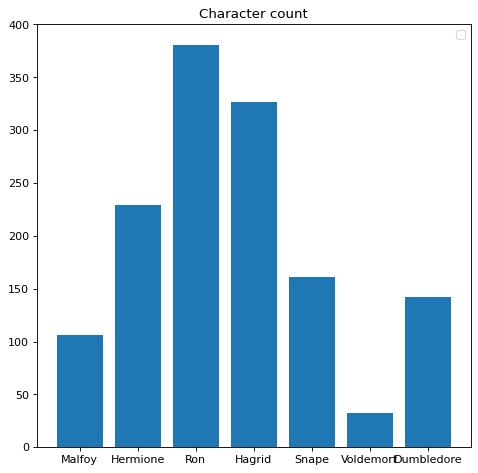
Code snippet:

Output:



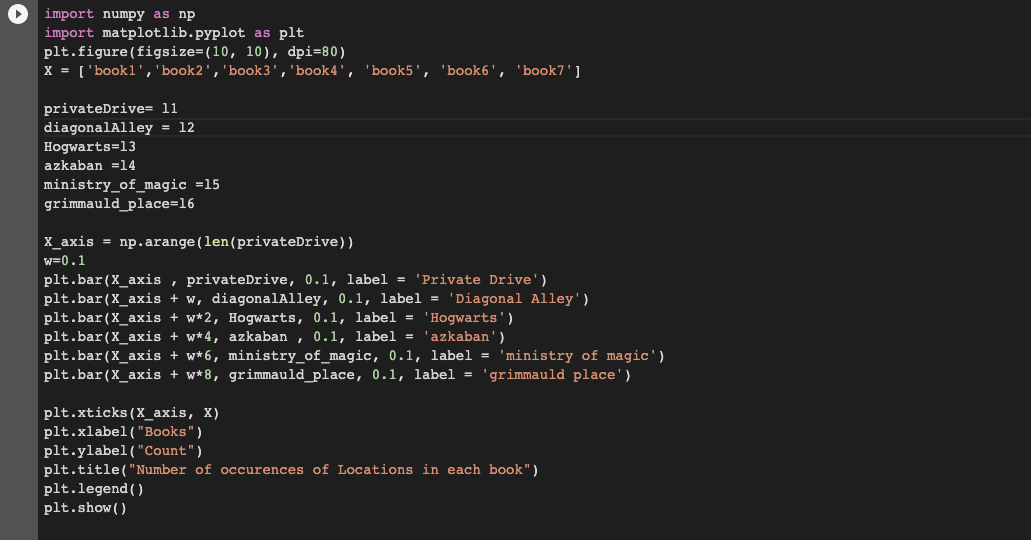
We attempted to illustrate the word dispersion of specific Harry Potter series words in this query. We provided the words, which are all the names of the characters in the series, in order to identify which character from the list will appear during the entire run of the series. We chose the following few well-known names instead of Harry because the book mostly revolves around him. Tokenization has been utilized to identify those particular words in the corpus by the code.

The below bar graph shows the character occurrences count in the first book. This was done by employing the FreqDist() function from the nltk module. Apart from Harry, the second most frequently occured character was found to be ‘Ron’, followed by ‘Hagrid’.

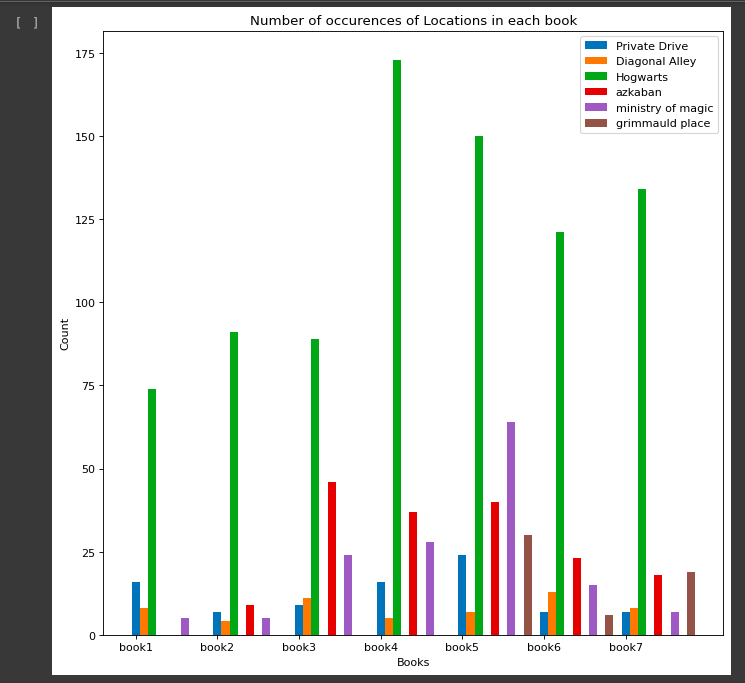


### **Number of times each location has appeared in each book**

Code snippet:



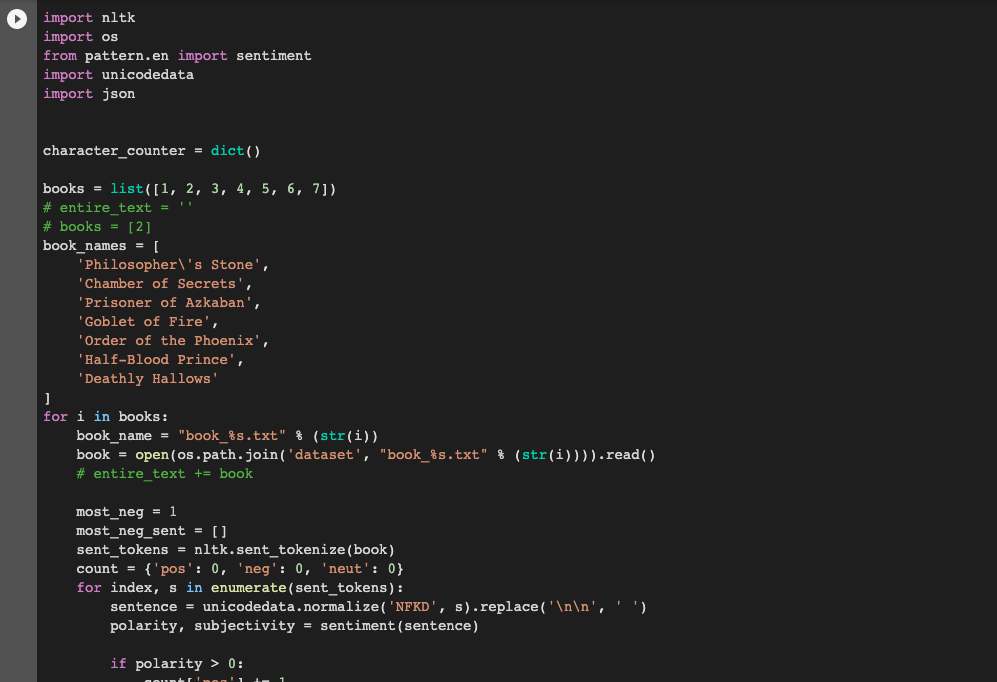
Output



In the above query,​we sought to comprehend how the locations in the series were distributed. Although Hogwarts is undoubtedly the setting that receives the most attention in the book, we can learn interesting details about some of the other places. For example, Azkaban gets its first mention in the second, and because the title of the third book is “Harry Potter and the prisoner of Azkaban”,it revolves around the location of Azkaban the most compared to the rest of the books, but still Hogwarts tops the list.

### **Sentiment Analysis on the Corpus**

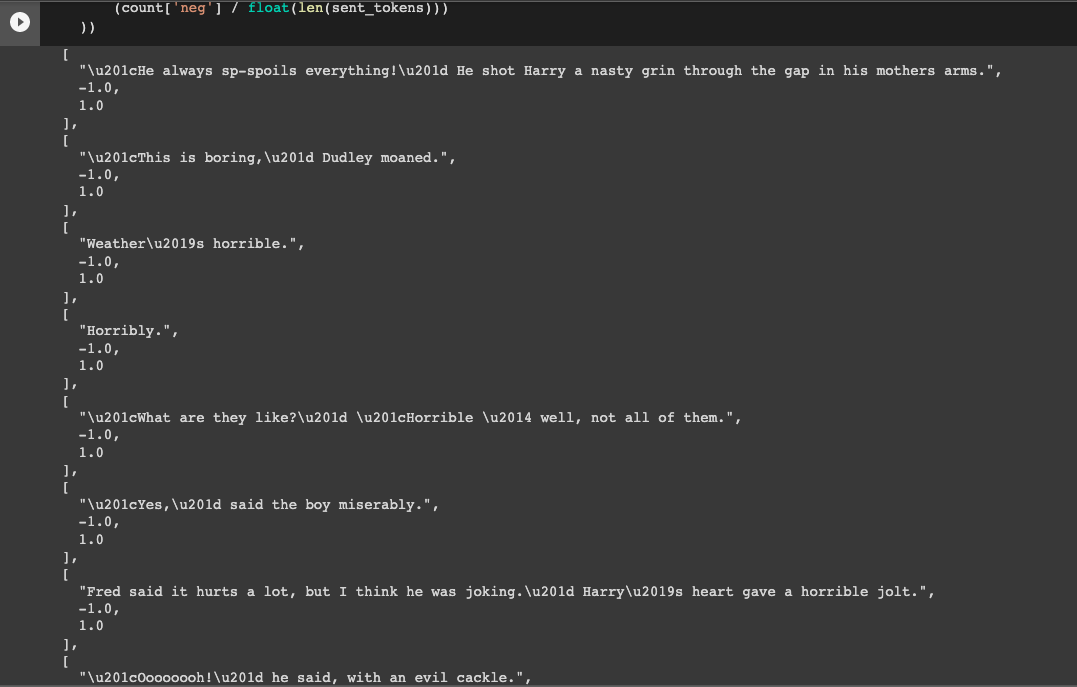
Code snippet part 1:

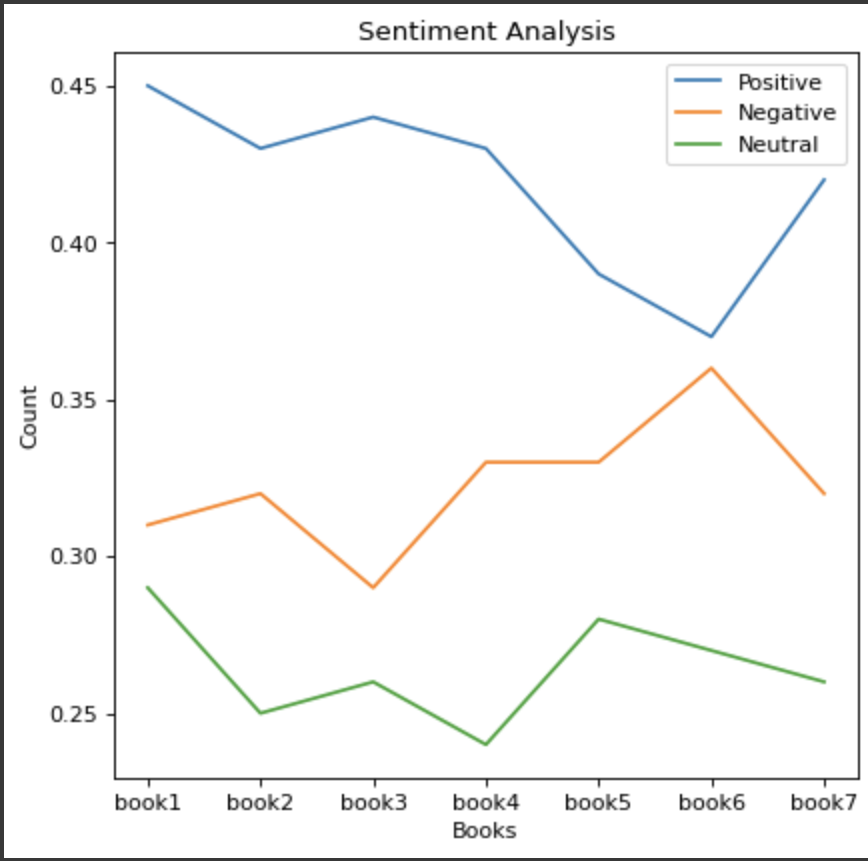


Code snippet part 2:



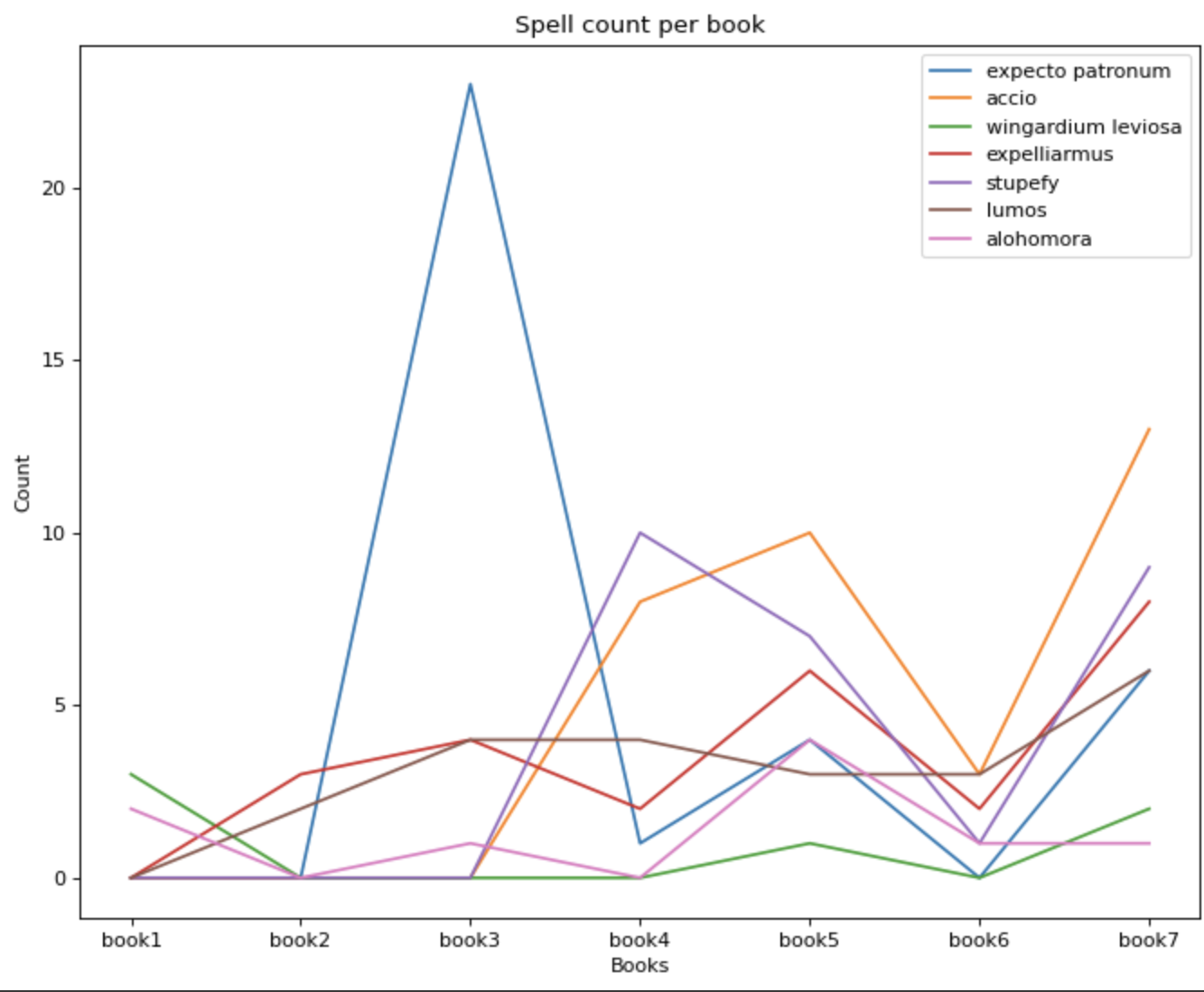
Output:



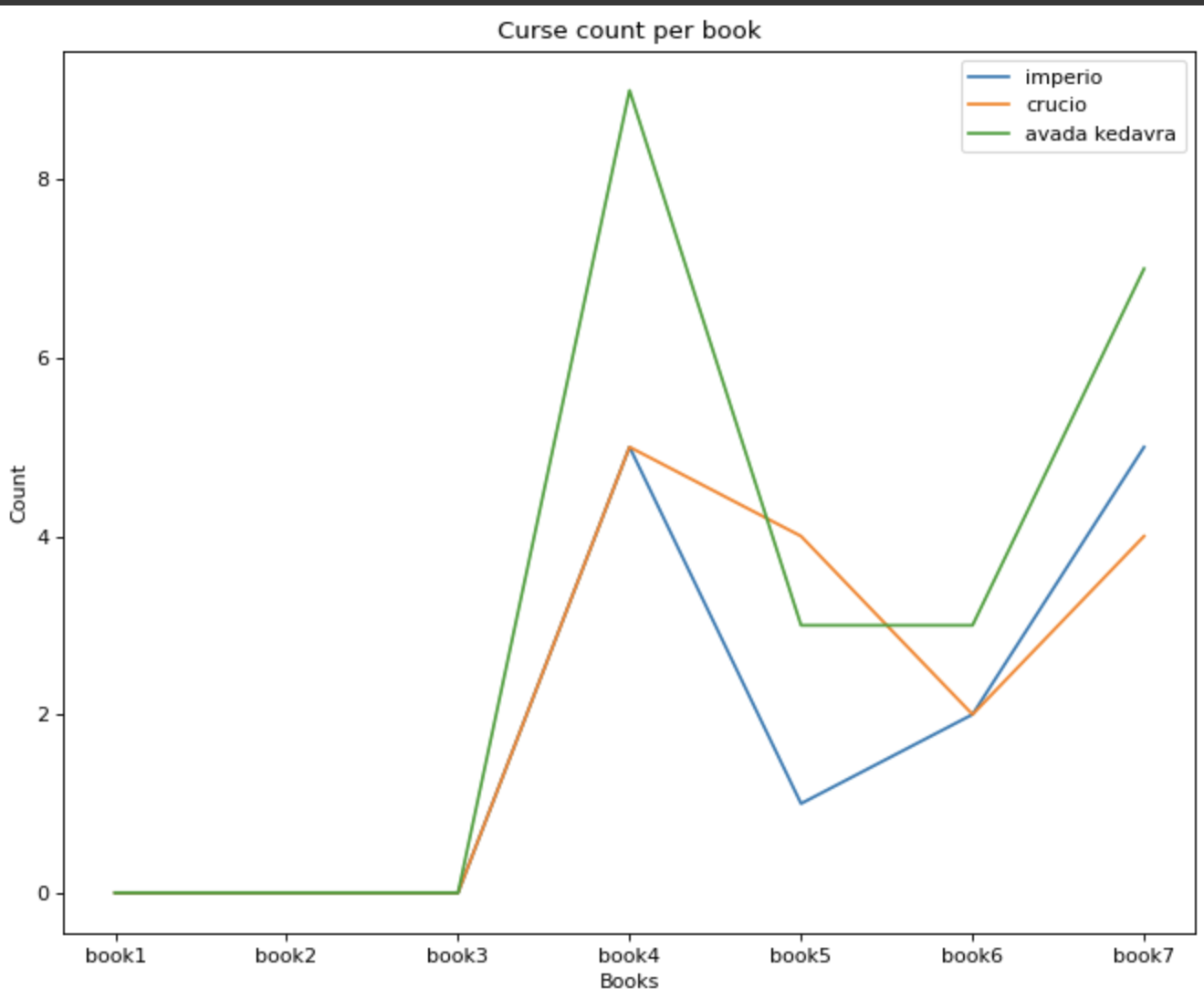


This analysis involves a sentiment analysis on the corpus, using the Python library called “Pattern”. There are two parameters that are being displayed here, which are Polarity and Subjectivity. Polarity basically indicates whether the sentiment is negative, positive or neutral, and the score ranges from -1(negative) to 1(positive). Whereas subjectivity tells us about the intensity of the emotion, or to put it in other words, whether the sentence is objective(0) or subjective(1). For example, in the sentence that involves only one word, “Horribly”, it tells us that it is a negative emotion, and also how strong the speaker’s emotion is because of the word used. Another interesting pattern we noticed is that, in book 6 the negative sentiment almost matches up with the positive, as compared to the rest of the series.

### **Analysis of number of spells and curses in all the 7 books**

Spells in the book:

Curses in the book:

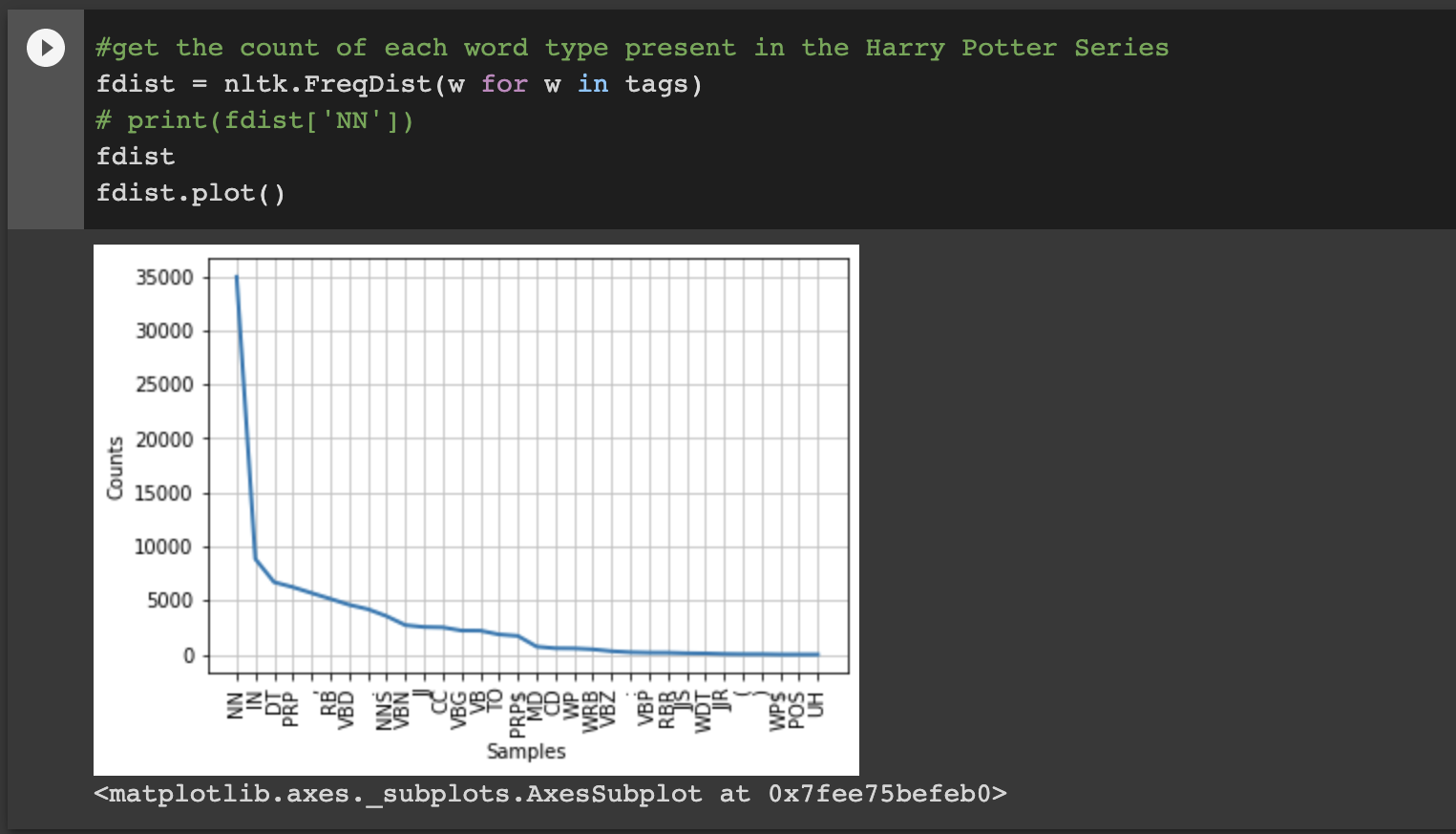


Any Harry Potter discussion is incomplete without mentioning the spells and curses used at various parts of the book. As of this query, we have plotted the various spells whose distribution we were interested in. We can see a sharp decline in the spells in the 6th book, as the author JK Rowling explores more dark themes, which explains the decline. Whereas for the curses, we were particularly interested in these three curses and analyzed their occurrences in the series.

### **Analysis of the most frequent bi-gram and tri-gram in the first book**

Our study revealed that “(‘of’, ‘the’)” was the most frequently occurring bi-gram in the series, with a count of 4901, while "(‘out’, ‘of’, ‘the’) " was the most frequently occurring tri-gram, with a count of 783.

### **Counting the word type (Parts-of-Speech) in the series**



In the analysis of each word type of book 1, we can conclude that the most frequent word type is NN, which stands for Noun followed by IN, DT and so on. Also when frequency distribution function was applied to the words in the first book, the characters ‘Harry’ and ‘Hagrid’ were one among the most common 50 words.

### Making word clouds for each book:





These word clouds are for the first and the third book in the series. As expected the word “Harry” has appeared the most in all the books, but the other words vary from book to book.

### **Conclusion**

By delving into one of our all-time favorite series, we have gained an exceptional understanding of linguistics through this project. We have examined several facets of the Harry Potter books and have discovered the solutions to some extremely intriguing questions that have persisted in our minds for years, such as the most often employed spells across the seven volumes. We've used some of the questions we learned how to ask in class to get answers to questions we never would have thought to ask. The series' analysis provided us with more material than only the mythical, fantastical world of Harry Potter, the boy who lived, and the other Hogwarts supporting cast members.

## Contributions:

Tasks 1,3,5 - Shravani Hariprasad

Tasks 2,4,6 - Sharanya Akkone

Task 7 and Report were equally divided between both of us.