Analyzing the Works of H.P. Lovecraft in Thematic Correlation to his Personal Life

Projektarbeit zum Modul: Einführung in Digital Humanities, Wintersemester 2022/23

Autor: Ferris Kleier
Matrikelnummer: 3732130
Studiengang: Informatik B.Sc.
Datum: 05. März 2023

1 Introduction

This project aims to analyze the works of Howard Phillips Lovecraft, one of the most known horror authors, regarding 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 horror and science fiction fans, 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 to 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 into Lovecraft's writings and identify recurring themes, patterns, and motifs in the big corpus of it. 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 to 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?), personal correlation (does he refer to personal life 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, the dangers of scientific progress, and the persistence of ancient evils.

2 Research Agenda

The research agenda of this project is 'Is there a correlation between H.P. Lovecraft's personal life events and the themes in his works of that time?'. The goal is to understand how they relate to Lovecraft's personal life experiences, eventually also 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 aim to find changes in the lore of his mythos as well as correlations to the main events in his life.

We will use topic modeling on a dataset of Lovecraft's works to represent the main themes of his 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. Clarke Ashton Smith died 1961). This is part of the reason why we decided only to 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 language R, we first format the corpus into a collection of all works, then process the texts to remove annotations and unnecessary titles and use topic modeling on the corpus to gather and represent the data. After that, we will analyze the gathered data represented as a plotted diagram in R. That way we can create a timeline-like graph of the main themes and motifs in Lovecraft's works and discuss them in retrospect to events of that time or his personal life.

To properly find correlations between his writings and private life or social events of his time, we will shortly examine personal events from his letters as well as his biographical records. After that, we include the ideas and insight into the discussion. We expect to find results regarding tragic events like his mother's death or homesickness to Providence during his stay in Brooklyn, which heavily influenced his writings since Lovecraft is known for his rejection of modern movements of that time. He was known for his dislike of New York City, after visiting the city during his stay in Brooklyn. This is one of the interesting parts of his life we expect to find reflected in his writings. We already know of the story 'The Horror at Red Hook' (written 1925) in which Lovecraft negatively reflects on Red Hook, Brooklyn. It is also important to mention that Lovecraft condemned ethnic 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 aim to find similar patterns in his writing through which one could assign 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 the 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 alterations 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 occurred with other datasets, where the earliest works of Lovecraft are missing. We decided to still go with this dataset and remove works that we don't want to include. Since our research question wants to only cover Lovecraft's ideas and writings, the influence of 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 the 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 regarding his mythos and lore. After only keeping the works of Lovecraft from the bibliography, the dataset has 70 stories remaining. 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 having 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 that some works 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 to 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 represent 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 Method 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 representing 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 by using a frequentist method with the most words used, but by 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 regarding not only the frequentist quantity of certain topics but, as already mentioned, the contextual frequency. We want to know what topic dominates 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 easily and even visualize them in plots. For example, the R package ggplot2 is an easy tool to plot the results in a timeline-style plot.

We will apply topic modeling to each text. This is done because if we would use chunks of text over the whole concatenated collection, we would lose the important context. Especially in this literature of fantasy and horror, we need to keep the topics regarding 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 'place', 'sculpture', and 'ancient'. When using chunks of a concatenated collection of his writings, we would use the context of 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. A measure we use to counter this is computing a topic model using paragraphs of similar size for every work. According to the word count of Lovecraft's works the length of his works linearly increases 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 focus on topics and words that are subject to the English language and can be covered by NLP. We will also keep in mind that in Lovecraft's style, specific words tend to occur more often than others in the literary context.

5 Related Work

In this chapter, we will cover some related work that represents use cases of topic modeling similar to our research question. They analyze either Lovecraft's works or compare topics between Lovecraft's works and films and other authors.

The first related work is the paper ‘Beyond the mountain of madness: a look at the shared themes of Edgar Allan Poe and H.P. Lovecraft’ by Kristoffer Gustafson. This paper compares the main themes of both H.P. Lovecraft and Edgar Allan Poe, another well-known author, and poet. Poe heavily influenced Lovecraft’s fiction and literary style, as he stated in his essay ‘Supernatural Horror in Literature’. The author examines the themes of insanity, death, and the gothic setting on both authors and projects similarities that suggest an influence of Poe on Lovecraft.

Another paper is ‘The Lovecraft Look: An Examination of Lovecraftian Themes in Film’ by Michael A. Church. The author analyzes the philosophical beliefs and life experiences that inspired Lovecraft’s ”weird fiction” and his literary philosophy of ”Cosmic Indifferentism” and compares them with films influenced by Lovecraft. Lovecraft’s philosophy has been heavily misconstrued, as evidenced by several films that purport to adapt his stories, but actually ignore or misinterpret Cosmic Indifferentism. However, some films successfully adhere to Lovecraft’s focus on cosmic horror and humanity’s insignificance, even if they are not direct adaptations of his work. The relevant task of uncovering topics gets applied in this project to compare Lovecraft’s works and these films.

The third related work, ‘Re-visioning Romantic-Era Gothicism: An Introduction to Key Works and Themes in the Study of H.P. Lovecraft’ by Philip Smith, is similar to our research question. The author examines the recurring themes of language, genre, literary influences, xenophobia, cosmic indifferentism, dreams, time, and the influence of Lovecraft. Contrary to our work, this paper focuses on the criticism of Lovecraft. As we said in our research agenda, Lovecraft was known for racist slurs and regarding our modern society, these have to be reflected. The paper does not cover all of Lovecraft’s work and the key topics to be uncovered correlate to major critical responses.

6 Experiment Design

We applied the computational approach of topic modeling by using the R programming language and orient on an approach provided by A. Niekler. The R scripts as well as the optimized dataset, stopwords, and a simple dictionary can all be found in our repository on GitHub. The file dict.txt contains a base form of English vocabulary with stemmed word forms used to stem words and get their original lemmatization. For the processing of data, we also removed stopwords (found in stopwords.txt) since they tend to occur as noise in the retrieved topics. Furthermore, we removed some words in the preprocessing step that occurs often in Lovecraft’s works but do not really represent specifically related motifs for each text. For this project, we defined the terms and names to drop because of frequency in top_words.txt after reviewing the word count and first iterations of the model. We dropped unimportant names as well since they tend to occur more frequently than other words and did not yield any more information in the first iterations. One name we did keep for this experiment is ‘Carter’. That’s because Lovecraft maybe used the character Randolph Carter as his alter ego in the stories since they shared many personality traits. We also kept some important names of the mythos, like ‘Gilman’ or ‘Charles’.

After preprocessing the data, we calculated the topic model. For the topic model calculation, we used a Latent Dirichlet Allocation model using the R package topicmodels. LDA (Latent Dirichlet Allocation) is a statistical model used for topic modeling in natural language processing. This model assumes that documents are generated from a mixture of topics, and each topic is represented as a distribution over words. In the calculation process, we get the topic model by only considering terms with a certain minimum frequency in the body of F=3. This is to reduce the overhead of topics that will certainly not be valuable at all and can already be dropped in this step. The next step included the already mentioned drop of domain-specific words. For LDA models, the number of topics is the most important parameter to define. If K is too small, the collection is divided into a few very general semantic contexts. If K is too large, the collection is divided into too many topics of which some may overlap and others are hardly interpretable. After consulting on 'Determining the Number of Topics to Retain using Tools from Factor Analysis' by Homles Finch, we decided to choose K=15 topics for our purpose. To facilitate a qualitative control between the retrieved topics and original texts, we worked with two corpus objects in this step. One is the preprocessed corpus to calculate the topic model overall and the original corpus. For the parameters of R packages like topicmodels used in the R script files, we varied them during the experiment process several times and stuck with the final version in our repository since they reflect the best use for our research question. The parameters are: a minimum of 6 counts per token in the corpus, 100,000 iterations, 5 terms per topic, verbose of 10, and alpha of 0.2. Other parameter options did not satisfy our purpose, were too short to give enough detail, or too long to compute without further advancements.

After calculating the topic model, the details got visualized as results in different forms like a plot representing a timeline using the ggplot package in R or using the LDAvis package in R to gain more information on topics. LDAvis calculates the importance of a topic by determining how much probability is assigned to the topics. We can also filter the topics in advance, e.g. to check which terms occur more frequently in a topic. With these results, we then picked the main themes of Lovecraft's works and compared them according to time and events in his personal life in the chapter 'Results and Discussion'.

7 Results and Discussion

For the results of this Project, we represented the retrieved topic model in plots. For one we used a simple plot package, ggplot in R, to represent every of Lovecraft's works in chronological order with their distribution of the topics. We also used the R package LDAvis to show correlations between the topics and use an interactive interface to see the importance of relevant terms with respect to a topic. The resulting topics are as follows:

1. thing, time, feel, mind, life
2. thing, horror, eye, black, sound
3. dream, city, day, night, strange
4. room, door, hand, floor, window
5. place, wall, stone, vast, foot
6. west, specimen, work, lake, body
7. letter, curwen, time, late, charles
8. street, house, hill, place, town
9. family, child, son, home, time

10. thing, folk, time, good, obed
11. voice, sound, begin, word, time
12. tomb, night, fear, thing, remain
13. land, city, sea, dream, water
14. thing, night, place, talk, people
15. carter, ghoul, cat, night-gaunts, ship

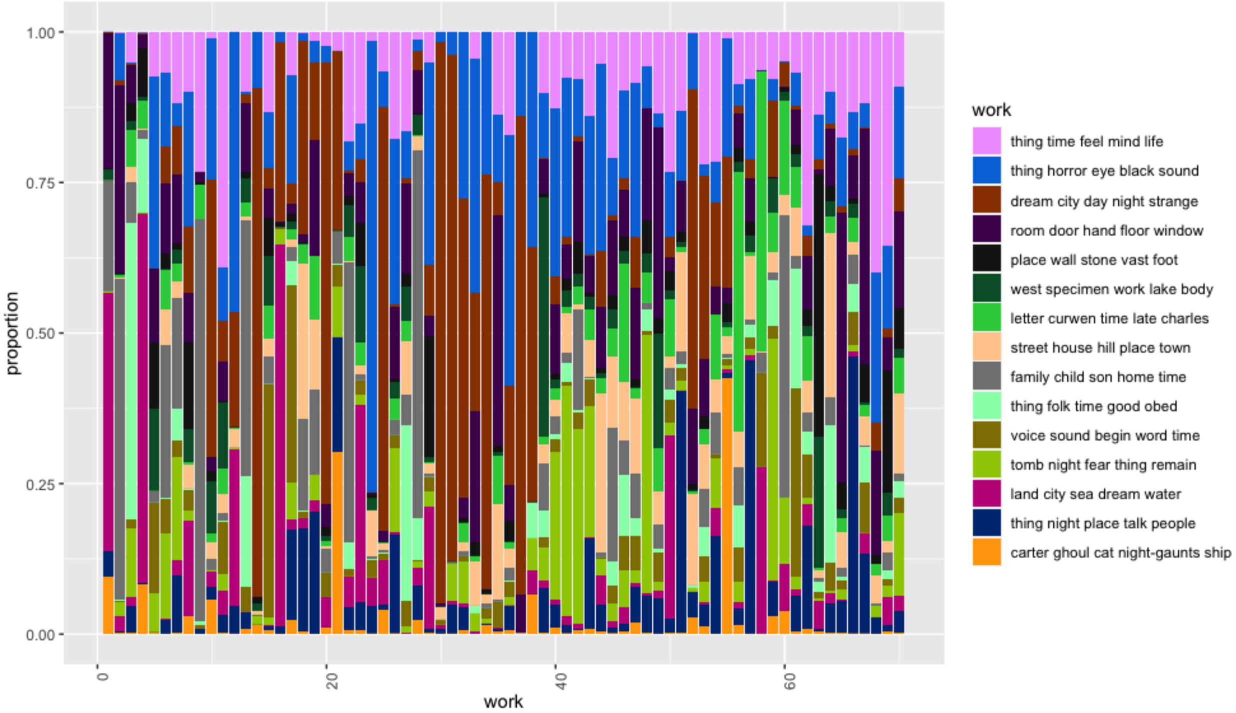

Figure 1: The plot assigning the proportion of a topic to each story of Lovecraft

Figure 1 shows the bar chart we created to plot every of Lovecraft's stories in chronological order with the percentage of the topic they share. On the x-axis, you can see the number of the story. Keep in mind that we used the date Lovecraft wrote the story and changed unspecified dates like just a year to the January of that year. For the enumeration of his works to compare to this figure, see stories_list.txt. On the y-axis, you can see the distribution in percentage. Each bar corresponds to one work and the percentage of a topic across each of the topics in that work. The order for each topic per entry is the same, it's not ordered by percentage but by topic. On the right-hand side, you can find the colors for each topic, pink being topic 01 consisting of the words 'thing, time, feel, mind, life'.

The bar chart clearly shows that some topics dominate a story and others have a fairly distributed amount of topics. For example, the first topic in pink is highly present in some works. For work 55, 'The Dream Quest of Unknown Kadath' (written 1927) one can see that topic 15 is highly present, which satisfies the expectation because it is a fairly long work with a unique setting and motif. Another example is work 63, 'At the Mountains of Madness' (written 1931), the topics 5 and 6 are more present than in any other work, which again satisfies the expectation for this work being unique in the antarctic setting and length. The bar chart also shows batches of topics for several time frames, like his works 30 to 38 which are dominated by topic 3. After reviewing the topic distribution for every work, we were very satisfied with the results and decided to use this resulting topic model for further examination.

To put this into context with Lovecraft's personal life, we took several events that may have had a significant influence on him and checked for the topics of that time. We reviewed 'An Epicure in the Terrible' (Edited by David E. Schultz and S.T. Joshi) which is a perfect collection of essays by different authors regarding Lovecraft's life. Topic 2 being blue clearly shows the time Lovecraft began to write the 'Dream Cycle', since the topics correlate well with the stories. For work 10, 'Polaris' (written 1918), and work 14, 'The White Ship' (written 1919), one can see the high share of topic 3, which just continues for later stories in the dream cycle as well. This is due to the fact that around this time, Lovecraft met Lord Dunsany, who Lovecraft admired and who heavily influenced his early works which led to the dream cycle. For the works 24 ('Nyarlathotep', written 1920), 26 ('From Beyond', written 1920), and 29 ('The Nameless City', written 1921) topic 2 is highly present and these works mark the first shift in Lovecraft's writings towards cosmic horror, which later caused the works of the 'Cthulhu Mythos'. Starting from work 30, 'The Quest of Iranon' (written 1921), topic 3 dominates parts of his writings. Topic 3 is a good indicator for writings covering the dream cycle, which matches the works. The first interesting correlation between Lovecraft's personal life and the graph can be seen starting from work 33, 'The Outsider' (written 1921), with topic 2 (blue) having a great share for some works and before. This may relate to the death of his mother in May 1921. Not only did he not write for a short period, but topic 2 consists of the keywords 'thing, horror, eye, black, sound' which suggests a coping to his mother's death as well as the time before may having an impact on this topic as well because she was admitted to a hospital. Lovecraft moved to Brooklyn in 1922 and married his wife, Sonia Haft Greene, which can be seen in a clear shift in topics during the time around work 39 and following, putting a break on the dream cycle and prompting topic 12 to be more present. Topic 12 ends being more present starting from work 44, 'The Festival' (written October 1923). That's interesting because he did also write just one story for almost two years during that time, which goes in hand with his marriage and time in New York. Lovecraft started writing again when in 1925 he moved to Red Hook, Brooklyn, living alone because his wife worked somewhere else. It is known that Lovecraft despised Red Hook and was negatively standing against minorities that lived there. The only observable difference is, that topics 4, 5, and 9 started to have a stable share starting from that time. This started with his first story after the break 'The Horror at Red Hook' (written 1925) which strongly suggests his cope with living there. The next observable pattern begins with work 52, 'The Strange High House in the Mist' (written 1926), which is represented with a higher amount of topic 3 being the topic that indicates the continuation of the dream cycle. This is very interesting, because Lovecraft moved back to Providence in April 1926, resulting from his growing homesickness and that he became increasingly depressed by his isolation and the masses of "foreigners" in the city. The observation of topic 3 being more present again due to the continuing dream cycle indicates this. Another interesting correlation with Lovecraft's stay in Brooklyn comes from topic 13. This topic covers terms like 'sea' and 'water' that are somewhat lacking during his time away from Providence (1922 to 1926, work 37 to 49). This comes from the fact that Lovecraft was inspired by Providence and the nearby sea, though Brooklyn was close to the sea too, it was the city that influenced Lovecraft's themes at that time. For 'The Dream Quest of Unknown Kadath' (work 55), it's probably the most important work of the dream cycle and covers the fictional character Randolph Carter, who was supposedly an alter ego of Lovecraft as mentioned earlier. The graph also suggests a lower share of the first topic during the timeframe for the works covered by that topic, work 56-61 from spring 1927 to 1930. This correlates to the fact that he divorced in 1929 and there may have been first signs of a shift in themes and motifs caused by unknown problems in his and Greene's relationship prior to the divorce. Topic 2 is present in some of the last works again as well as topic 1 having a higher share starting at work 62, 'The Thing on the Doorstep', which was written in the summer of 1930. This and the higher share of topic 4 starting from work 65 ('Dreams in the Witch House', written 1932) indicate an observable shift in topics following a break from writing because of the death of his aunt Mrs. Clark in 1932 and moving with his other aunt Mrs. Gamwell into small quarters in Providence. He was very close with aunt Mrs. Clark and her death could be represented in the shift starting at work 65 where topics 4 and 14 become highly present and topic 1 having a noticeably higher share from work 68. Another explanation for this mix of topics during the last years could be that Lovecraft had a hard time selling his longer stories and concentrated on ghostwriting and non-fiction, resulting in the fictional works he wrote being less frequent but longer and more focussed on different topics.

8 Conclusion

9 References