Ryan Folks
ETA: Final Project Report
6th of May, 2022

Exploring the Literary Works of HP Lovecraft using Text Analytics

“The oldest and strongest emotion of mankind is fear, and the oldest and strongest kind of fear is fear of the unknown.”
- Howard Phillips Lovecraft

Introduction:

HP Lovecraft was a controversial yet highly influential author of pulp fiction in the early 20th century. So influential and distinct were his stories that they, defying categorization at the time of their publishing, are now considered the foundation of a subgenre of their own: “Lovecraftian horror” (alternatively “cosmic horror”). The goal of this project is to perform a statistical text analysis of the short stories and poetry of HP Lovecraft to find whether the statistical results agree with the opinions of classical text critics and my own knowledge of the corpus, having read a significant portion of it. (note: I suggest viewing the full res images on the github repo)

Data Processing:

The contents of these text files were broken down into the OHCO levels “title/sentence/token”. This OHCO skips many large delimiters like chapter and paragraph for two reasons. First, no work of Lovecrafts is much longer than a chapter of a novel, so the title is functionally equivalent to an unordered set of chapters. Second, Lovecraft’s work has a wide variety of structure. Some works are rigorously structured poetry about cultural topics, others are prose poems about gardens filled with ghouls. Some are short stories delimited by section headers, and others are too brief to need them. Due to this, the paragraph was omitted from the OHCO because it cannot be applied to all works.

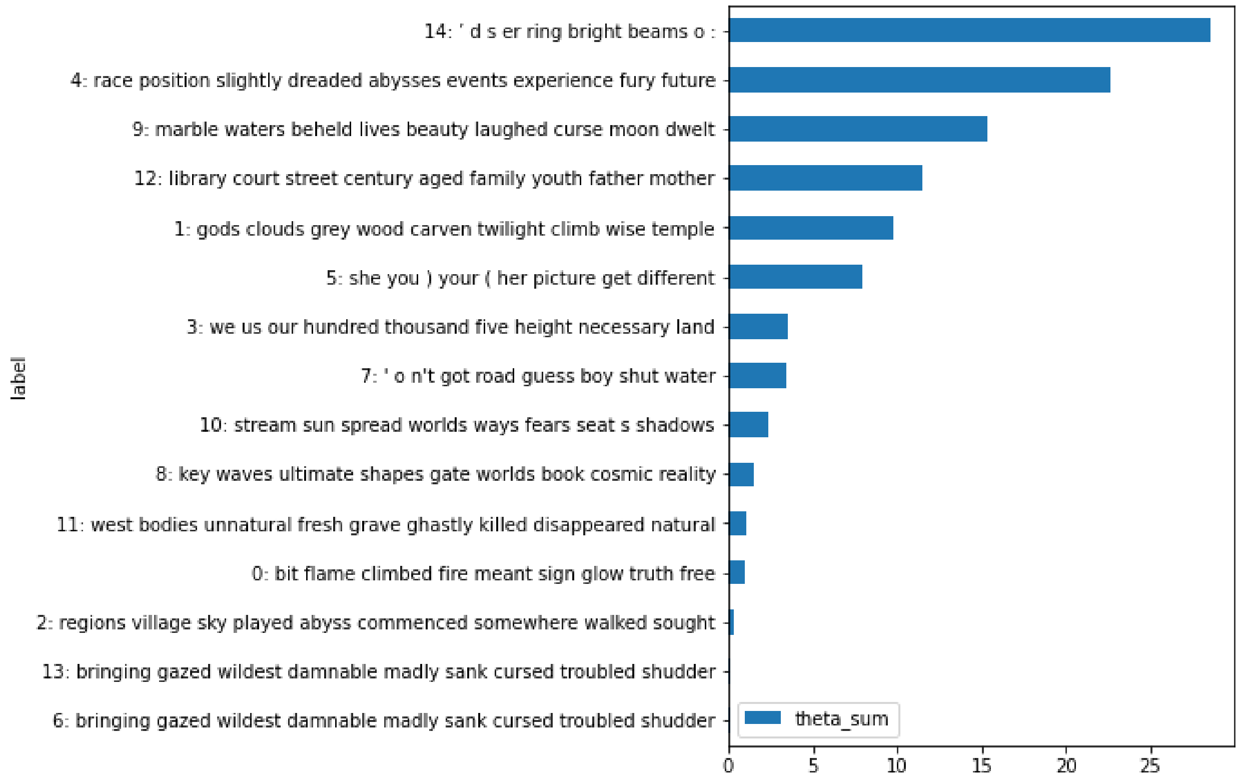
Now collected and structured, the corpus was tokenized, tagged for part of speech, stemmed via the porter, snowball, and lancaster stemmers, and used to compute the statistics zipf’s K, tfidf (using the sum method of term frequency and ‘standard’ idf), and dfidf.

Whenever graphs generated using noisy data, the savitzky-golay filter was applied. This filter interpolates a polynomial with a convolutional operation using a certain number of points before and after the target (the kernel/filter size). This filtering method doesn’t force smoothness or continuity upon the data, and avoids issues like ringing artifacts with things like the fourier transform of discrete cosine transform by not forcing these assumptions onto the data.

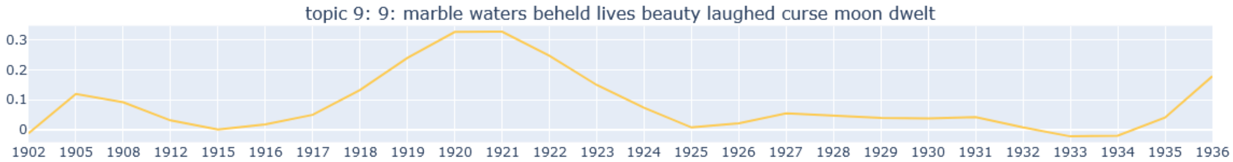
Topic Modeling:

Topic modeling is generally useful for corpora with *distinct* categories. Remarkably, nearly all of Lovecraft’s short stories were singularly focused on the same topics from the beginning of his career to the end. In a sense, he merely honed his abilities and clarified the theme of cosmic horror throughout his career.

Even still, topic modeling did unearth a few interesting things within the Lovecraft corpus which are of note. Using LDA and deciding through trial and error on fifteen topics, the resulting categories, while not very distinct, did provide some insight into slight shifts in his themes and work over time.



For example, topic nine deals with words associated with stories that feature characters stumbling upon ancient civilizations with unknown creatures (see Dagon, The Temple, Ex Oblivione). By grouping by date and taking the mean, then smoothing with a savitzky-golay filter that uses the five years before and after the target year and a fifth degree polynomial for the convolution, we find that Lovecraft was more interested in these types of stories earlier in his career.

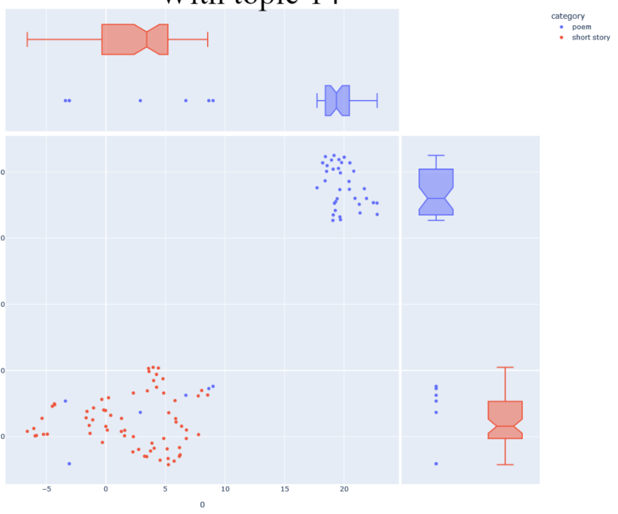


Some topics only detect single stories. Topic eleven is the Herbert West: Reanimator topic, and topic eight is the Through the Gates of the Silver Key topic. Topic three seems to heavily focus on At the Mountains of Madness, but other associated stories are roped into it as well (Hypnos, The Temple). Strangely, the poem Pacifist War Song, which has nothing to do with his other stories, is associated most strongly with this category and topic fourteen. The likely reason for this is that the word "submarines" and "old" appears in the short poem, which has direct ties to the plot of The Temple where a German u-boat breaks down and drifts into the center of an ancient underwater city with supernatural, maddening effects on the crew.

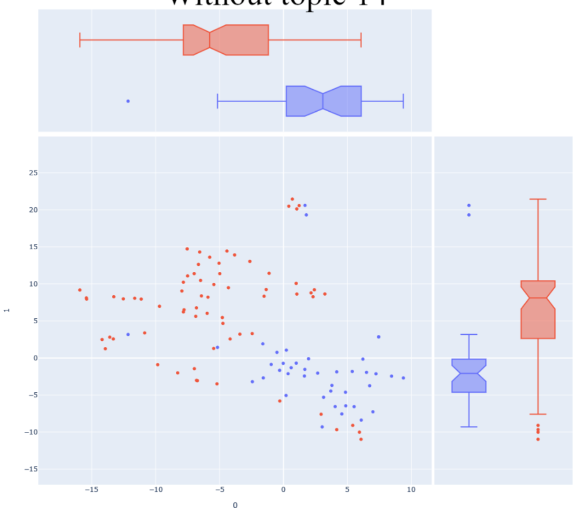
Topic fourteen is the most divisive topic in the corpus. Nearly everything associated with topic fourteen is totally associated with it such that no other topics get any probability. Topic fourteen's top words are "ring, bright, beams, o, 'd". This topic is picking out poetry, since Lovecraft used his poetry to explore other, non-horror topics and often used distinct grammar in them. For example, 'd is a suffix that rarely appears in any of Lovecraft's short stories, but one which features quite often in his poetry as a more archaic way of spelling words that end in 'er' (ex: quench'd, liv'd, learn'd). Therefore, to get a better understanding of the content rather than the style of the corpus, topic fourteen was removed from the list of topics, then the rest of the topics were adjusted such that each work still summed to one across all topics.

Using tSNE to plot a dimensionally reduced version of the topics reveals how influential topic fourteen was. When comparing the plot which includes fourteen to the one without it, we can see that it drastically separates almost all of the poetry from the short stories. The only poems which remain within the short story cluster are prose poetry that tell weird stories, and Arcadia, which is very brief and therefore is highly influenced by only a few words.

With topic 14

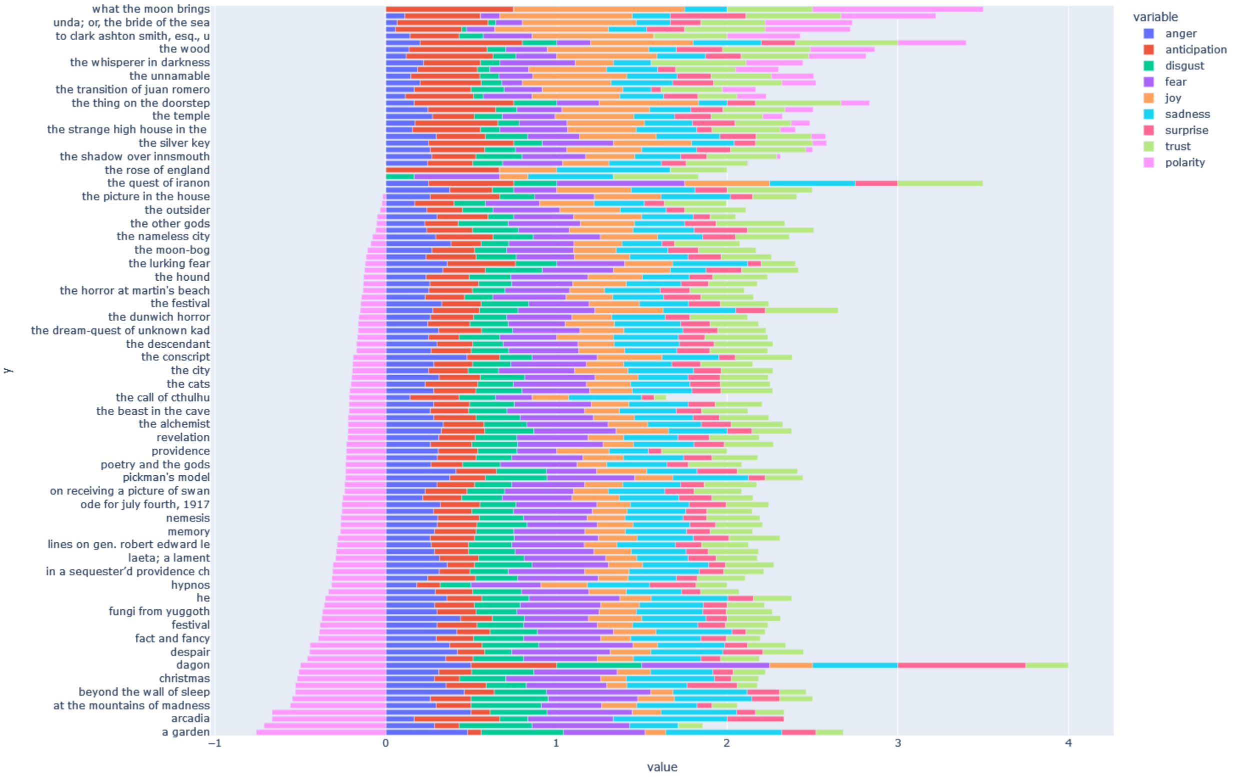


Without topic 14



Sentiment Analysis:

As previously mentioned, Lovecraft's work is almost exclusively horror. So, we should expect that the emotions contained in his works are primarily ones of fear and disgust. Lexicon based sentiment analysis methods are usually unreliable for small stories like these (as any statistical method is with low sample sizes), but they ought to be more correct than incorrect across the whole corpus. Thus, general trends that emerge from all of Lovecraft's works should be trustworthy and give useful insight. Using the NRC lexicon, sentiment analysis was performed on the whole corpus.



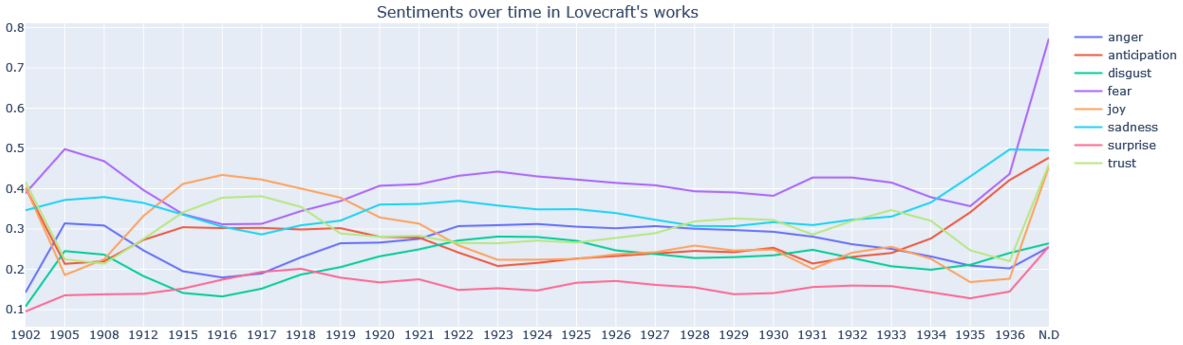
Examining the plot of all works, we see that the vast amount of them have high negative polarity; a good sign that the sentiment analysis is performing well. However, for any particular story, the sentiments may be totally wrong. For example, Christmas, a short poem Lovecraft wrote that is steeped in warm, cozy feelings, is perhaps the single most positive writing Lovecraft ever produced, yet has the fifth most negative polarity.

Title: Christmas
Author: H. P. Lovecraft
Year: 1920
+++
The cottage hearth beams warm and bright,
The candles gaily glow;
The stars emit a kinder light
Above the drifted snow.

Down from the sky a magic steals
To glad the passing year,
And belfries sing with joyous peals,
For Christmastide is here!

So, sentiment is not very useful for any individual story. However, stepping back and examining the trends in sentiment across his career, we can see some useful information emerge.

Using the Savitzky-Golay filter again, we can smooth the curve with a kernel that examines the five years before and after the target year to get a general trend line again.



Fear trends higher than any other sentiment, increasing across his career slowly as he focuses on his themes further. Sadness follows as the second highest sentiment, maintaining a steady level across all his works.

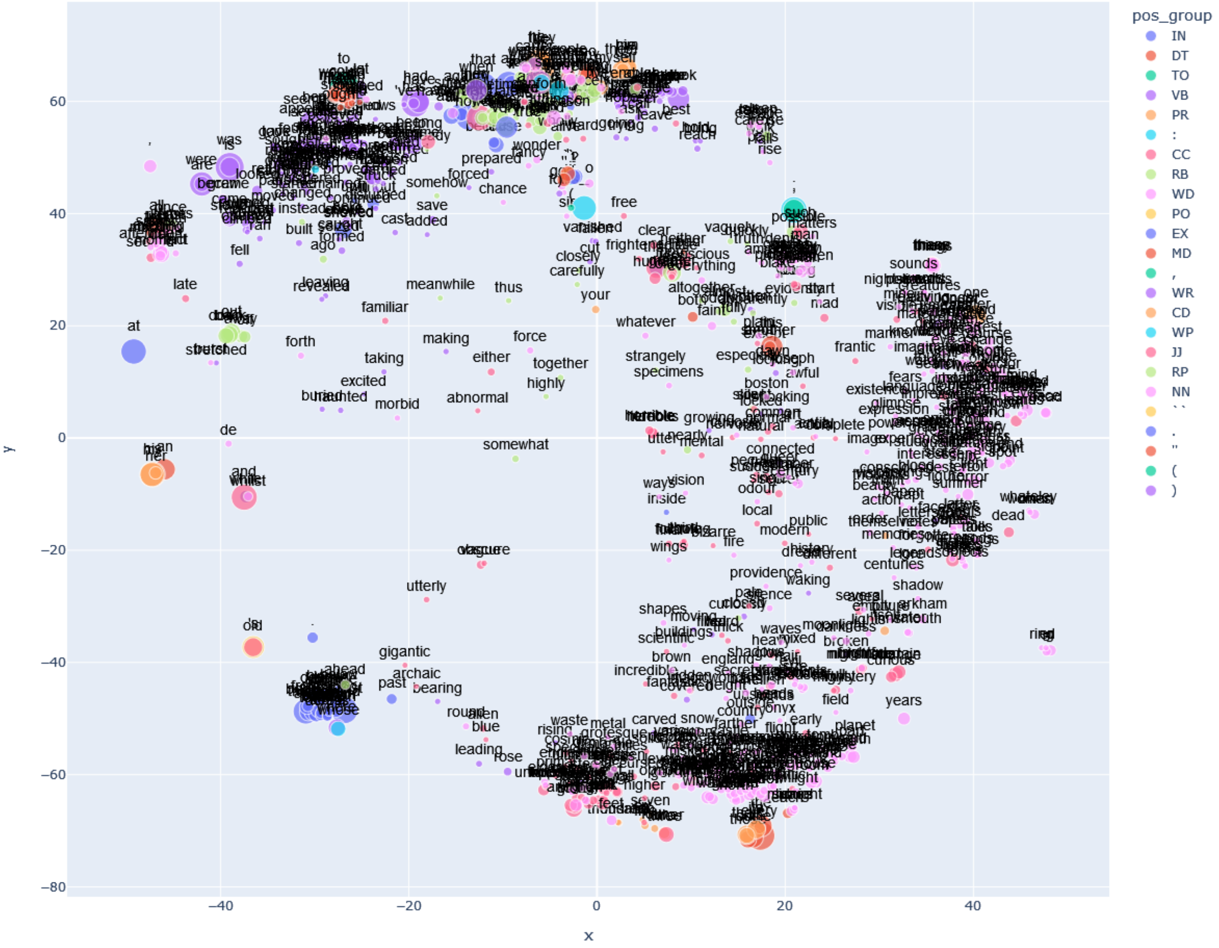
Interestingly, disgust remains low across his whole career. This is surprising, since Lovecraft text critics claim his work has a “preoccupation with visceral textures, protean semi-gelatinous substances, and slime” (Carlin). Perhaps here this critic is wrong, and Lovecraft’s horror is cleaner than we assume.

Another unexpectedly low sentiment was surprise, which ranked the lowest out of all of the emotions. Lovecraft’s earlier stories often feature one large twist to shock the reader, as was the custom of the pulp fiction community (Facts Concerning the Late Arthur Jermyn, The Beast in the Cave, Dagon). It is more likely, however, that surprise is a difficult emotion to assign to words, and that the NRC lexicon doesn’t do a good job of assigning it.

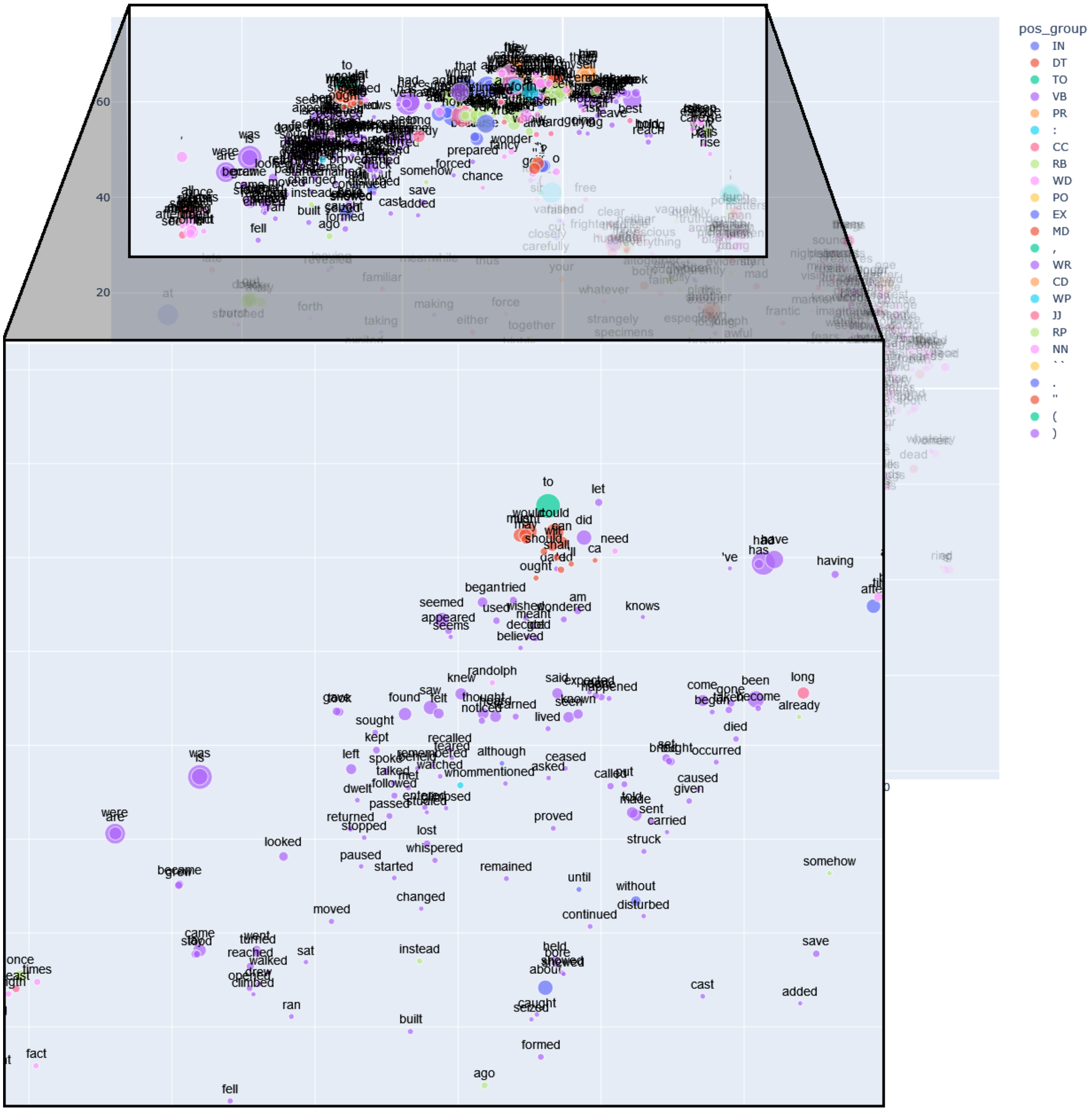
Overall, sentiment analysis was useful in confirming what we already knew about the corpus, as well as challenging the assumption that Lovecraft’s work contains words that evoke disgust, but our analysis seems to fall apart at the level of the individual story, with the most positive work of all having an erroneous negative polarity.

Word Embedding:

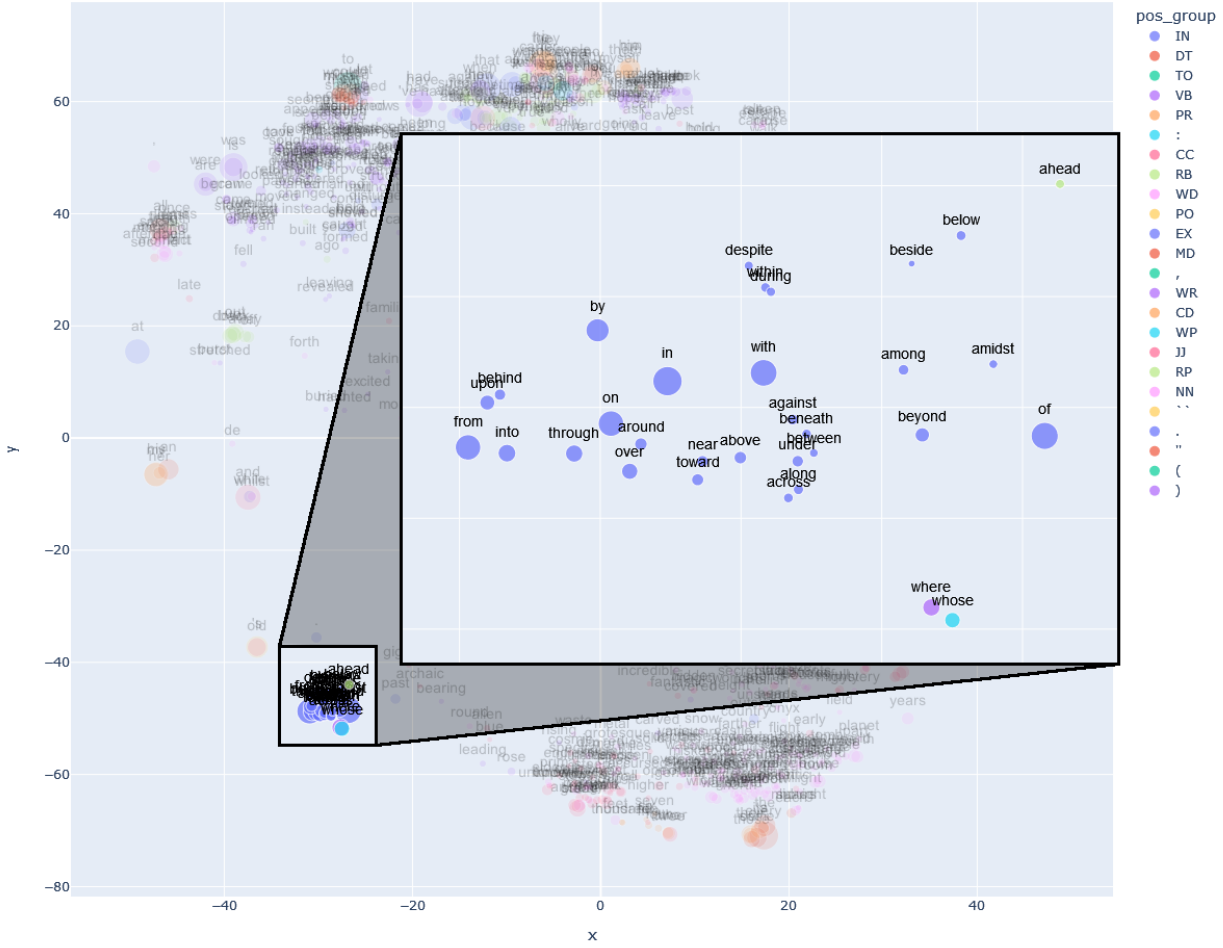
Word embedding is the most useful tool for exploring the themes of this corpus. Using word-2-vec and tSNE to plot the 1000 most significant words by dfidf, we can see a neat landscape beginning to form. This section will explore this geography piece by piece.



The first area is what I call the “functional lagoon”. Here we mostly find verbs which perform only slightly more than functional work. For example, words like “was”, “knew”, and “said” appear in the lagoon, and don’t tell us much about the texts themselves. Some words do carry more semantic meaning though. “Whispered”, “feared”, and “lived” show up alongside these more functional words, implying that they are used frequently enough to be considered roughly interchangeable. This suggests Lovecraft used the word “feared” nearly as frequently and in the same context as the word “spoke”, showing again how singularly focused his works were. More broadly, the types of words that appear here can inform us about how the stories in the corpus generally unfold. Many of the words found here suggest that the main character has little agency in the story, and is merely the subject to which the story happens. Words like “dwelt”, “paused”, “ceased”, “noticed”, and “seen” are more prominent by dfidf than more sparse active words like “moved”, “changed”, “continued” or “disturbed”. This choice of verbage corroborates the critical idea that characters in Lovecraft’s stories have very little agency. Even in stories where characters should be leading the exploration, like At the Mountains of Madness and The Temple, Lovecraft often likes to rob from them any plot device which gives them the ability to act upon the world, thrusting them suddenly into a place where they are at the mercy of forces greater than themselves, leaving them with only the ability to freeze and observe in horror.



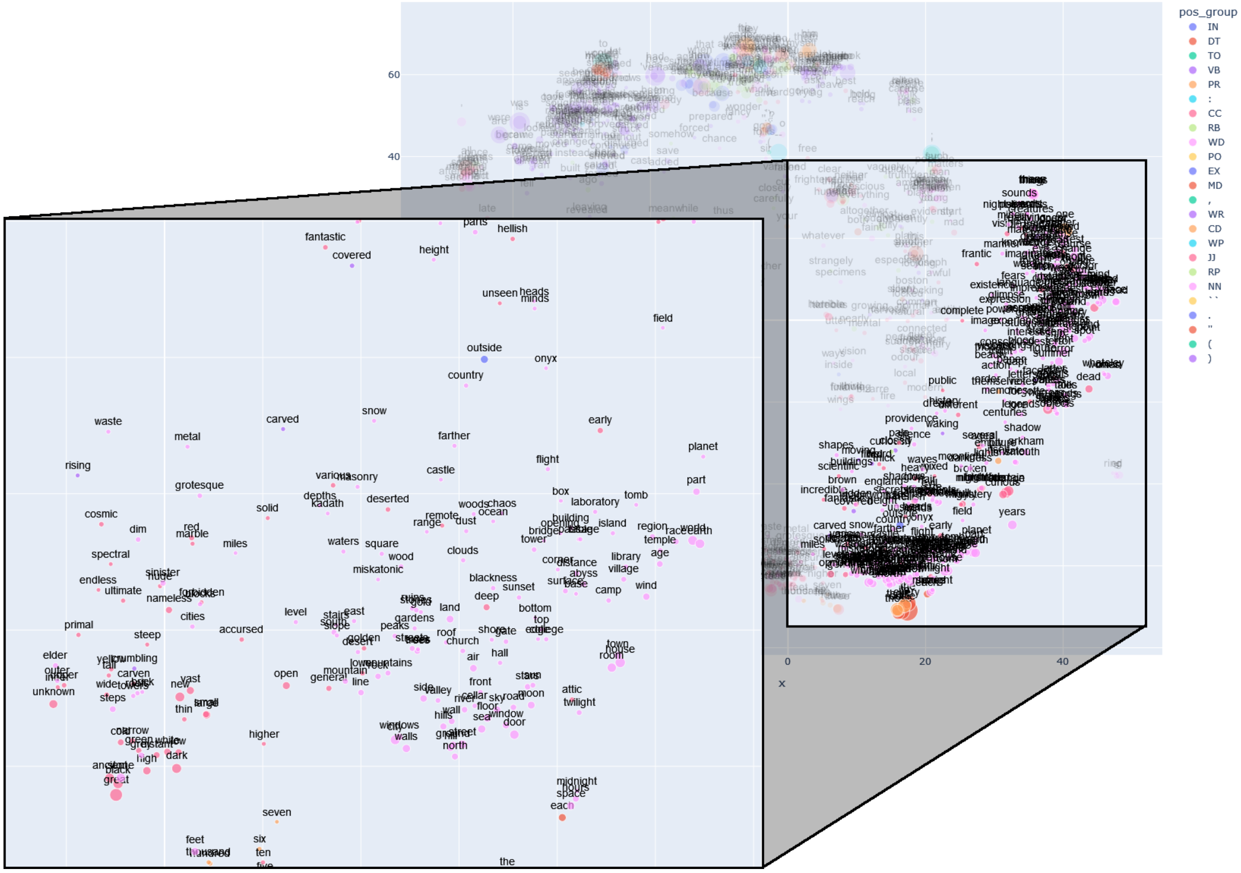
Before moving to the most important place in this landscape, one other area should be noted. There is a dense cluster of high dfidf words I am calling the positional well. These are all functional words which reference some location in space relative to another (ex: among, over, between). If Lovecraft's characters can only freeze and observe, it makes sense that their long descriptions of ancient temples and giant creatures would need a lot of help from these words. This section, combined with the functional lagoon completely support the assertion of Lovecraft text critics that protagonists are "helpless in the face of unfathomable and inescapable powers, which reduce humans from a privileged position to insignificance and incompetence."(Fredriksson)



Finally, and most importantly, we come to the "noun miasma" where the most information about the corpus can be found. Here we see what these stories are about: "cosmic", "endless", "grotesque", things which dwell in places like a "castle", the "depths", or an "abyss". Lovecraft text critics assert that his horror is based upon the "fear and awe we feel when confronted by phenomena beyond our comprehension, whose scope extends beyond the narrow field of human affairs and boasts of cosmic significance" (Ralickas), and these locations physically embody these concepts. They are literally "dark", "vast" places where the "unknown" dwells (and all these words appear prominently in the noun miasma). Interestingly though, the word "gardens" appears semantically close to all these other places. In Lovecraft's work, garden's are strongly associated with dreams, and the thin barrier between them and reality, which is likely why they appear here.

Other places which appear in the noun miasma are "churches", "stairs", "building", "towers" and "bridges". In addition to character's being thrust into strange places, strange places are often thrust upon the characters, who more often than not live in a fictitious mirror to New England (Arkham, Dunwich, Miskatonic university). The horror relied upon here is not physically embodied by the location of the story, but rather lies just underneath a facade of normality (often inhabited by foreigners that Lovecraft saw as strange and dangerous people).

Overall the noun miasma confirms the textual critic's ideas about Lovecraft's works, and these settings are still highly associated with stories of cosmic horror that have been created nearly eighty years later (the manga works of Junji Ito, the video game Bloodborne, the film Prometheus, and the collaborative creative writing collection called the SCP Foundation).



Conclusion:

To answer the question in the introduction: yes, the statistical analysis did confirm a majority of the text critic's views of Lovecraft's work. His work is mostly centered around the theme of madness and existential crisis, as shown by the word embeddings. The primary emotions in his works are those of fear and sadness which he refined across his career as shown by sentiment analysis. And, finally, although he wrote both poetry and short stories as part of his creative writing, they are detectably different using stylistic elements and thematic ones which was shown using LDA topic modeling.

References:

Carlin, Gerry; Allen, Nicola (2013). "Slime and Western Man: H. P. Lovecraft in the Time of Modernism". In Simmons, David (ed.). *New Critical Essays on H.P. Lovecraft*. Palgrave Macmillan. pp. 73–90.

Fredriksson, Erik (2010). *Hidden knowledge and Man's Place in the Universe : a study of human incompetence and insignificance in the works of H.P. Lovecraft* (Bachelor thesis). Luleå University of Technology. Retrieved 29 March 2021.

Ralickas, Vivian. "'Cosmic Horror' and the Question of the Sublime in Lovecraft." Journal of the Fantastic in the Arts 18, no. 3 (2008): 364.