Coursework Report for Module INM433 “Visual Analytics”

Stefan Diener - 210008139

**Abstract**—Put here a brief summary of your work: analysis task, data, approach, main findings. Length: up to 200 words.

# Problem Statement

The shift of modern communication to the digital domain is the catalyst for an ever-increasing volume of digital texts. Computational methods for visualizing and analyzing text data that are able to help humans understand these large text corpora are therefore becoming increasingly important in the scientific literature [1].

This study uses the famous Lord of the Rings (LOTR) novels by J. R. R. Tolkien, to show how the application of text processing techniques can lead to a deeper understanding of the text at hand. More precisely, it aims to answer the following research questions:

1. Does the visualization sentiment and keywords across time detect key parts and tuning points of the story?
2. How can text summarization with word clouds help identify and understand certain concepts in the text?
3. Can network analysis help detect the main characters in the story and reveal insights about their relationships to each other?

The data used for the following analysis is composed of two parts: first, the books of the LOTR trilogy in text format, and second, additional data on all characters. This data is suitable for answering the research questions because it covers the entire LOTR story, because it is in an unstructured text format and because it contains a comprehensive list of all characters, which makes it possible to find the most important characters and analyze their relationships without prior knowledge of the story.

# State of the Art

This section provides an overview of three papers that leverage computational analysis in tandem with effective visualizations to provide insight into the text corpus at hand.

In the first study, Mohammad [3] shows how sentiment analysis can be used to quantify and track the polarity and emotions of mail and books. Here, polarity refers to the positivity positive or negativity of a text, whereas emotions refer to the eight distinct measures: joy, sadness, anger, fear, trust, disgust, surprise, anticipation. To quantify each text along any of those dimensions, a lexicon-based approach is used, where each word of a text is compared to a dictionary that associates a vocabulary of words with a certain polarity or emotion. With this approach he compares the polarity and emotional content of different kinds of mail, like love letters, hate mail, suicide notes, or mails written by men versus women.

One of the reasons why Mohammad’s paper is highly relevant the subsequent analysis of this study is because he not only applies these techniques to short texts like mail, but also shows how these techniques can be applied to entire books. By calculating the emotions for each line in a novel, Mohammad is able to track the development of certain emotions across the story. Furthermore, uses a word cloud to visualize the words that are associated with a certain emotion, which helps to convey a deeper understanding of what these emotions are based on in the context of the story.

In the second paper, Heimerl, Lohmann, Lange, and Ertl [4] build a system called Word Cloud Explorer that aims to improve the power of basic word clouds for text analysis, by leveraging interactivity, natural language processing techniques and context information, like e.g., part of speech tagging. To test their hypothesis, they conduct a qualitative user study, where a group of analytics professionals completed analytics tasks on three different corpora using the new software. They concluded that its main advantages were increased flexibility and intuitiveness. Hence, this study shows that combining word clouds with additional text processing techniques can be an effective way for exploring and analyzing text data.

In the third paper, Rydberg-Cox [5] analyzes a corpus of Greek tragedies with social network graphs. His goal is to discover quantifiable patterns about the tragedies and utilize visualizations of these networks to communicate the patterns. By representing each character as a node and connections between characters as edges in the network graph, he was able to identify four distinct patterns across the corpus.

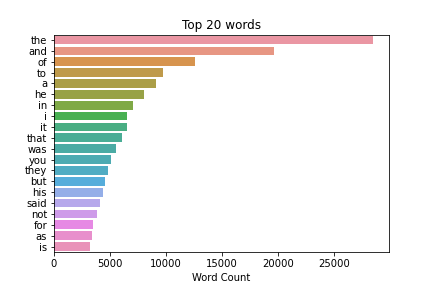
This work shows how network analysis can be used to model relationships between characters, however, Rydberg-Cox also states that the limited number of characters in each tragedy strongly contributed to the emergence of clear patterns. Thus, when applying network analysis, it seems to be useful to think about whether the number of characters included in the analysis can be limited in a meaningful way in order to increase the interpretability of the network graph.

# Properties of the Data

The data of two main components: first, the books of the LOTR trilogy, and second, demographic data on all characters in the LOTR universe. The data was downloaded from a public GitHub repo that used [https://archive.org](https://archive.org/) and [www.ageofthering.com](http://www.ageofthering.com) to scrape this information from the web [2].

Each part of the LOTR trilogy is stored in a text file that represents the entire physical equivalent, from the title and contents to the footnotes at the very end. The structure of the LOTR series is somewhat unconventional and is as follows: First, the trilogy divided into three parts “The Fellowship of the Ring”, “The Two Towers”, and “The Return of the King”. Secondly, each part is divided into two so-called books, which adds up to a total of six books in the complete trilogy. Lastly, each of the six books contains between 9 and 12 chapters. In total, the LOTR series contains about 470 thousand words. And uses a vocabulary of ca. 12 thousand words.

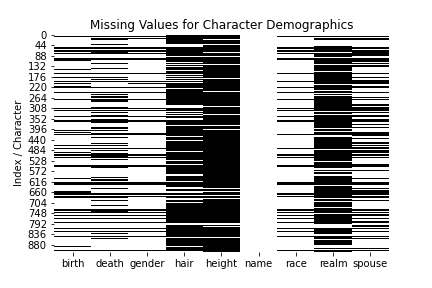
One of the biggest challenges with this data is that it is inherently unstructured, with each of the three text files essentially containing only a very long string of characters. In conclusion, for further analysis it is important to represent the data in a more structured way, that allows for the calculation of metrics e.g., over time or per chapter.

By manually skimming through the text data, it seems like the data is a very accurate representation of the original books. Encoding errors, like missing, swapped, or double characters, additional white spaces within a word, or missing white spaces between words, seem to be very rare or potentially non-existent. However, typical for text data, it contains punctuation, capitalization and many stop words, which are often not useful for analysis. This is exemplified by Figure 1, which shows that the top 20 words in the data are exclusively stop words.   


Hence, the text data will require additional cleaning in the early stages of analysis in order to reduce the ratio of noise to useful information and to effectively answer the research questions.

The second data component, the demographic data on the characters, contains 911 rows, each corresponding to a distinct character, as well as 9 columns, record information about their date or birth, date of death, gender, hair color, height, name, race, realm of origin, and name of their spouse.

However, the data is characterized by incompleteness. This is shown by figure 2, which plots the data frame and marks every cell with a missing value in black.



Counting the number of times each name is mentioned in the trilogy revealed that 729 characters never appear in the LOTR books. Further investigating revealed that the data also includes characters other books about middle earth like ‘The Hobbit’ or the 12 volume book series ‘History of Middle-earth’. In conclusion, to analyze the character dynamics in the LOTR books in a meaningful way, the characters must be filtered down to only include the most important ones.

# Analysis

## Approach

This section explains the general analytical approach that is used in the subsequent chapters and furthermore goes more in depth on it is applied to answer the research questions of this study. Figure 3 shows a graphical representation of the approach. It is divided into three distinct steps: Data Preparation, Analytics, and Human Reasoning.

The data preparation step aims to convert the data into a format that facilitates further processing and functions as the foundation for all the subsequent analyses. The analytics step, then uses more complex data derivation and modelling techniques to uncover latent patterns in the data. Here, visualization bridges the gap between the purely computational methods, and the last step in the analytical approach, human reasoning. Finally, human reasoning uses the capabilities of human cognition, pattern recognition, and domain knowledge to derive conclusions about the data and the analysis methods used. These new insights then allow the analytics step to be refined and more insights to be generated. This iterative process is illustrated by the feedback loop that links the analytics and human reasoning step. To answer each research question, multiple iterations are necessary, before the final conclusions are reached.

Diagram

Description automatically generated

For this study, the data preparation step has a very high significance, given that the majority of data used in this study consists of unstructured text. In order to conduct meaningful exploratory data analysis (EDA) and facilitate further analysis the text data must be divided into distinct units like e.g., words, sentences, lines, paragraphs, pages, or chapters. This already requires some human reasoning, as the optimal level of aggregation is not obvious and could also vary depending on the use case.

The first research question relates to the performance of word clouds. Here, the role of human reasoning is two-fold. First, it is essential to maximize the potential of word clouds, e.g., by minimizing the amount of uninformative stop words in the text and by using filtering the corpus in a meaningful way to analyze certain parts corpus more in-depth. Secondly, human reasoning is required to recognizing patterns in the final word cloud outputs, as this requires an understanding of the meaning and the context of the shown words.

Second, network graphs are an effective tool for visualizing the co-occurrence of keywords and characters and the context in which they appear together. However, human reasoning is required for choosing the hyperparameters that optimize the interpretability of the final graph. Additionally, to derive new knowledge and confirm or deny existing hypothesis about the relationships of the analyzed keywords contextual knowledge is required that only a human analyst can provide.

Similarly, sentiment values and key words occurrences in the story timeline can highlight unique moments and developments in a text. However, to identify overarching patterns and to understand why these highlighted parts might be important, they must be analyzed by a human analyst who has additional information about the context of the story.

In a last step, human reasoning decides when the iterative analytical approach is stopped, and the final results of the analysis can be compiled.

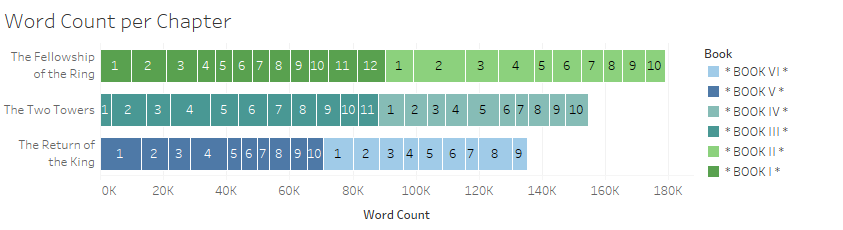
## Process

In this section the analytical approach described above will be applied to answer the following three research questions:

1. Does the visualization sentiment and keywords across time detect key parts and tuning points of the story?
2. How can text summarization with word clouds help identify and understand certain concepts in the text?
3. Can network analysis help detect the main characters in the story and reveal insights about their relationships to each other?

*Data Preparation*

The analytical process begins with data preparation. As described earlier, the majority of the data used for this study consists mainly of unstructured text data, which is difficult to analyze and investigate in its raw form. To mitigate this the three text files – each corresponding to one part of the LOTR trilogy – were joined and loaded into a pandas data frame. Each cell of the data frame corresponds to one new line in the text file, which is achieved by finding and splitting the text at the ‘\n’ symbol. However, with a total of 38’443 lines in the text, this representation seemed too granular for meaningful analysis. Thus, regular expressions were used to find all lines that contain a book or chapter heading and stored them in an extra column. With these markers in place, the data was grouped by chapter and the text for each chapter was joined together with a whitespace. The resulting data frame thus contained 62 rows, corresponding to the 62 chapters, and 4 columns containing the chapter text, and additionally the names of the corresponding, part, book, and chapter.

Figure 4 visualizes this structure and the word count per chapter.

Just by itself, this representation of the data provides insights, e.g., it reveals that there are no missing books or chapters in the data, that the chapters are in the correct order and are similar in length. Additionally, the 62 chapters are granular enough to conduct analysis over the story’s timeline, while having the additional benefit that the content of each text block is uniquely summarized the chapter title. Thus, this structure is chosen as the baseline for further analysis.

Following paragraphs...

*<1500 words, <=7 images*

## Results

First paragraph...

Following paragraphs...

*<200 words, <=2 images*

# Critical reflection

First paragraph...

Following paragraphs...

*<500 words*

Table of word counts

|  |  |
| --- | --- |
| Problem statement | 250 |
| State of the art | 500 |
| Properties of the data | 500 |
| Analysis: Approach | 500 |
| Analysis: Process | 1500 |
| Analysis: Results | 200 |
| Critical reflection | 500 |

References

The list below provides examples of formatting references.

1. M. Gentzkow, B. Kelly, and M. Taddy (2019). Text as Data. Journal of Economic Literature, 57(3), 535–574. https://doi.org/10.1257/jel.20181020
2. T. Gu, Lord\_of\_the\_ring\_project, (2018), GitHub repository, https://github.com/tianyigu/Lord\_of\_the\_ring\_project
3. S. M. Mohammad (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. Decision Support Systems, 53(4), 730-741.
4. F. Heimerl, S. Lohmann, S. Lange, and T. Ertl (2014). Word cloud explorer: Text analytics based on word clouds. *47th Hawaii International Conference on System Sciences*, 1833-1842. IEEE.
5. P. Mutton (2004). Inferring and visualizing social networks on internet relay chat. *Eighth International Conference on Information Visualisation*, 35-43. IEEE.