Analyzing the Works of H.P. Lovecraft

Projektarbeit zum Thema Topic Modelling in Digital Humanities
von Ferris Kleier

Einführung in Digital Humanities, Wintersemester 2022/23

Autor: Ferris Kleier
Matrikelnummer: 3732130
Studiengang: Informatik B.Sc.
Datum: 14. Februar 2023

1 Introduction

This project aims to analyze the works of Howard Phillips Lovecraft, one of the most known horror authors, in regard to their themes and patterns. This will be done using digital humanities and topic modeling on all of his writings to collect a timeline-like collection of the most common themes in his works. H.P. Lovecraft highly contributed to the modern horror genre with his detailed descriptions of cosmic horror (a term defined by his writings) and stories about monsters and scary events. While his writings are known for complexity and style, his thematic and contextual depth is equally impressive and the subject of this project. Therefore, Lovecraft is well-known by fans of horror and science fiction literature, like Stephen King, who stated that Lovecraft heavily influenced his style and ideas.

Even though Lovecraft lived from 1890-1937 and did not encounter computers or other devices of our current age, his works and letters are digitalized and preserved for reading and research. One key aspect of this project is to apply the digital humanities method of topic modeling on all of Lovecraft's works. The digital humanities is a field of study, research, teaching, and invention concerned with the intersection of computing and the disciplines of the humanities. One can think of it as a bridge between the two cultures of natural science, which tries to explain what is going on, and humanities, which tries to understand what is happening. One key figure in this field is Roberto Busa (1913 - 2011), who used computational methods to create a collection of all the words of Thomas Aquinas, later known as the Index Thomisticus. He achieved his goal with the help of IBM, at that time one of the biggest computer manufacturers, and connected his studies with the help of digital computing to process the millions of words he strived to process.

By using the digital humanities, we can gain new insights on Lovecraft's writings and identify recurring themes, patterns, and motifs in the big corpus of his writings. The methods used in the digital humanities include Stylometry, Topic Modeling, Network Analysis, and Geovisualization. For the purpose of this project, we will use topic modeling to gather insight into the large corpus and structure the results like a timeline for every one of his works. Regarding digital humanities, this project aims to explore the humanities of literature and apply computational methods on it. That way we can see the most occurring themes from his first works to his last ones and compare them in terms of lore consistency (how did his themes and mythos change?), social correlation (does he refer to social events of his time?) and comprehensive insight on the different features of his

works. Some works already covered features of his writings and life like the fear of progress, dangers of scientific progress, and persistence of ancient evils.

2 Research Agenda

The research agenda of this project is 'what are the most dominant topics in H.P. Lovecraft's writings over his lifespan?'. The goal is to understand how they relate to Lovecraft's personal life experiences, historical events, cultural and social influences, as well as to the broader literary and intellectual traditions of his time. To answer that question we will apply topic modeling to each of his works in chronological order. By comparing the resulting main topics we hope to find changes in the lore of his mythos as well as correlations to the main events in his life.

We will use computational methods like text processing and tokenization to represent the main themes of Lovecraft's writings from a corpus of all of his works. Because Lovecraft died over 70 years ago, many of his writings are public domain and can easily be accessed from various institutions providing good coverage for the material. Nevertheless, some of the works he contributed are still not public domain, since some authors he worked with lived on longer than he did (e.g. Clark Ashton Smith died 1961). This is part of the reason why we decided to only cover works solely written by Lovecraft himself. The more important reason is that mixing up his writings with ghostwritten or co-written works could alter the results since we only want to cover Lovecraft's themes. Using the programming languages Python and R, we first format the corpus into a collection of all works (using Python), then process the texts to remove annotations and unnecessary titles and use topic modeling on the corpus to gather and represent the data (using R). After that, we will analyze the gathered data represented as a plotted diagram in R. That way we can create a timeline-like graphic of the main themes and motifs in Lovecraft's works and discuss them in retrospect to events or that time or his personal life.

Lovecraft works can be distincted into phases. We will stretch the phase 'Early Work' up to his first writings, when he was a child. That gives us the following phases:

- Phase 1: Early Work (1905-1920)
- Phase 2: First Mythos (1920-1923)
- Phase 3: Middle Years (1926-1928)
- Phase 4: Later Mythos (1930-1934)

We will take these phases of his life into account when comparing his lifetime events with the results from the topic modeling later.

To properly find correlations between his writings and private life or social events of his time, we will shortly examine his letters as well as his biographical records. After that, we include the ideas and insight into the discussion. We expect to find results in this aspect, since Lovecraft is known for his rejection of modern movements of that time. He was known for his dislike of New York City, after living there for a period of his life. This is one of the interesting parts of his life we hope to find reflected in his writings. We already know of the story 'The Horror at Red

Hook' (1925) in which Lovecraft negatively reflects on Red Hook, Brooklyn. It is also important to mention that Lovecraft condemned ethnical groups he encountered during his time in New York, mainly Afro-Americans and Asian immigrants. It is no secret that Lovecraft used racist slurs and sometimes even antisemitic stereotypes in his writings, all of which will be analyzed according to his personal events to explore how his prejudices shaped his writing. For all of these aspects of his life, we hope to find a similar pattern in his writing through which one could pair a batch of his works to a phase of his life.

3 Data Overview

The data for this project can be found on Kaggle. We decided to go with this dataset because it features all of Lovecraft's written texts in .txt. format, which makes it easy to process the data. All stories exist in public domain and are not subject to copyright. Due to the stories being public domain, and more datasets being available online as well as raw HTML texts available, we had many sources to choose from. A comparison with other sources and physical prints of stories does not indicate any biases or alteration in the texts. Though we have found some works in the dataset that are not solely written by Lovecraft, since he collaborated with many other authors of his time and even took ghostwriting commissions. Another problem also occured with other datasets, where the earliest works of Lovecraft are missing. We decided to still go with this dataset and remove works which we don't want to include. Since our research question wants to only cover Lovecraft's ideas and writings, the influence from other authors or commissioners could alter the results. That's because the ideas of Lovecraft are not really present in these specific writings, though his stylometry may be the same.

The dataset contains 102 works of Lovecraft, including all collaborations. As we just mentioned, collaborations, revisions and ghost writings are not subject of this project and will be removed. An overview of all the works written solely by Lovecraft, his fiction, can be found in his bibliography on Wikipedia. In this case only his 'fiction' will be included in this project as well as some of his works found under 'Juvenilia', which include his earliest writings when he was young. Also not included in this project are his poetry works, philosophical works and scientific works since they are either too short or provide no information in regard to his mythos and lore. After only keeping the works of Lovecraft from the bibliography the dataset has remaining 70 stories. These works can be distinguished into short stories, novellas, or even fragments from letters or novellas. Lovecraft's works largely vary in length, some just a few pages, and other whole novellas. Fragments could also contain just a few paragraphs. One example is the story 'Azatoth' (written June 1922), which is very short in length and was supposed to be part of a whole novella, though only this fragment survived. It is currently unknown how many such fragments were lost, but his wife admitted to have burned a lot of his letters. If these included other fragments or whole works of Lovecraft not yet published or copied at that time, that means there were works which are lost forever.

The 70 works in this modified dataset are .txt documents. The text files contain the title and the raw text. We formatted the files using R and pasted them into a .csv format. In this process we also added the dates as a second row to the text and manually put another line between paragraphs process them using regular expressions. The file to process the dataset can be found in the repository as 'format.r'. The .csv file contains for every text the text_id, then for every paragraph of the

text the doc_id, the title, the date, and the paragraph containing the text. This way we can later use topic modeling on the csv to create a timeline-like plot of the main themes. The dates found in different sources can be misleading, since some stories were published after Lovecraft's death. One example is the story 'The Case of Charles Dexter Ward', which was written in 1927, partly published in 1941, and fully published in 1943, while Lovecraft already died in 1937. The dates we manually added to each file represents the date a work was written as provided on WikiSource. This corresponds better to his personal life and events. For dates where only the year is provided (see 'The Alchemist', 1908) or no explicit month (see 'The Street', late 1919) we used proper months according to quarters or the beginning of seasons. For plain years we added the January of that year. Taking into account publishing dates is no option, since they do not reflect Lovecraft's personal events, may not be in order, or even have been published after his death as already mentioned.

4 Methods Overview

In the digital humanities, there are four main methods to work on data and achieve information. Stylometry is a method to find textual similarities between texts, Network analysis can be used to find relations in data, Geovisualization is used to represent data on maps, and the fourth method, which we will use for this project, is topic modeling. Topic modeling is a method to gather information from texts regarding the content. Using this computational approach we can uncover the main themes and topics from texts, not just using a frequentist method with the most words used, but putting context in the found information. A topic model achieves this while it discovers the degree to which each document exhibits those topics. That way we can build a statistical lens that encodes our specific knowledge, theories, and assumptions about texts.

Using topic modeling on the collection of Lovecraft's works will help us uncover the most important themes per text in regard to not only the frequentist quantity of certain topics but, as already mentioned, the contextual frequency. We want to know what topic dominate each text, and topic modeling is exactly the tool we need for that. By writing short scripts in the programming language R, we can compute the main topics easy and even visualize them in plots and well-known word clouds. For example, the R library ggplot2 is an easy tool to plot the results in a timeline style plot. We will apply topic modeling on each text. This is done because if we would use chunks of text over the whole concatenated collection, we would loose the important context. Especially in this literature of fantasy and horror, we need to keep the topics in regard to their text. In one text like 'The Thing on the Doorstep' (written 1933) the main topic may refer to 'creature', 'darkness', house' or others. But another text like 'The Nameless City' (written 1921) may yield topics like 'desert', 'race', 'ancient'. When using chunks of a concatenated collection of his writings, we would use the context to each text. Therefore we work on each text separately, even if some texts are significantly shorter than others.

To reflect on some biases with this approach, we have to particularly emphasize this problem: Some works of Lovecraft are shorter than others. Though we do not expect any major differences in contextual topics between short and long texts, this has to be noted. According to a word count on Lovecraft's works the length of his works linearly increase with the years of writing. Another major bias in topic modeling is the approach of Natural Language Processing (NLP). For our purpose we

will work with english libraries, stopwords and dictionaries. This could lead to problems when it comes to topics that do not reflect the english language. Lovecraft is known for his monsters and cosmic words like 'Necronomicon' or 'Cthulhu'. For this, we will create a separate list of the most important words regarding Lovecraft's deities and locations. That will allow us to work on these words as well and include them as topics if they occur dominantly in a given text.

5 Related Work

6 Experiment Design

7 Results and Discussion

8 Conclusion

9 References