**Analysis of the “Harry Potter” books**

**Report**

**Group members:** Guseva A., Malysheva T., Oreshnikova V.

**Problem Statement:** The aim of the study is to learn as much as possible about the plot of the first Harry Potter book and to compare it with other two books of that series only with the help of Python codes and NLP libraries.

**Dataset:**

For our analysis, we utilised the “[Harry Potter books in pdf (1-7) (kaggle.com)](https://www.kaggle.com/datasets/hinepo/harry-potter-books-in-pdf-1-7/data)” dataset, which encompasses 7 individual PDF files, containing the complete collection of Harry Potter books in English. Our approach involved leveraging the first book for an in-depth study of the plot, while books 1, 2, and 3 were subjected to a similarity analysis. To facilitate our analysis, we transformed each book into a txt file format, enabling further processing and examination.

**Steps for making this project:**

1. Find a suitable dataset
2. Transform the data into txt.files
3. Compile all methods which will be suitable for revealing and comparing the plot of the books
4. Distribute the responsibilities among the members of the group
5. Set deadlines for writing the codes
6. Join all the parts of the code together, which was made by different members of the group
7. Distribute responsibilities among group members connected with making a presentation, adding the code to GitHub and writing a report
8. Set deadlines

**Methods of the project:**

1. Lowercasing
2. Punctuation removal
3. Stop words removal
4. Parts of speech tagging
5. POS-tagging
6. Tokenization
7. Stemming
8. Lemmatization
9. Exploratory data analysis (WordCloud)
10. N-grams
11. Term Frequency-Inverse Document Frequency (TF-IDF)
12. Text similarity
13. Information extraction – NER – Entity recognition
14. Topic modelling
15. Sentiment analysis

The detailed explanation, how the methods worked and how we applied them, is made on Github in the section with our project: [Final-Project/The\_analysis\_of\_Harry\_Potter.ipynb at main · Nastya-00/Final-Project (github.com)](https://github.com/Nastya-00/Final-Project/blob/main/The_analysis_of_Harry_Potter.ipynb)

**Problems during the project and how we solved them:**

1. During the TF-IDF code writing process, an issue arose with the expected iterable over raw text documents, as a string object was received instead. It was resolved by converting the text to a list of strings using [text] from sklearn.feature\_extraction.text.
2. The code for sentiment analysis took a considerable amount of time to process. However, we discovered that this is acceptable, considering that it encompassed the analysis of the entire book.
3. In an effort to avoid the need to "clean" the dataset each time and import libraries, our initial plan was to clean it once and subsequently allow groupmates to use it in their own code cells for further manipulations. Regrettably, the new code cells did not have access to the operations from previous code cells when added, resulting in the necessity to forgo the initial idea and instead import the necessary libraries anew in each new cell while also cleaning the code where necessary.
4. Some of the outputs were quite big with a lot of data, that’s why it was difficult to interpret the results and with some methods to avoid it we asked the code to show the most frequent results or to check if the answer appears more than time, and ask in that situation only once to show the results.

About **the results of the project** you can learn during our defence and on the GoogleSlides: [Analysis of the book 'Harry Potter' - Google Презентации](https://docs.google.com/presentation/d/1FBYvfLlwrvjcnHC1e7vr4IVvan9xCpD_yyidNnqp1Vc/edit#slide=id.g2b4d6b14468_0_285)

**Further perspectives of the project:**

We are currently concentrating on a single book and have used the similarity method to compare the first three books with each other. However, in the future, we can extend the analysis to encompass the entire series of Harry Potter books, enabling us to interpret the results and compare the plotlines across the entire series.