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

First paragraph...

Following paragraphs...

*<250 words*

# State of the Art

First paragraph...

Following paragraphs...

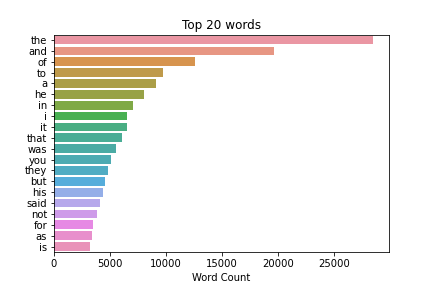
*<500 words*

# Properties of the Data

The data used for the following analysis is composed of two main components: first, the books of the LOTR trilogy, and second, additional demographic data on all characters of 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 [1].

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.

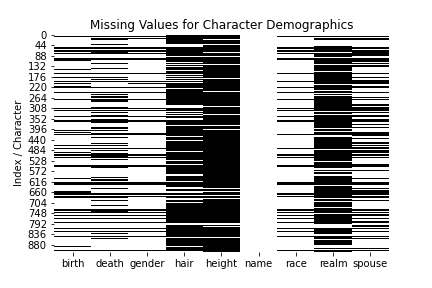
One of the biggest challenges with this data is that it is inherently unstructured, with each of the three text files essentially containing only a very long string of characters. In conclusion, for further analysis it is important to represent the data in a more structured way, that allows for the calculation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

First paragraph...

Following paragraphs...

*<500 words, 1 diagram*

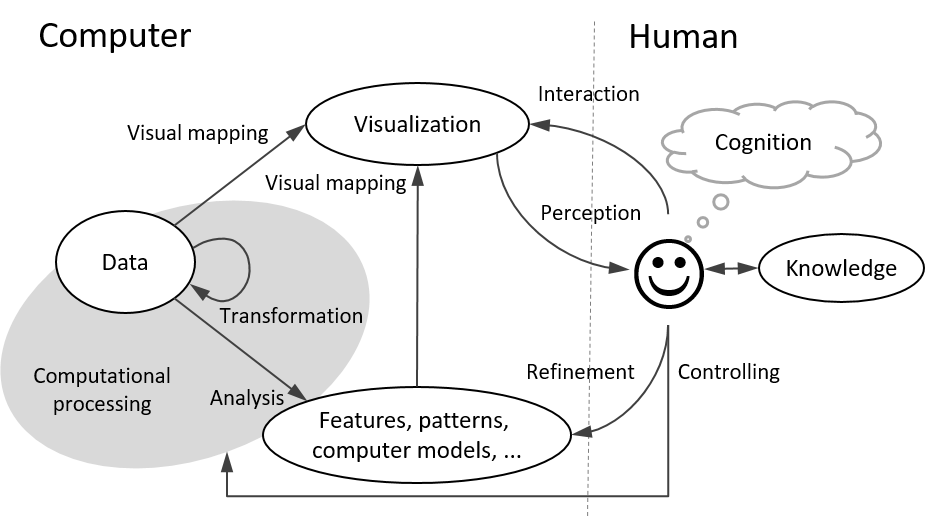


Fig. 1. An example of including a diagram in the document.

## Process

First paragraph...

Creating Dataframe:

To mitigate this the three text files loaded into a pandas data frame, where each new line recognized by the ‘\n’ symbol corresponds to one cell in the data frame. Next, the three data frames were joined, and regular expressions were used to find all lines that contain a book or chapter heading. With these markers in place, the text per chapter was joined. 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. 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, and that they are in the correct order. Additionally, this structure allows for the calculation of statistics per chapter, which is a useful characteristic for many of the subsequent analyses.

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. T. Gu, Lord\_of\_the\_ring\_project, (2018), GitHub repository, https://github.com/tianyigu/Lord\_of\_the\_ring\_project
2. M. Bögl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch, and A. Rind. Visual Analytics for Model Selection in Time Series Analysis, *IEEE Trans. Visualization and Computer Graphics*, 19(12): 2237-2246, 2013.
3. T.F. Cox, M.A.A. Cox. *Multidimensional Scaling*. Chapman and Hall, 2001.
4. S. van den Elzen and J.J. van Wijk. BaobabView: Interactive construction and analysis of decision trees, *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST’11)*, pp. 151-160, 2011.
5. M.Harrower, C.A.Brewer: Colorbrewer.org: An online tool for selecting color schemes for maps. *The Cartographic Journal* 40(1): 27–37, 2003.
6. T. Mühlbacher and H. Piringer. A Partition-Based Framework for Building and Validating Regression Models, *IEEE Trans. Visualization and Computer Graphics*, 19(12): 1962-1971, 2013.
7. D. Phan, L. Xiao, R. Yeh, P. Hanrahan, T. Winograd. Flow map layout. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis 2005)*, pp.219-224, Oct. 2005.
8. J.W. Sammon. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, 18(5): 401–409, May 1969.
9. J. Sewall, D. Wilkie, and M.C. Lin. Interactive Hybrid Simulation of Large-Scale Traffic, *ACM Transactions on Graphics*, 30(6), Article 135.
10. N. Willems, H. van de Wetering, J.J. van Wijk. Visualization of vessel movements. *Computer Graphics Forum*, 28(3): 959-966, Jun. 2009.