Term Project: Comparison between Harry Potter 1 and 7 – What changed?

Motivation and Research Question

The novels about Harry Potter developed from being called light fiction in the beginning to the most influential book series of all times. As there was a time span of 10 years between the release of the first book "Harry Potter and the Sorcerer's Stone" and the seventh and last book "Harry Potter and the Deathly Hallows" and a "fictional time span" of 7 years in Harry's life it is an obvious question if any changes can be investigated. These changes can be literal – What changed in Harry's life? – as well as on the level of authorship – Did the author change her writing style?

For investigating these issues several questions were posed:

1. Which literary figures do occur most frequently?
2. What are the proportions of stop words and content words?
3. What is the type-token-ratio? I.e. How high is the vocabulary richness?
4. What are the proportions of positive and negative words? I.e. did the atmosphere change?

The Code – Prepatory work

I. For excluding the capitalized stop words in the search of named entities later in the main program, I wrote a code (uppercase.pl) that capitalizes all the stop words.

II. Lists of stop words, negative and positive words have been copied from the web. As the words were all written among each other, I wrote a program (wordlistprocessing.pl) that opens the document with this list and removes all the newlines in it. The words were then pushed into an array. For doing so, I used a subroutine as the process is the same for the four lists.
Having the lists processed, I could easily copy the output word lists and insert them into the main program.

The Main Code

Apart from opening the directories and the welcoming of the user, the code consists of one big foreach-loop i.e. the exact same processes were applied on both texts.

I. First, newlines and punctuation were removed from the texts (line 53-56). The apostrophe has been left in in order to avoid having "*Im" as one of the most frequent capitalized words.

II. If a word was capitalized it was added to a hash for making named entity recognition possible later on (line 68-74).

III. All the words were being decapitalized (line 77-79).

IV. Every word was put into a hash for counting the types. Note that this is not a scientifically correct way of calculating the type-token-ratio as no lemmatizer has been applied (line 81-86) but it can be seen as a toy-TTR-approach.

V. With help of the stop words array and the array of all the words, the stop words and the overall amount of words were counted. By subtracting the stop words from the number of all the words, the number of content words could be calculated, too (line 90-95).

VI. Also, the amount of negative and positive words was counted (line 125-135) and used for creating a ratio (line 138-145).

VII. The hash of the capitalized words and the hash of capitalized stop words were iterated. If a capitalized word was not part of the stop words and if it did not contain a genitive-s (e.g. “Harry’s”), it was assumed a named entity and the 10 most frequent of them were output (line 157-168).

During the processing of the texts, one CSV file and three txt-files for an overview of the data were created.

THE OUTCOME

infos.csv and relations_HP1.txt/relations_HP7.txt

The txt-files about the relations in the text describe what is put in the CSV file shortly. If we compare the two texts, you can see that the 7th Harry Potter has a lot more words than the first one. Anyway, the type-token-ratio in the first book is higher, which means that the vocabulary is richer in the first book. Also, it becomes obvious that more stop words are being used in the 7th book. Therefore, one can assume that the author’s style indeed has changed even though I do not want to give a normative evaluation of this. Also, the result of the processing says that more negative words are being used in the 7th Harry Potter. It seems like the atmosphere in this last book is darker and gloomier than in the first one.

personsOfHarryPotter.txt

In this file, the most frequently occurring persons in the novels are shown. There, one can see how the name “Harry” is used in the exact same amount in both of the books. Apparently, Ron and Hagrid were mentioned more often than Hermione in the first book, but Hermione played an even bigger role in the 7th book.

Of course, “Professor” is not a named entity by itself, but it shows that the children’s teachers are mentioned a lot in the first book and not so much anymore in the last one (remember: in large parts of the 7th book Harry and the others are not students at Hogwarts any more). It’s interesting that Dumbledore, although he is already dead in the 7th book, plays a bigger role in that very book. Also, “Death” is not a named entity but its occurrence in this list shows how important the concept of death is in the last book. A further hint for the story becoming darker is the appearance of Voldemort only in the second list and Dudley (as he sure is Harry’s enemy, but by far not as bad as Voldemort) only in the first list. Also, the last names “Potter” and “Weasley” coming up in the second list might show how Harry’s and all the other’s lives might have turned a little more serious in the course of time.

EVALUATION OF THE CODE

The code surely could be more elegant and lean but I tried to keep it as universal as possible by not excluding certain values with regard to only Harry Potter (e.g. I could have excluded “Harry’s” in the list of persons, but I rather took the little detour by first leaving apostrophes in the text and then excluding genitive-s in the listing of persons). Like this, the program can be applied on other texts without losing its power.