Harry Potter 1-3. Explorative analysis of movies' transcription.

Olga Bystrova

Introduction and Data Description

In this project we explore the first three movies related to the Harry Potter novel: Harry Potter and the Philosopher of Stone, Harry Potter and the Chamber of Secrets, Harry Potter and the Prisoner of the Azkaban.

Hypothesis:

- Harry Potter should be one of the most common words / bigrams during three movies
- The proportion of positive words decreases from film to film

The main analysis is done with the use of the R programming language. It is shown to be a very effective language to deal with statistical research. As for the specific libraries, we use tidyverse, tm, nlp and wordcloud2. All these packages allow us not only to manipulate data but also to visualize extracted knowledge points.

Let us go to the dataset description. For every movie transcript there are two informative columns: 'Character' and 'Sentence'. The first column shows the name of a character that said the sentence and the second shows the sentence itself. For example:

Character	**Sentence**
Hagrid	You are a wizard, Harry

Table 1. Dataset Example

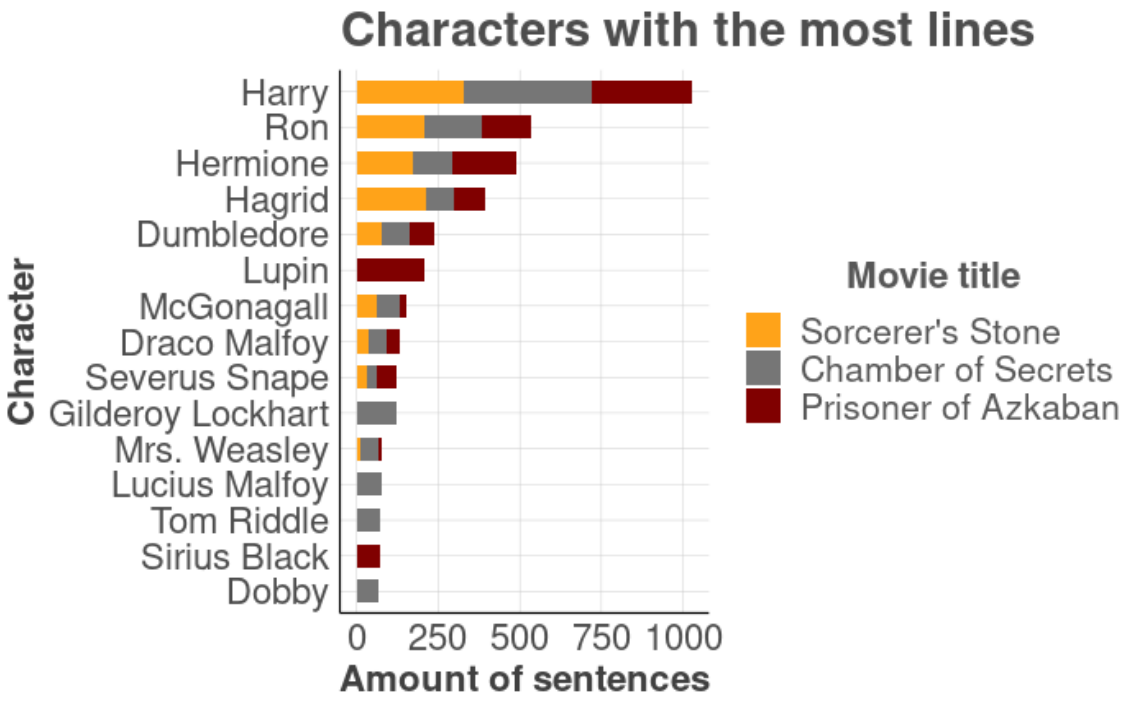
To analyse data through a concrete movie perspective, we add the third column to the dataset with the title of the film. Overall, we have 1587 sentences with 79 characters. After that we made some preprocessing steps to analyse data: we removed punctuation and stop-words. The final step would be to lowercase all text.

For sentiment evaluation there is an extra database with a dictionary for evaluating the opinion or emotion in text named "Bing". This database contains almost 7,000 popular words with assigned sentiment: positive or negative.

The full code is available at
https://github.com/ovbystrova/hse_practicals/tree/main/R/project

Main Analysis

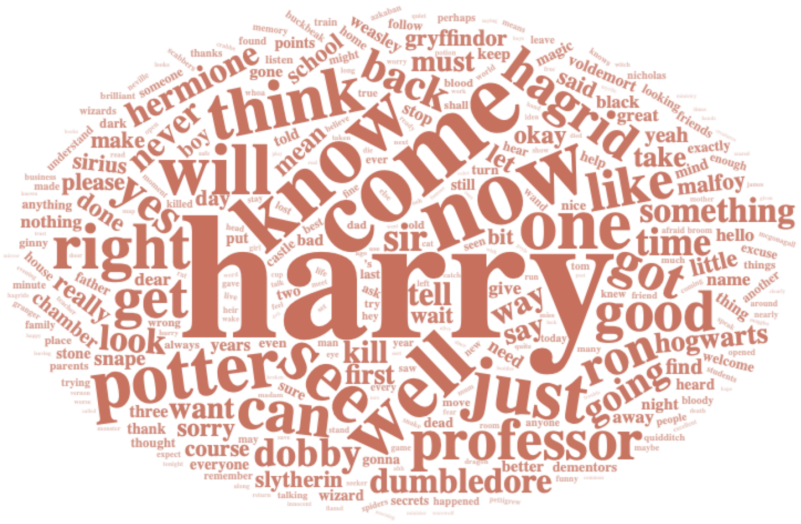
The first obvious thing to do is to check which character said more sentences than others. From *Picture 1* you can see that Harry Potter has around 1 000 lines during three movies which is about twice bigger than Ron Weasley, who took the second place. The first three places belong to the three main characters. It is quite surprising that Lucius Malfoy is in the top 15, because he appeared only in the second book and was not a main character at all.



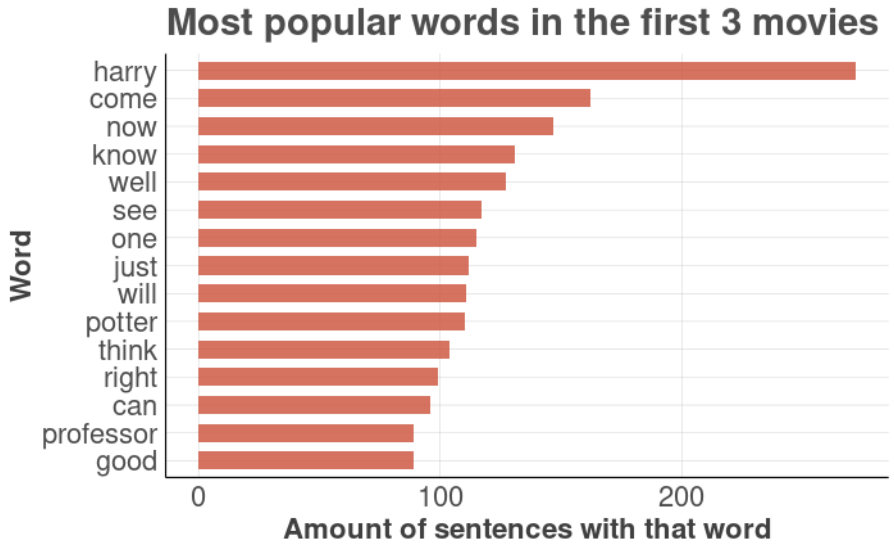
Picture 1. Top 15 characters with the most sentences from all three movies.

There are several characters that appear only in one of the three movies: Lupin, Gilderoy Lockhart, Top Riddle, Sirius Black. Another surprising fact is that Lupin, who appears only in the third movie, has more lines than Professor McGonagall who appears in all three parts of Harry Potter movies.

The next step would be to see what words, bigrams, and trigrams are the most popular in our data. To do so, a Term Document Matrix was made. And by using word cloud library we managed to see the most popular tokens. In both *Pictures 2-3* we can see that the name of the main character absolutely dominates with more that 250 appearings (It is quite interesting that for 455 minutes of movie footage it appears 272 times, which is one time per 1.4 minute). Harry's last name is also in the top 15, it is in 10th place. For wordcloud illustration we took only those words that appear more than 10 times. From these two illustrations we may also conclude that the language of the first three books is quite simple and prove that these books are for children.



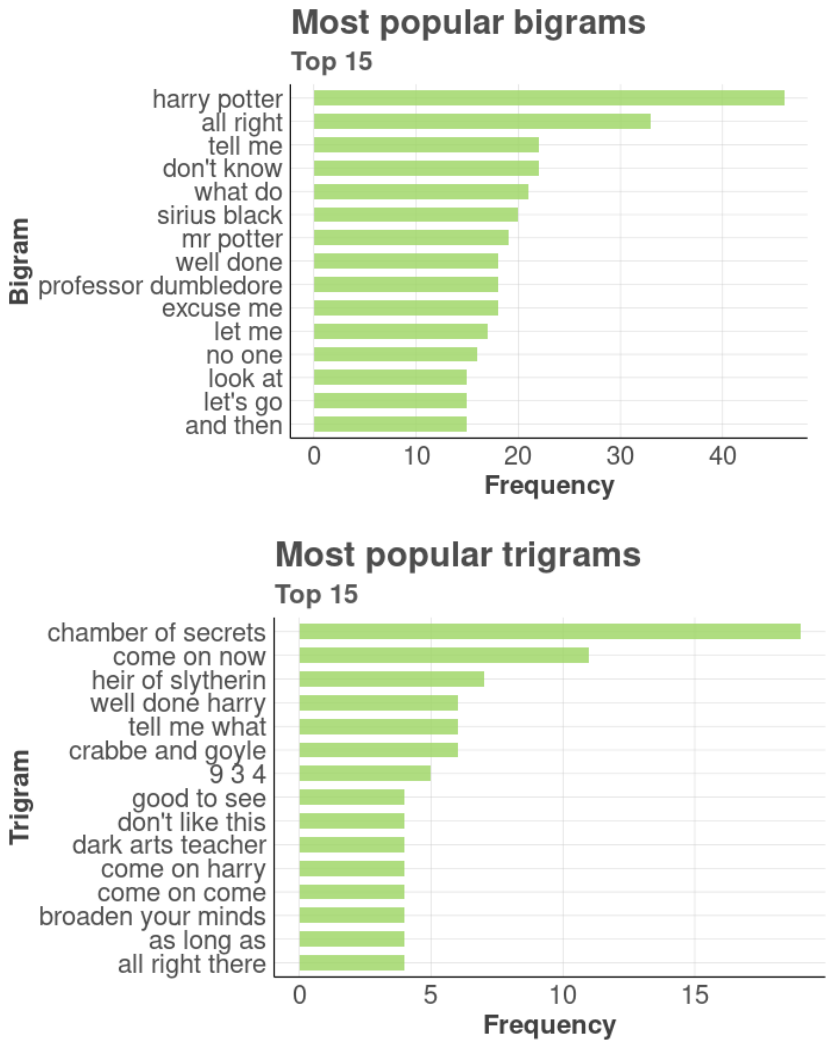
Picture 2. WordCloud of the most frequent unigrams in all three movies



Picture 3. Top 15 words that appear in the first three Harry Potter movies.

If we check the proportions of these top 15 words according to the movie part, we would notice an almost balanced distribution. If we exclure Harry's name, we would notice that there is no word that orruces more than 70 times in one movie.

Top 15 bigrams and trigrams are presented in *Picture 4*. First of all, we have proven our hypothesis number 1 that Harry Potter is the most common bigram and its parts are at the top of unigrams as well. Even its version of 'mr potter' appears on the 7th place). Secondly, there are also only two names in the top list: professor Dumbledore (although his actual name is Albus Percival Wulfric Brian Dumbledore)



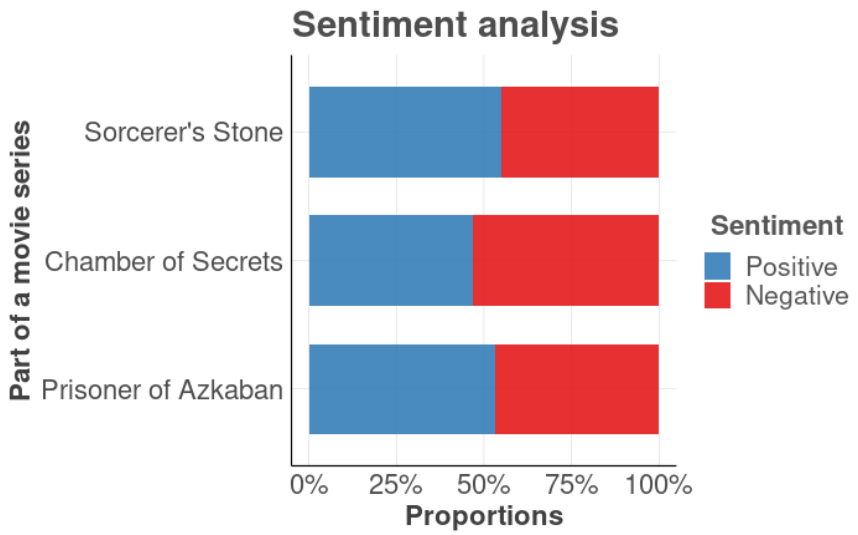
Picture 4. Top 15 bigrams and trigrams in the first three movies.

and Sirius Black. As for trigrams, there are several phrases with 'com on' in it. Another interesting fact is that neither Crab nor Goyle appear in the top unigrams but the trigram 'Crab and Goyle' is at the 6th place among trigrams. Based on that we can see how inseparable these characters are. 'Heir of slytherin' and 'chamber of secrets' are at the top of trigrams although the main action related to them happens only in the second movie.

The next step in our analysis would be sentiment analysis of given words. With the use of Bing dictionary (it contains around 7 000 words with their sentiment alignment). Our hypothesis was that the proportion of positive words decreases from film to film. If you watch all Harry Potter movies one by one you would notice that the first movies are lighter and more naive than the last one. Let us check this for the first three movies.

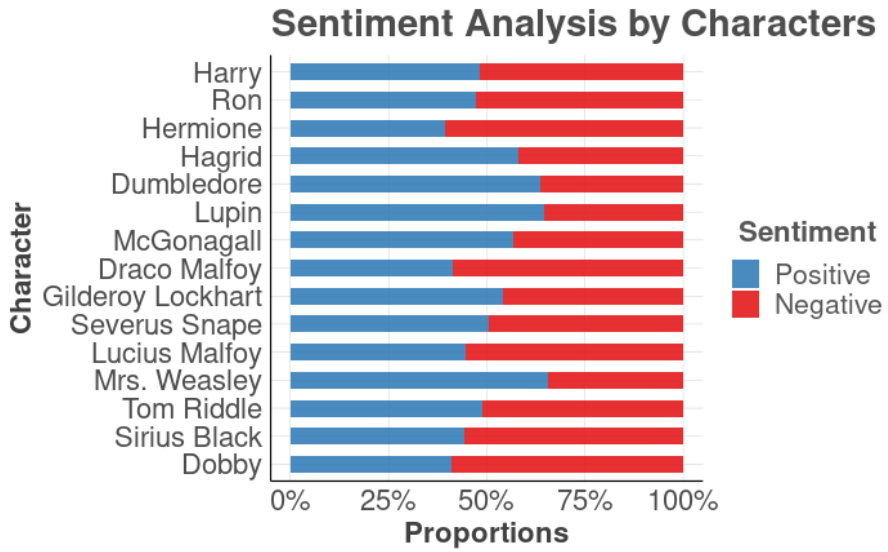
From *Picture 5* we can see that this hypothesis is not true for all movies. The decrease in the proportion of positive elements is noticeable between the first and the second movie, but in

the third movie the situation reverts itself. In Prisoner of Azkaban there are more than 50% words that are positive according to the Bing dictionary. Of course, the full dynamic patterns should be analysed based on all eight movies together.



Picture 5. Proportions of words with positive and negative sentiment

The final step in our analysis is to check the distribution of positive and negative words in characters' perspective.



Picture 6. Proportions of positive/negative words according to movie characters

There are several insights that we can see: 1) Dumbledore, Lupin and Mrs. Weasley (positive characters themselves) have more positive words in their lines than negative, which is not surprising. 2) Draco Malfoy and Lucius Malfoy (negative characters themselves) have more

negative words in their vocabulary, which is also not surprising. 3) Hermione and Dobby (positive characters themselves) have more negative words in their sentences than positive. And this is surprising. If for Dobby we could suggest that this happened because he made some mean thing to Harry at first and was a questionable character. But for Hermione, there is no explanation at this point. Finally, Tom Riddle, although he is a negative character (or rather his incarnation in Lord Voldemort) has a sentiment very similar to Harry and to the average, close to an equal share of positive and negative words in his statements.

Conclusion

Overall, we have proven that our first hypothesis is completely true: ‘Harry’ is the most popular word to appear in the first three movies, ‘Harry Potter’ is the most common bigram as well.

As for the second hypothesis, it is true only for the first two movies out of three. First and third movies have the advantage of words with a positive character, while the second one with a negative character, but the values in all cases are close to 50%.

Moreover, several interesting insights came up during the analysis: in bigrams and trigrams we can find expressions known from everyday speech, as well as expressions that we only meet in this movie saga; Hermione has more negative words in her lines than positive.