Mini-project in NLP - Analysis of Howard Phillips Lovecraft's literary works

---

Authors: Ivgeniy Daskovskiy, Anton Novikov, Shir Daniel.

Introduction

Howard Phillips Lovecraft, commonly known as H.P. Lovecraft, born in 1890, was an American writer who gained posthumous fame for his influential works of horror fiction. Even non-fans of the horror genre will have heard of one of his most known titles, like "The Call of Cthulhu".
Lovecraft faced various personal struggles throughout his life, including the death of his father at a young age, the death of his mother later on, financial difficulties, social isolation and his own struggles with mental illness. It's speculated that the traumatic events he experienced may have shaped the dark and mysterious themes present in his work.
Lovecraft's writing style evolved over time, shifting from traditional Gothic horror to cosmic horror, which explores themes of psychological and existential dread, and humanity's insignificance in the universe.
Lovecraft continued writing up until his death at the age of 46.

Experiment goal

Considering Lovecraft's evolving literary output throughout his career, we would like to determine to what extent can the variations in his authorial style, be identified by statistical analysis.

Research literature

- Data
  - Wikipedia page on H.P. Lovecraft - *gave us insight into Lovecraft's life and career*
  - Works Of H.P. Lovecraft - *served as a one-stop source of our dataset*
- Tools and algorithms
  - Spacy docs - *framework documentation, code examples, source of information about nlp methods*
  - Article: NLP with spaCy in Python - *succinct guide into using the spacy framework*
  - Top2Vec - *framework documentation*
  - BERTopic - *framework documentation*

    - Gensim - *framework documentation, used for their LDA and TF-IDF implementations*
    - pyLDAvis - *framework documentation, used for visualizing LDA topics' distance*
    - Article explaining LDA - *served for understanding the algorithm*
    - Topic Modeling Coherence Score - *served as a guide on optimizing an LDA model*
    - Logistic Regression - *served for understanding the algorithm*
    - PyTorch docs - *framework documentation*
- Related research
    - A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts - *served as an insight into the differences between the different available topic-modeling frameworks*
    - Corpus Periodization Framework to Periodize a Temporally Ordered Text Corpus - *served as an example and inspiration*
    - Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts - *served as an example and inspiration*

Examining the data

For our research topic, it seemed most relevant to us to use time as the basis for comparing text features.

Our dataset includes a total of 63 of Lovecraft's works, spread over the years 1917-1935:

Year	Stories
1917	"The Tomb", "Dagon"
1918	"Polaris"
1919	"Beyond the Wall of Sleep", "Memory", "Old Bugs", "The Transition of Juan Romero", "The White Ship", "The Doom That Came to Sarnath"
1920	"The Statement of Randolph Carter", "The Terrible Old Man", "The Tree", "The Cats of Ulthar", "The Temple", "Facts Concerning the Late Arthur Jermyn and His Family", "The Street", "CelephaTs", "From Beyond", "Nyarlathotep", "The Picture in the House"
1921	"Ex Oblivione", "The Nameless City", "The Quest of Iranon", "The Moon-Bog", "The Outsider", "The Other Gods", "The Music of Erich Zann"
1922	"Herbert West — Reanimator", "Hypnos", "What the Moon Brings", "Azathoth", "The Hound", "The Lurking Fear"
1923	"The Rats in the Walls", "The Unnamable", "The Festival"
1924	"The Shunned House"
1925	"The Horror at Red Hook", "He", "In the Vault"

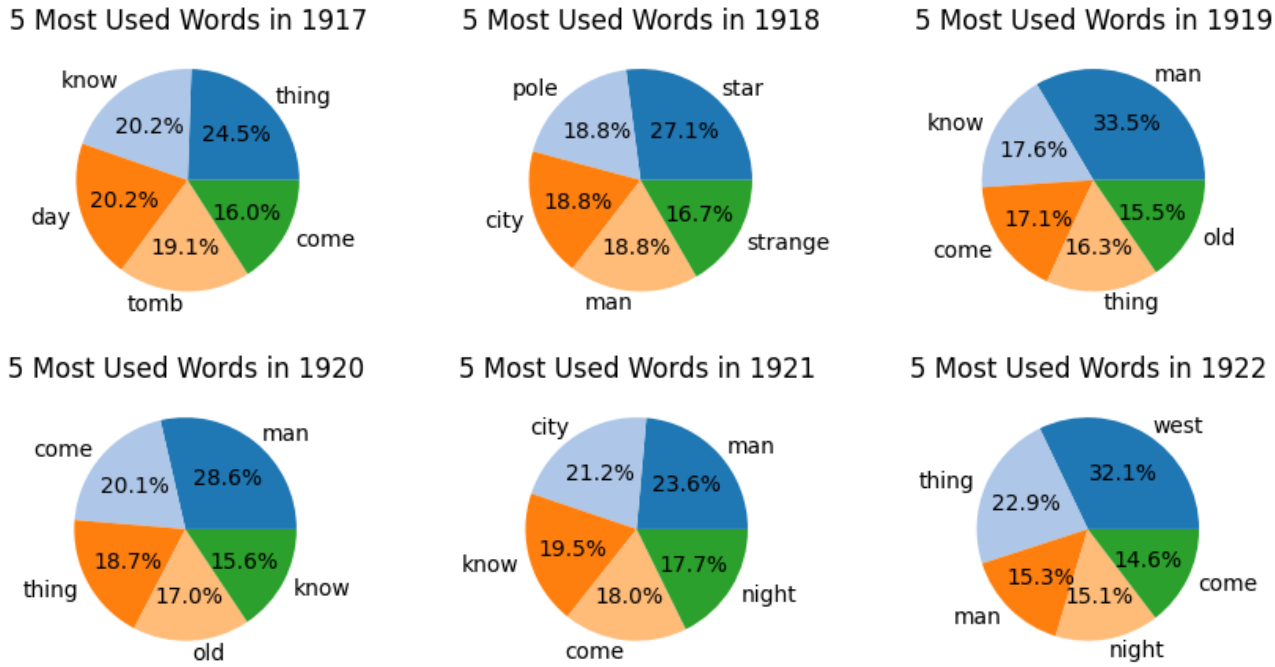
Year	Stories
1926	"The Descendant", "Cool Air", "The Call of Cthulhu", "Pickman's Model", "The Silver Key", "The Strange High House in the Mist"
1927	"The Dream-Quest of Unknown Kadath", "The Case of Charles Dexter Ward", "The Colour Out of Space", "The Very Old Folk", "The Thing in the Moonlight", "The History of the Necronomicon"
1928	"Ibid", "The Dunwich Horror"
1930	"The Whisperer in Darkness"
1931	"At the Mountains of Madness", "The Shadow Over Innsmouth"
1932	"The Dreams in the Witch House"
1933	"The Thing on the Doorstep", "The Evil Clergyman", "The Book"
1934	"The Shadow Out of Time"
1935	"The Haunter of the Dark"

Shallow statistical analysis

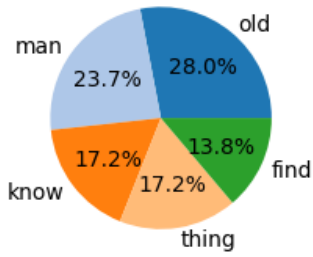
We tokenized our texts, and added bigrams and trigrams tokens.

Using the `spacy` framework we parsed the text to gather some initial yearly statistical information. Like top words used by year, top bigrams/trigrams by year and top parts-of-speech by year.

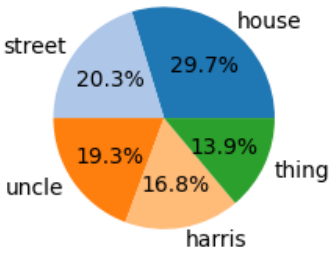
Words



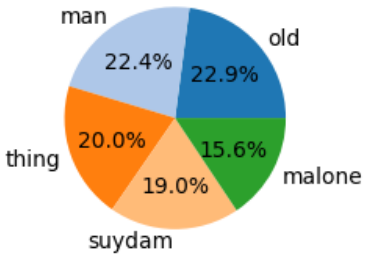
5 Most Used Words in 1923



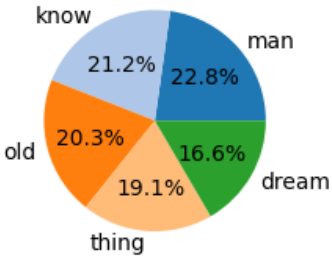
5 Most Used Words in 1924



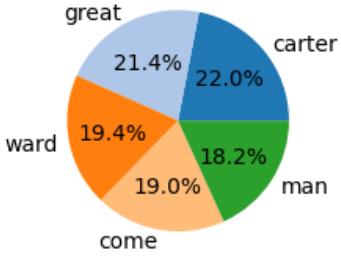
5 Most Used Words in 1925



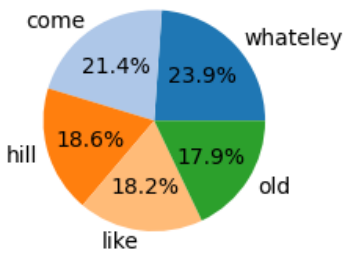
5 Most Used Words in 1926



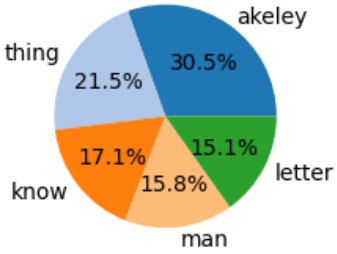
5 Most Used Words in 1927



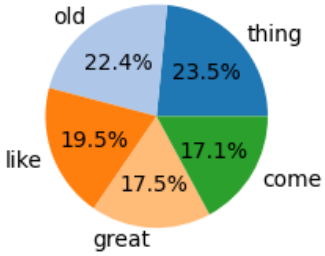
5 Most Used Words in 1928



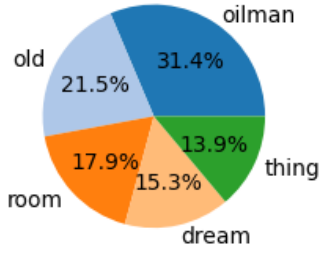
5 Most Used Words in 1930



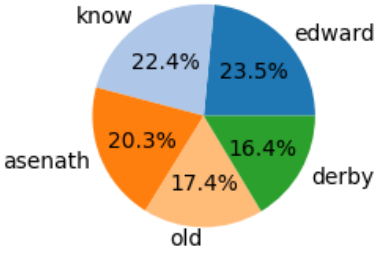
5 Most Used Words in 1931



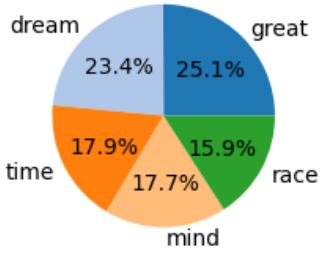
5 Most Used Words in 1932



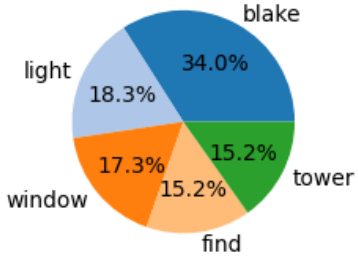
5 Most Used Words in 1933



5 Most Used Words in 1934

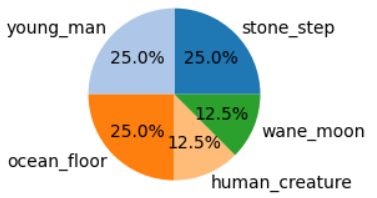


5 Most Used Words in 1935

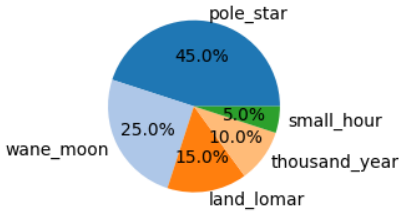


Bigrams & trigrams

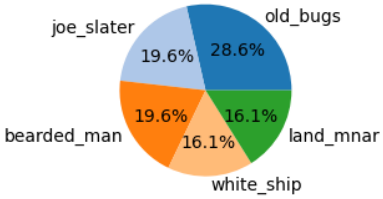
5 Most Used Bigrams/Trigrams in 1917



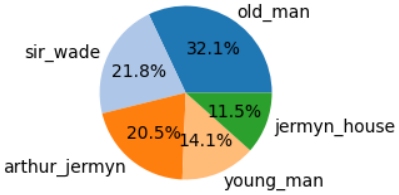
5 Most Used Bigrams/Trigrams in 1918



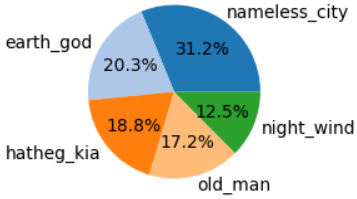
5 Most Used Bigrams/Trigrams in 1919



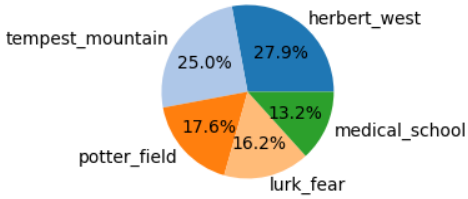
5 Most Used Bigrams/Trigrams in 1920

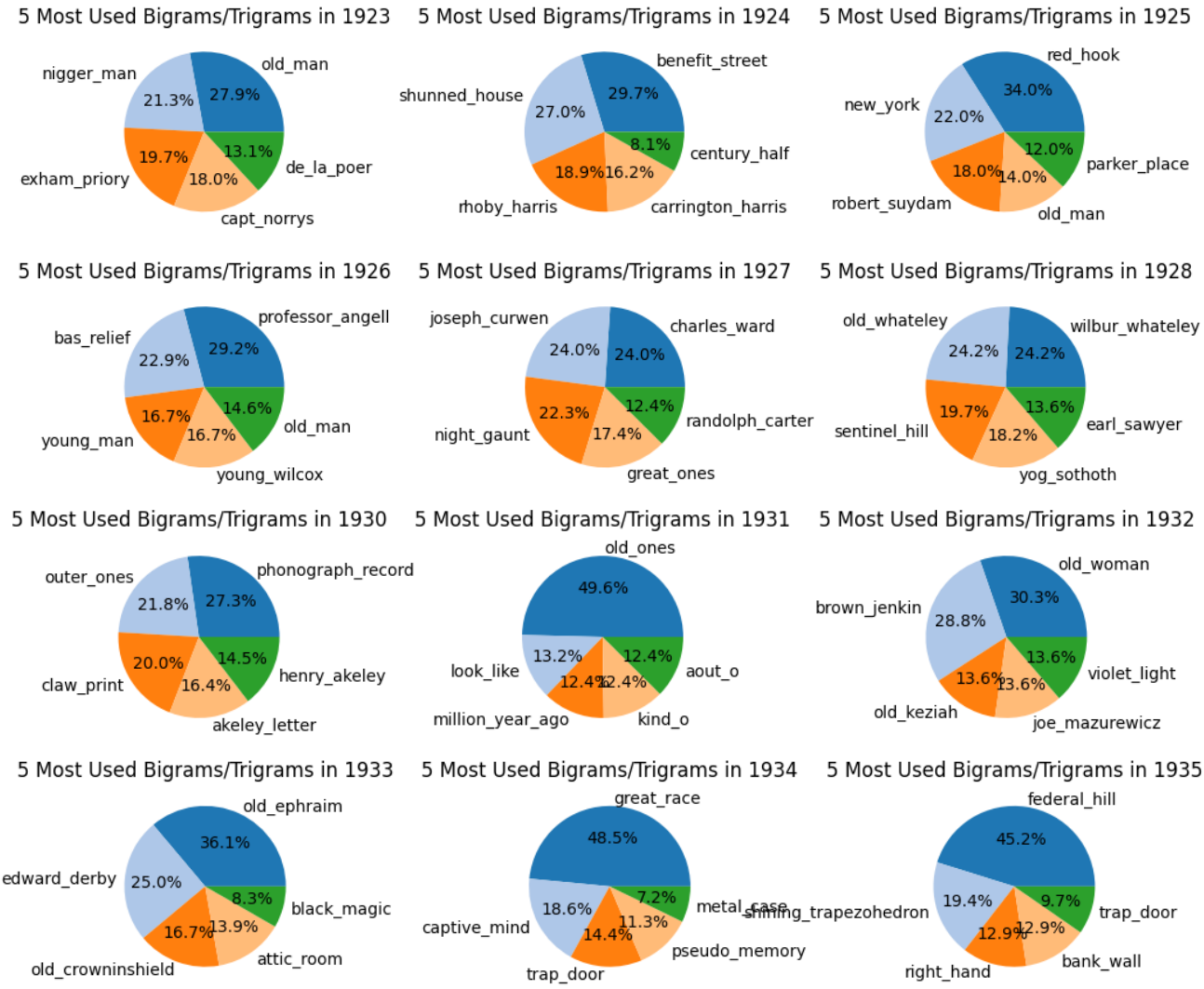


5 Most Used Bigrams/Trigrams in 1921

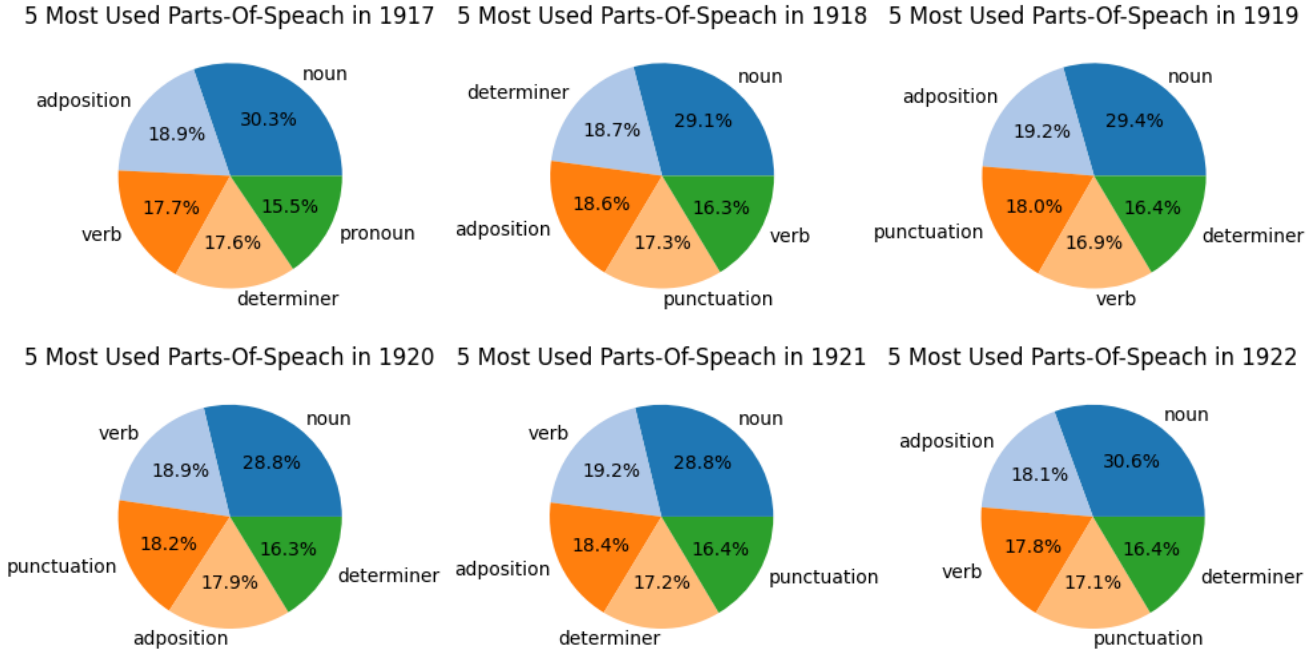


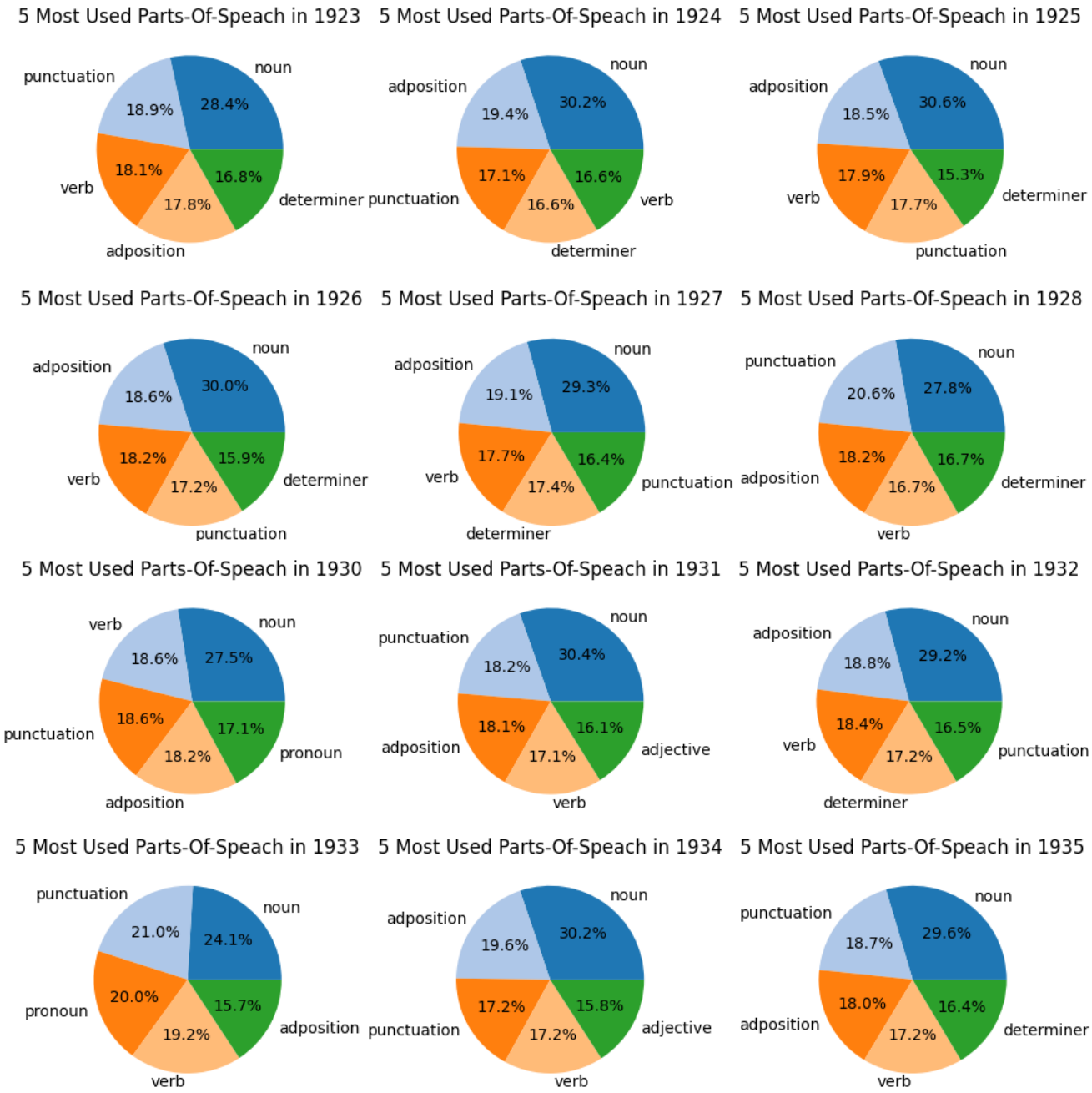
5 Most Used Bigrams/Trigrams in 1922





Parts-of-speech





Some observable trends

- Words:
  - "man" appears in top-5 in almost every year from 1918-1927. It seems to coincide with the fact that his stories frequently dealt with experiences of isolation/solitary experiences.
  - "old" is recurring in top-5 over the years 1931-1933. It's possible that it had something to do with the fact that Lovecraft was ill at the later stages of his life (an illness that would later claim his life) and weaved his concerns into his stories.
  - "come" is recurring from 1919-1922. Coincides with the recurring themes of travel/journey in his stories.
  - "thing" makes inconsistent appearances throughout the years. Probably stems from the super-natural themes of his work.
- Bigrams/Trigrams:

    - It's noticeable that unlike in top-5 words' charts, here the spread over the phrases is less even. In some years, there are phrases that dominate at almost 50% usage.
    - From contrast between phrases such as "great ones", "million year ago", "earth god", "century half", "thousand year", "tempest mountain" and phrases such as "nameless city", "small hour", "old man", "old woman", "nigger man", "old Ephraim", "wane moon" we can notice the marks of Lovecraft's themes of the grand vs. the insignificant, on the texts.
- Parts-of-speech: the only significant trend that we can observe is that the usage of nouns in his work always prevailed over other parts-of-speech. There isn't any other observable trend binding a period into one.

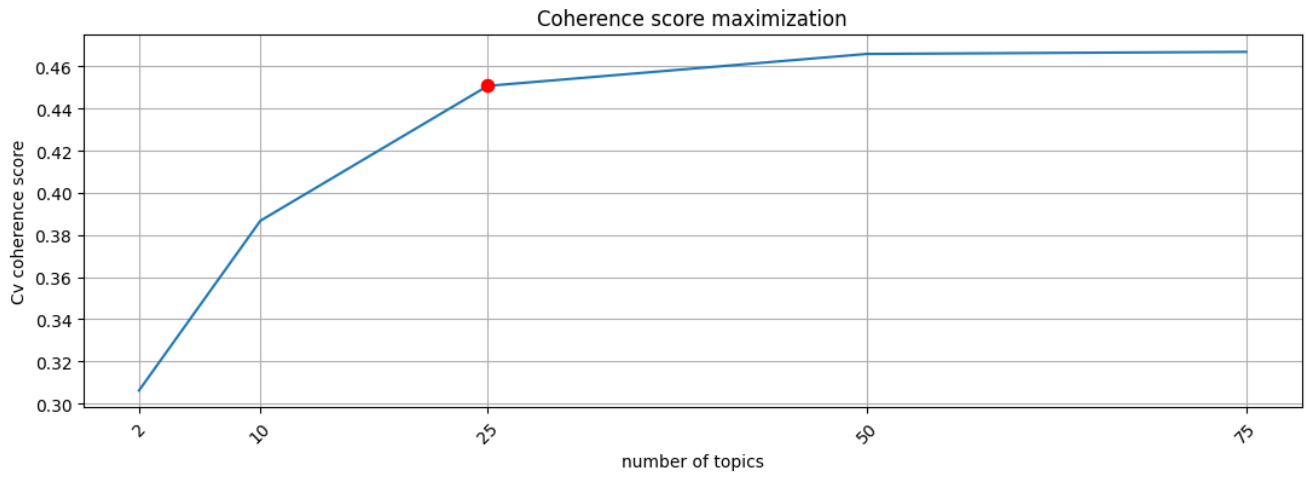
Topic Modeling

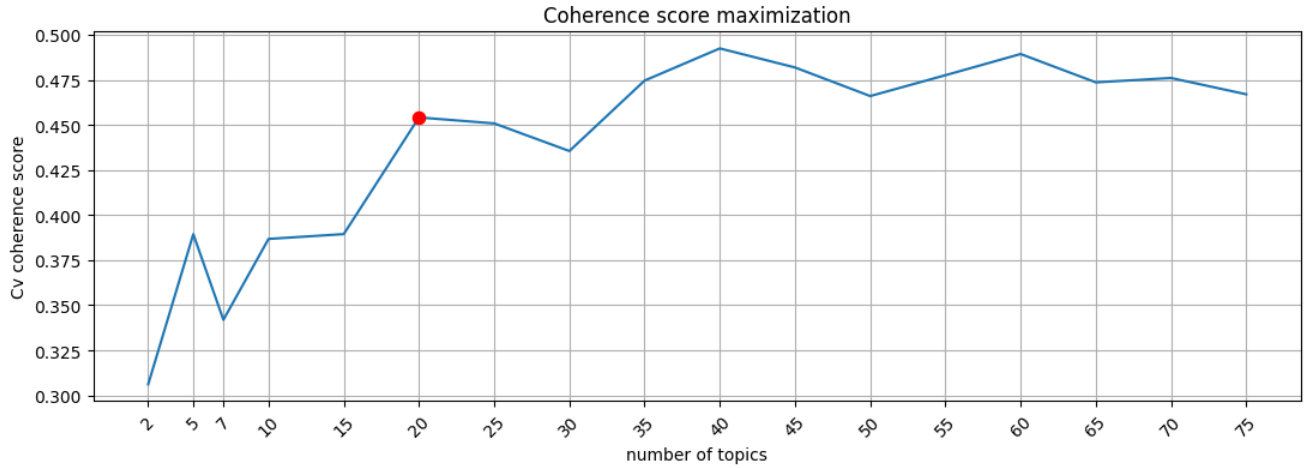
Initially, we planned on using more modern topic modeling tools, like `Top2Vec` or `BERTopic` which make use of transformer-encodings in their training process, don't require prior text-tokenization, and provide richer topic-related information. Unfortunately employing these tools didn't work for our dataset as it is too small for them to be able to extract information from it.

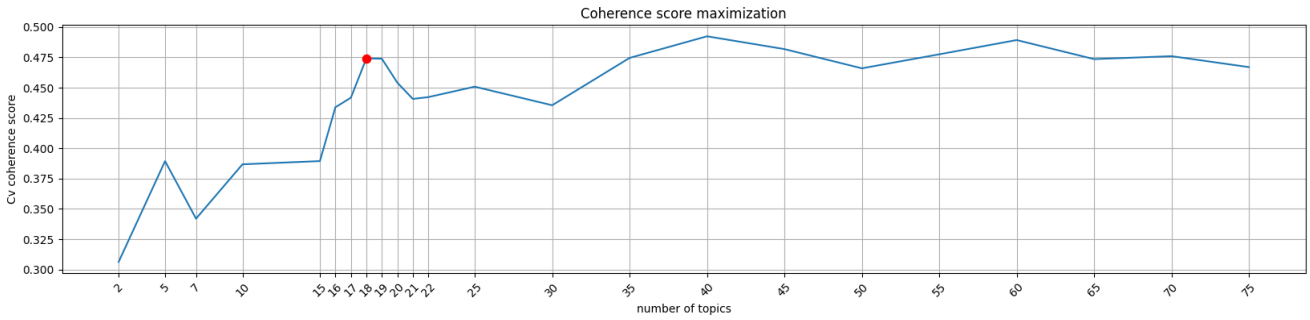
Thus we settled for a more traditional approach to topic modeling - with LDA.

To get better separation of topics, we discarded tokens occurring too frequently across the corpus. To get better topic coherency we discarded infrequent tokens per document. We then transformed the tokenized documents into "bags-of-words", and trained an LDA model (from `gensim` python package) on the resulting corpus.

In choosing the number of topics for the model to detect, we tried to maximize topic-coherency. We plotted the C_v coherence score as function of an increasing number of topics and used the elbow rule to determine what number's best to pick.

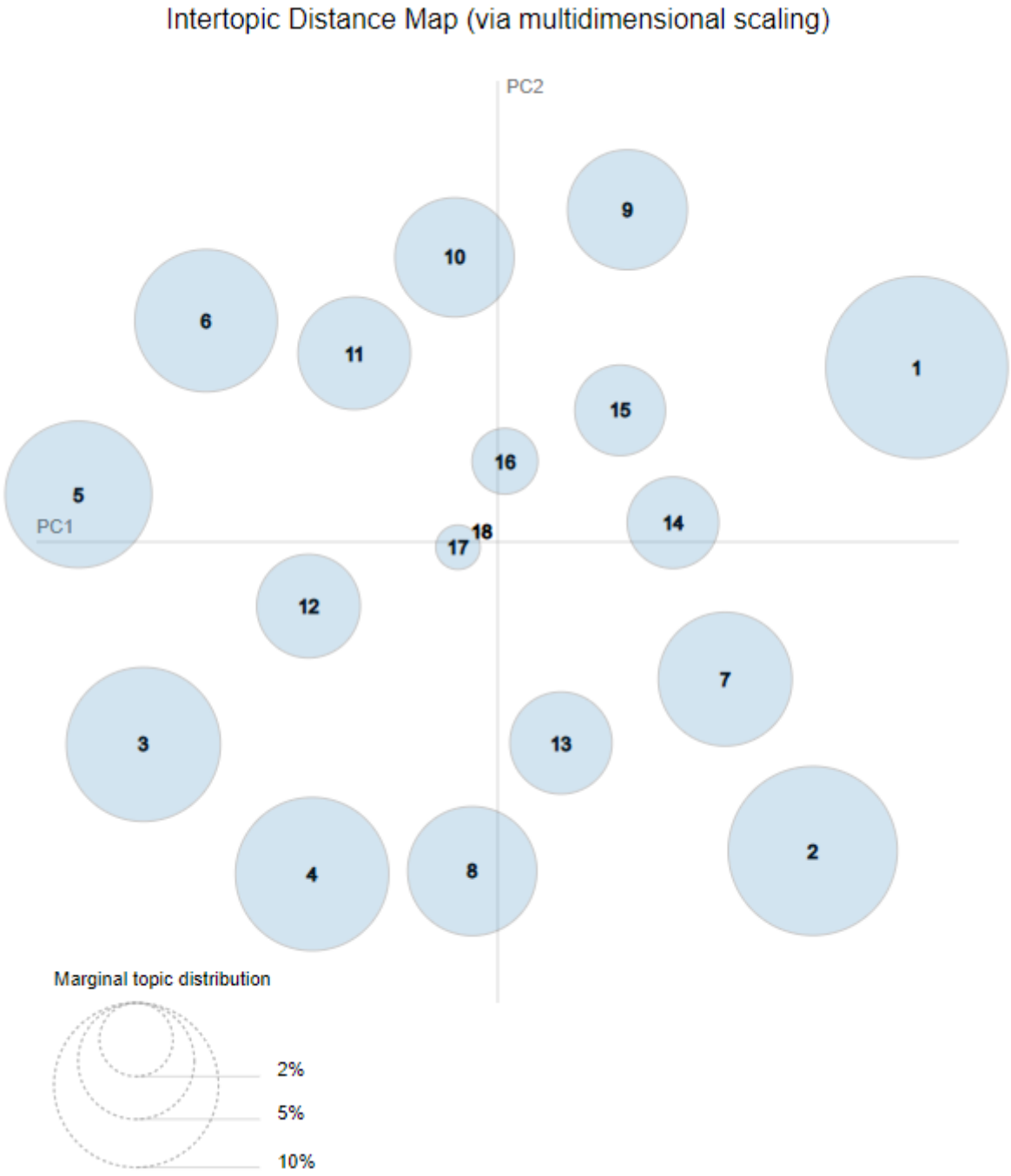






We've chosen to extract 18 topics.

To visualize the resulting topics' separation we used `pyLDAvis`'s integration with `gensim`'s models.



As we can see from the visualization of topic embeddings, they have little-to-no overlap between them. Which should indicate successful generation of corpus and good choice of number of topics.

The following topics were detected ("raw" format):

Topic	Distribution
0	0.000*"dream" + 0.000*"sea" + 0.000*"carter" + 0.000*"street" + 0.000*"west" + 0.000*"innsmouth" + 0.000*"specimen" + 0.000*"case" + 0.000*"arkham" + 0.000*"jermyn" + 0.000*"ship" + 0.000*"body" + 0.000*"lake" + 0.000*"o" + 0.000*"temple" + 0.000*"uncle" + 0.000*"cult" + 0.000*"land" + 0.000*"tower" + 0.000*"town"

Topic	Distribution
1	0.021*"cat" + 0.009*"ulthar" + 0.006*"wife" + 0.006*"villager" + 0.005*"caravan" + 0.005*"wanderer" + 0.004*"boy" + 0.004*"cottage" + 0.004*"yard" + 0.003*"old_man" + 0.003*"folk" + 0.003*"prayer" + 0.003*"cotter" + 0.002*"horn" + 0.002*"kitten" + 0.002*"lean" + 0.002*"slay" + 0.002*"kranon" + 0.002*"black_kitten" + 0.002*"burgess"
2	0.013*"nameless_city" + 0.010*"passage" + 0.009*"stone" + 0.009*"temple" + 0.009*"low" + 0.009*"sand" + 0.009*"desert" + 0.009*"city" + 0.007*"race" + 0.007*"crawl" + 0.007*"wind" + 0.007*"shew" + 0.006*"moon" + 0.006*"corridor" + 0.006*"torch" + 0.006*"reptile" + 0.005*"abyss" + 0.005*"valley" + 0.005*"ruin" + 0.005*"rock"
3	0.017*"iranon" + 0.013*"city" + 0.013*"aira" + 0.010*"song" + 0.010*"sing" + 0.008*"slater" + 0.007*"teloth" + 0.007*"thou" + 0.007*"sleep" + 0.006*"dream" + 0.006*"romnod" + 0.006*"nyarlathotep" + 0.005*"joe_slater" + 0.005*"golden" + 0.005*"shall" + 0.005*"marble" + 0.005*"beauty" + 0.005*"oonai" + 0.004*"toil" + 0.004*"thy"
4	0.037*"dream" + 0.020*"case" + 0.019*"mind" + 0.017*"great_race" + 0.015*"age" + 0.012*"block" + 0.011*"vision" + 0.010*"study" + 0.009*"sand" + 0.009*"masonry" + 0.009*"wholly" + 0.009*"future" + 0.009*"impression" + 0.008*"book" + 0.008*"course" + 0.008*"torch" + 0.008*"myth" + 0.008*"race" + 0.008*"knowledge" + 0.008*"legend"
5	0.064*"carter" + 0.027*"ghoul" + 0.024*"god" + 0.015*"cat" + 0.013*"land" + 0.013*"cloud" + 0.012*"night_gaunt" + 0.012*"ship" + 0.012*"rock" + 0.011*"mountain" + 0.011*"peak" + 0.011*"cliff" + 0.011*"tower" + 0.009*"great_ones" + 0.009*"galley" + 0.009*"onyx" + 0.009*"merchant" + 0.008*"inganok" + 0.008*"ngranek" + 0.008*"ulthar"
6	0.029*"street" + 0.019*"innsmouth" + 0.013*"o" + 0.012*"town" + 0.011*"sea" + 0.011*"people" + 0.011*"folk" + 0.010*"obed" + 0.009*"water" + 0.009*"church" + 0.009*"fish" + 0.008*"malone" + 0.008*"marsh" + 0.008*"suydam" + 0.008*"arkham" + 0.007*"cross" + 0.006*"fer" + 0.006*"reef" + 0.006*"bad" + 0.006*"island"
7	0.022*"west" + 0.022*"lake" + 0.018*"city" + 0.016*"specimen" + 0.014*"foot" + 0.014*"land" + 0.013*"danforth" + 0.012*"old_ones" + 0.012*"mountain" + 0.012*"course" + 0.012*"camp" + 0.011*"sculpture" + 0.011*"point" + 0.011*"plane" + 0.009*"ice" + 0.009*"antarctic" + 0.009*"sea" + 0.009*"body" + 0.007*"arkham" + 0.007*"dog"
8	0.028*"dream" + 0.019*"cult" + 0.013*"uncle" + 0.013*"johansen" + 0.010*"star" + 0.010*"note" + 0.009*"professor" + 0.008*"legrasse" + 0.008*"wilcox" + 0.007*"swamp" + 0.007*"manuscript" + 0.007*"professor_angell" + 0.006*"sea" + 0.006*"image" + 0.006*"cthulhu" + 0.006*"march" + 0.005*"alert" + 0.005*"bas_relief" + 0.005*"sculptor" + 0.005*"april"
9	0.012*"dream" + 0.008*"carter" + 0.008*"key" + 0.007*"tree" + 0.007*"kalos" + 0.007*"musides" + 0.005*"gate" + 0.005*"john" + 0.005*"box" + 0.004*"baying" + 0.004*"hill" + 0.004*"valley" + 0.004*"amulet" + 0.004*"grave" + 0.004*"beauty" + 0.003*"latin" + 0.003*"grove" + 0.003*"text" + 0.003*"marble" + 0.003*"tyrant"

Topic	Distribution
10	0.027*"oilman" + 0.016*"room" + 0.012*"elwood" + 0.010*"space" + 0.009*"rat" + 0.006*"old_woman" + 0.006*"pull" + 0.006*"brown_jenkin" + 0.006*"witch" + 0.006*"tomb" + 0.005*"gilman" + 0.004*"loft" + 0.004*"slant" + 0.004*"landlord" + 0.004*"garret" + 0.004*"organic" + 0.004*"vault" + 0.003*"ceiling" + 0.003*"angle" + 0.003*"dimensional"
11	0.032*"akeley" + 0.019*"hill" + 0.018*"voice" + 0.015*"letter" + 0.012*"human" + 0.011*"pickman" + 0.009*"record" + 0.008*"brattleboro" + 0.008*"street" + 0.007*"road" + 0.007*"vermont" + 0.007*"machine" + 0.007*"old_man" + 0.006*"noyes" + 0.006*"cylinder" + 0.006*"dog" + 0.006*"house" + 0.006*"warren" + 0.006*"speech" + 0.005*"romero"
12	0.029*"blake" + 0.014*"window" + 0.013*"tower" + 0.013*"jermyn" + 0.012*"church" + 0.011*"lightning" + 0.009*"mansion" + 0.008*"thunder" + 0.008*"steeple" + 0.008*"box" + 0.008*"earth" + 0.008*"squatter" + 0.008*"tempest_mountain" + 0.008*"horror" + 0.008*"sir_wade" + 0.008*"diary" + 0.007*"arthur_jermyn" + 0.007*"mountain" + 0.007*"mound" + 0.006*"hamlet"
13	0.047*"ward" + 0.046*"willett" + 0.032*"curwen" + 0.025*"charles" + 0.023*"ye" + 0.020*"doctor" + 0.017*"youth" + 0.016*"charles_ward" + 0.016*"joseph_curwen" + 0.016*"letter" + 0.013*"library" + 0.013*"allen" + 0.011*"providence" + 0.010*"laboratory" + 0.009*"pawtuxet" + 0.007*"party" + 0.007*"salem" + 0.007*"bungalow" + 0.006*"weeden" + 0.006*"capt"
14	0.026*"ammi" + 0.022*"nahum" + 0.017*"house" + 0.013*"uncle" + 0.011*"birch" + 0.010*"colour" + 0.010*"street" + 0.009*"horse" + 0.008*"bog" + 0.008*"castle" + 0.007*"coffin" + 0.007*"harris" + 0.007*"arkham" + 0.007*"wood" + 0.006*"road" + 0.006*"cellar" + 0.006*"zenas" + 0.005*"ruin" + 0.005*"gardner" + 0.005*"family"
15	0.010*"kuranes" + 0.009*"window" + 0.008*"friend" + 0.008*"dream" + 0.007*"street" + 0.006*"music" + 0.006*"city" + 0.006*"viol" + 0.005*"zann" + 0.004*"sky" + 0.004*"chair" + 0.004*"garret" + 0.004*"rue" + 0.004*"village" + 0.004*"room" + 0.004*"table" + 0.004*"drug" + 0.004*"note" + 0.003*"host" + 0.003*"playing"
16	0.015*"hill" + 0.014*"armitage" + 0.011*"whateley" + 0.011*"dunwich" + 0.009*"wilbur" + 0.007*"o" + 0.006*"ye" + 0.006*"glen" + 0.005*"boy" + 0.005*"city" + 0.005*"cohort" + 0.005*"old_whateley" + 0.005*"whippoorwill" + 0.005*"wilbur_whateley" + 0.005*"frye" + 0.005*"whateleys" + 0.004*"land" + 0.004*"road" + 0.004*"cattle" + 0.004*"big"
17	0.025*"edward" + 0.025*"asenath" + 0.016*"derby" + 0.015*"body" + 0.013*"rat" + 0.012*"cat" + 0.011*"servant" + 0.008*"norrys" + 0.007*"car" + 0.006*"roman" + 0.006*"family" + 0.006*"arkham" + 0.006*"wife" + 0.006*"dan" + 0.006*"old_ephraim" + 0.006*"book" + 0.006*"priory" + 0.006*"nigger_man" + 0.005*"exham_priory" + 0.005*"innsmouth"

Attempting to assign meaning from the topics' definitions:

Topic	Distribution
0	landscapes
1	country/village themes
2	post-apocalyptic, desertedness
3	music, beauty
4	knowledge, belief
5	mythology, travel, adventure
6	ruralness
7	landscapes
8	cult
9	mysticism
10	confined spaces
11	communication
12	gothic structures
13	providence
14	family
15	music
16	whateley hill
17	household

Topics common by year:

Year	Topics
1917	[10, 13]
1918	[8]
1919	[3, 9, 13]
1920	[11, 9, 1]
1921	[14, 9, 2]
1922	[7, 15, 6]
1923	[11, 17]
1924	[14]
1925	[6, 15, 14]
1926	[16, 13, 8]
1927	[5, 13, 14]
1928	[5, 16]

Year	Topics
1930	[11]
1931	[7, 6]
1932	[10]
1933	[16, 17, 4]
1934	[4]
1935	[12]

Observable trends in identified topics

From the topic information extracted, we can identify a couple of trends in the topics of Lovecraft's works:

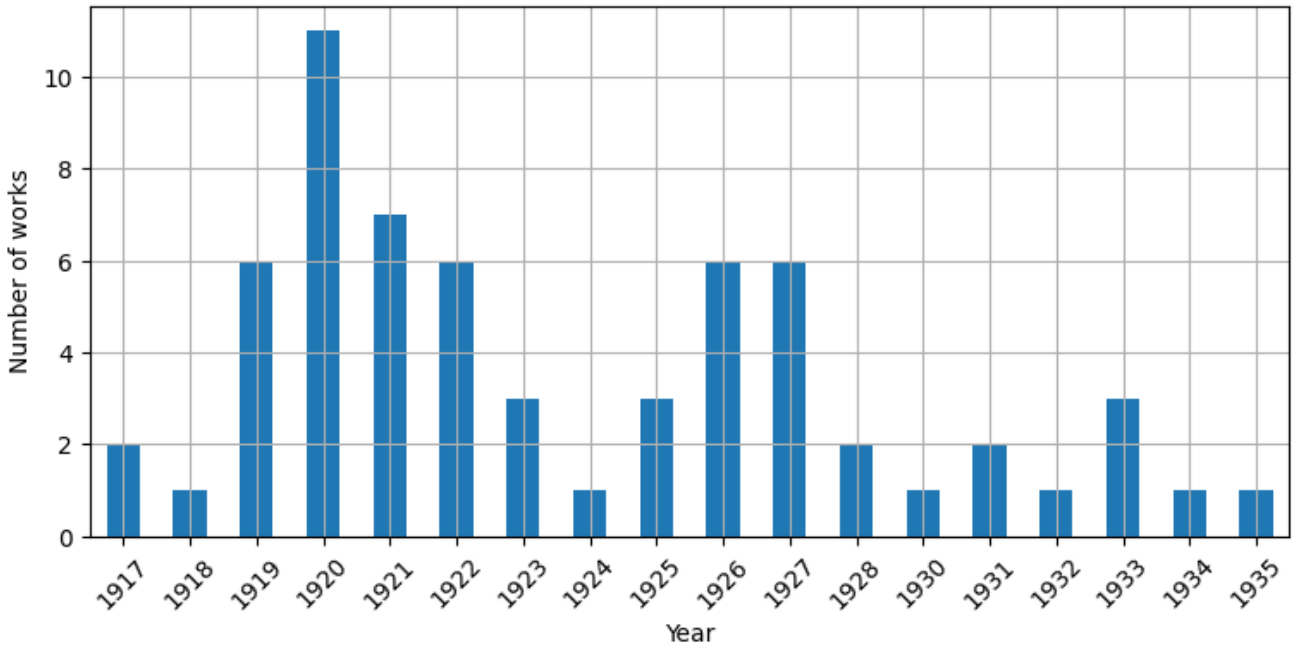
1. Topic 14 (family) appearing once in 1921 and recurreing between 1924-1927.
   Coincidentally in the year 1921 his mother, Sussie had died; an event that, according to his biography on Wikipedia page on H.P. Lovecraft, has effected his state of mind immensely.
   Also, in the year 1924 Lovecraft got married to his spouse.
2. Topic 9 seems to be a recurring theme between the years 1919-1921. those were the earlier years of his career. We assume that this was an experimental topic for him.

Automatic recognition of style change

Getting inspiration from the paper on Corpus Periodization Framework to Periodize a Temporally Ordered Text Corpus, we initially tried to make use of the outlined methods, to periodize the dataset automatically. This proved unsuccessful, since any threshold we set resulted in either the whole corpus identified as a single period or each year was assigned a period of its own.

This led us to periodize the dataset manually.

Judging from Lovecraft's activity over the years, initially we decided to split the dataset into 3 periods, based on spikes of activity.

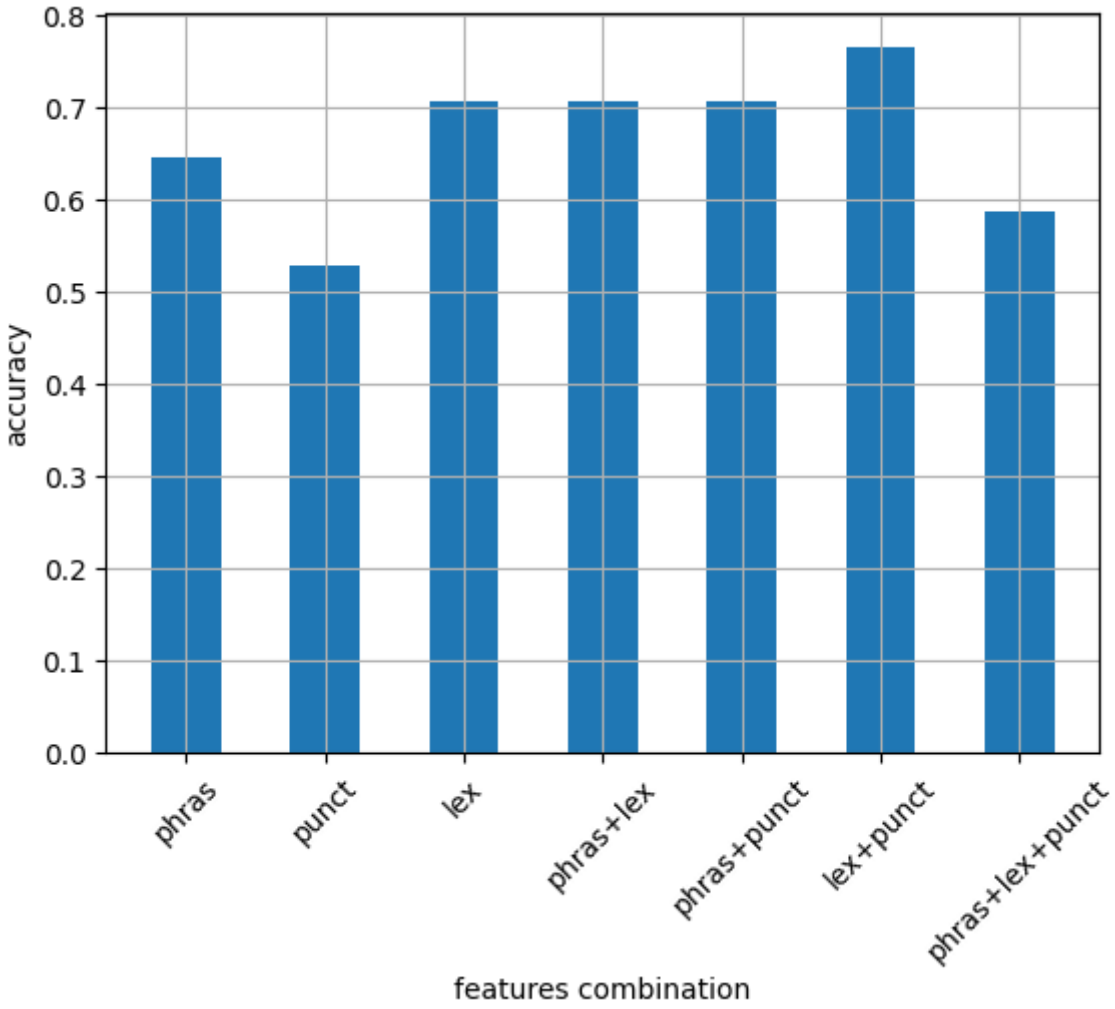


The 3 periods we decided on are: 1917-1922, 1925-1928, 1931-1935, skipping 2 years in-between.

Next we used the features, recommended in Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts, to create classification data. We created feature-vectors from phraseology, punctuation, lexical statistical info of the texts.

The classifiers we picked for our experiment are Logistic Regression classifier (from the popular `scikit` python package) and our own simple neural-network classifier.

The logistic regression classifier was run on different combinations of the statistical info from the texts.



The features that got us the most accuracy with logistic regression were the combination of lexical features and punctuation features of the texts.

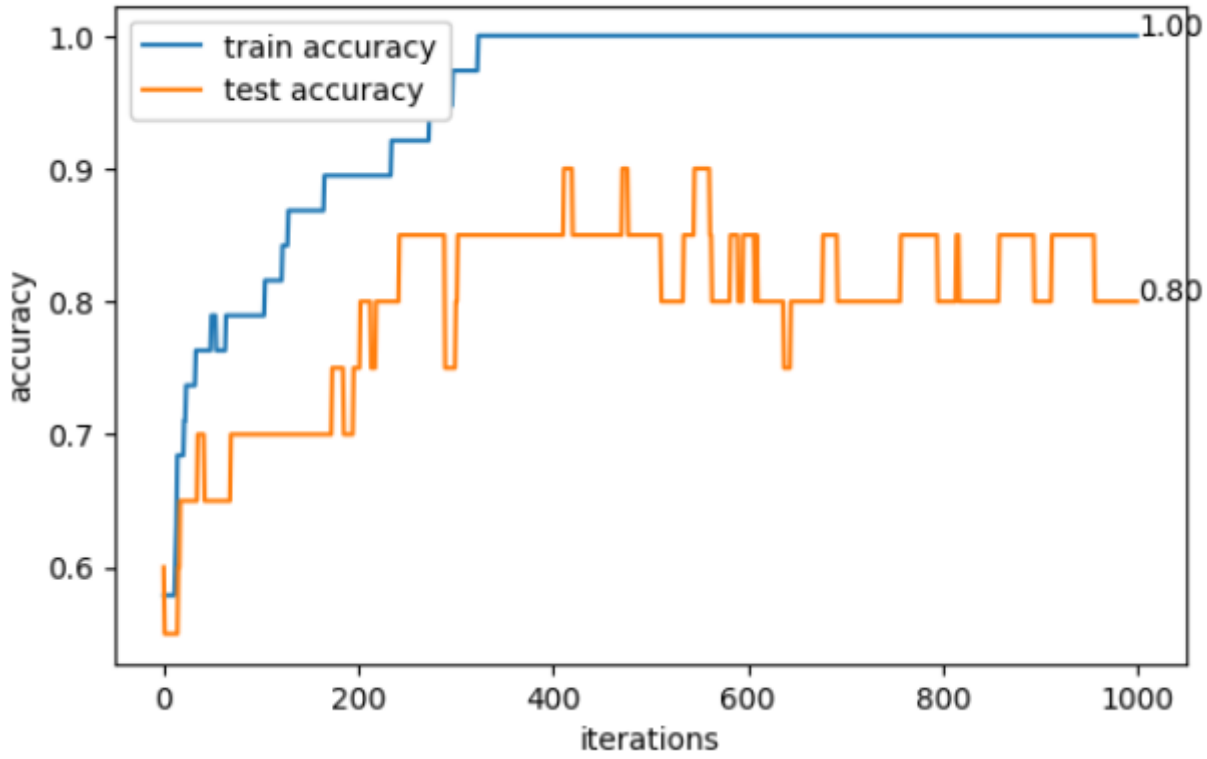
Seeing as we got decent accuracy with logistic regression, we wanted to see if can increase it by training a NN.

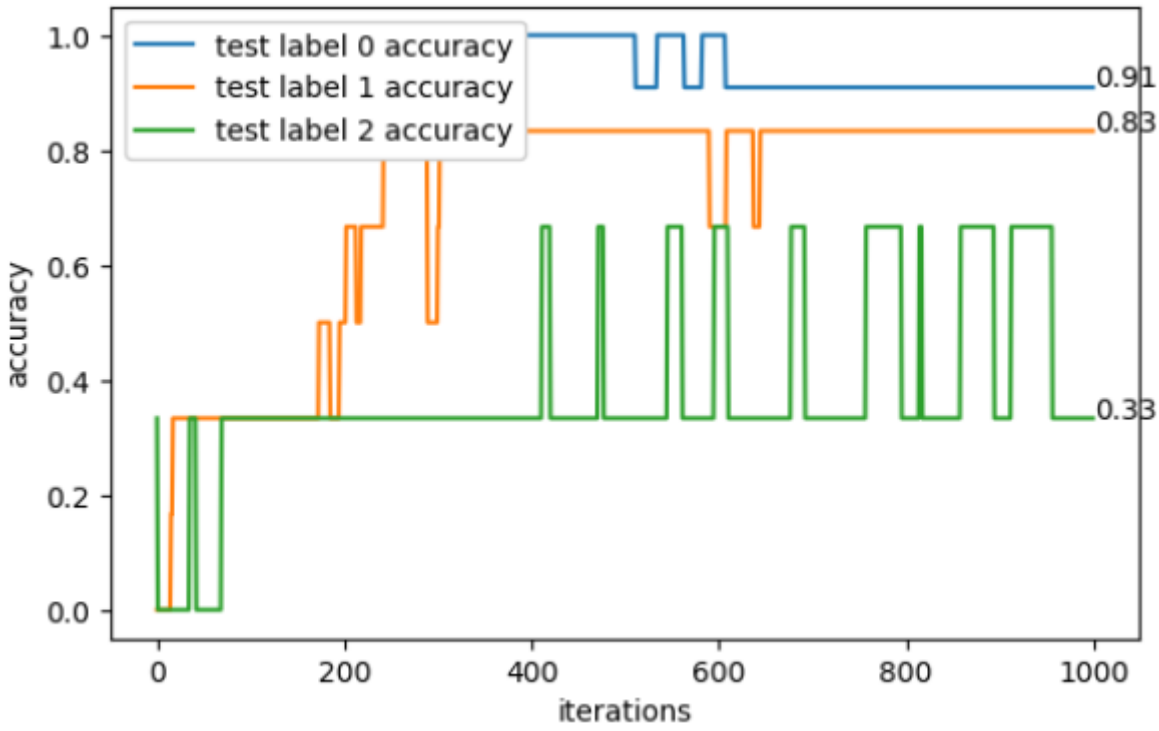
We implemented it using `pytorch` python library.
The architecture we chose for the NN consists of the following:

- a fully-connected layer which increases the features' dimension by a factor of 3,
- ReLU activation
- a droupout layer with probability of 0.95,
- a fully-connected layer, reducing the dimension to 3 (number of periods)

For the nn's training process we used cross-entropy for the loss function and `ADAM` for the optimization algorithm.

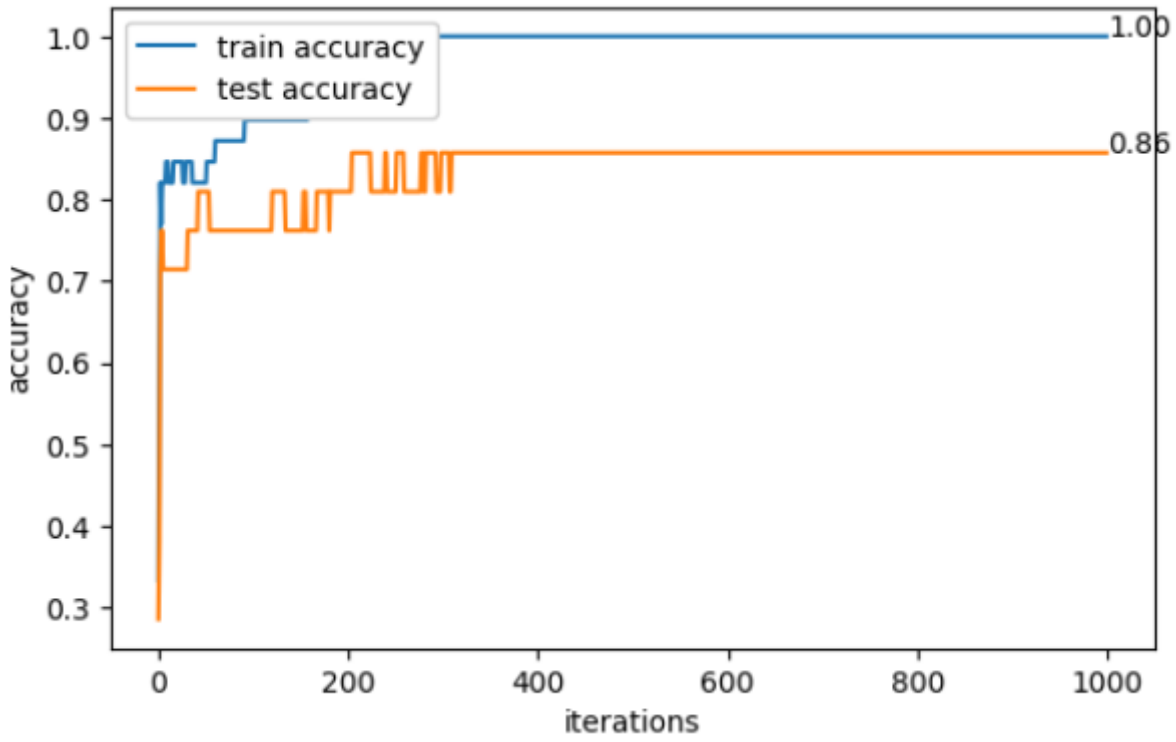
The following are the training results for the nn model (the numbers on the right are the final accuracies the model achieved):

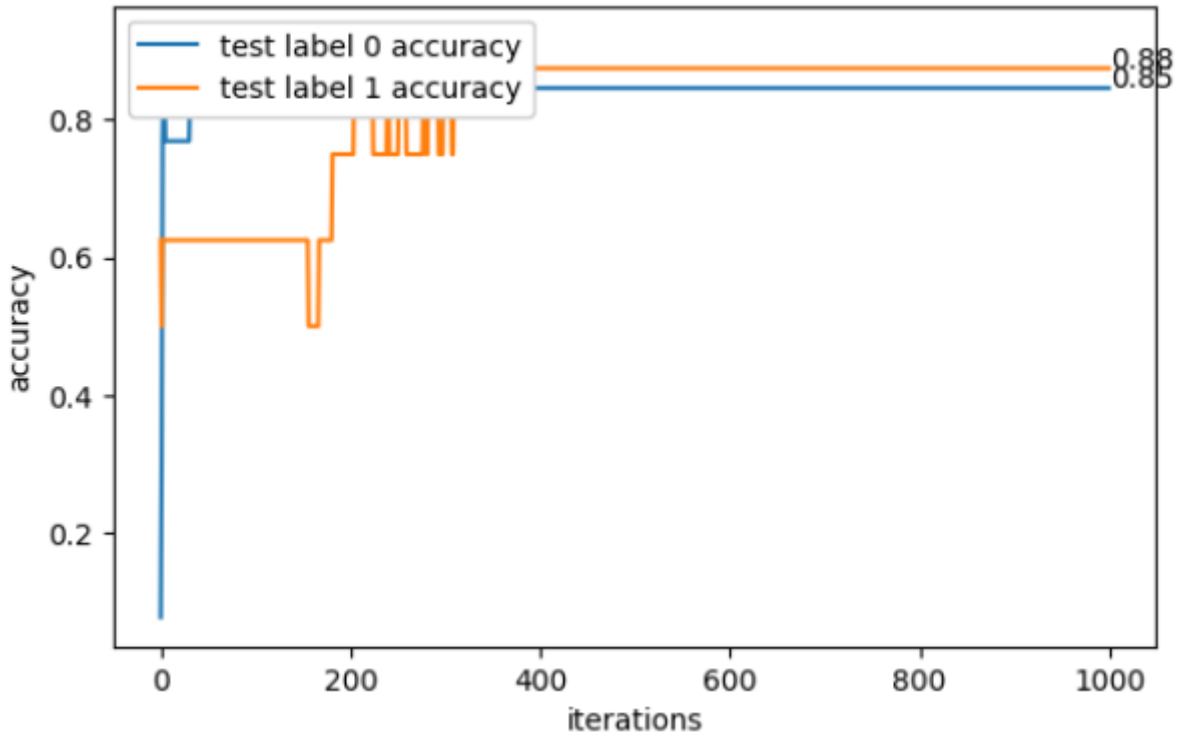




Seeing how the model struggled with differentiating between the second and third period, we decided to test what accuracy will be achieved we defined 2 periods instead of 3. We considered the best suitable periods for that division to be 1917 - 1924, and 1926 - 1935, which would align with the timeline of before and after Lovecraft writing of his more-popular works.

Using that division we got the following results (the numbers on the right are the final accuracies the model achieved):





Conclusions

We've shown that using extracted text features like phraseology, punctuation, lexical statistical info, different writing periods of H.P. Lovecraft can be differentiated between each other.

Final conclusions

From the results we showed in different parts of the report, we can say that we can indeed identify different stages of Lovecraft's changing literary style throughout the years of his career, through various statistical analyses of his works.