Culture of Conspiracy: Clustering Reddit's r/Conspiracy Board

Carolyn Seglem

Abstract

Reddit is a social media platform in which users may anonymously discuss conspiracy theories. Though conspiracy beliefs abound on the internet, their convoluted and sometimes illogical nature makes it difficult to predict where beliefs may converge or diverge. I sought to discover connections between common internet conspiracy theories over the past year through K-means and hierarchical clustering. These methods generated clusters according to various conspiracy topics, including the U.S. Presidency, vaccines, and Reddit culture.

Introduction

Conspiracy theories are inherently tangled and interwoven. The beliefs of conspiracy theorists rarely share one unified narrative. Most theorists develop their own unique beliefs composed from a set of common building blocks: a general set of people, technologies, historical events, and current news believed to be suspicious.

Reddit is a social media website driven by discussion amongst anonymous users. Currently, Reddit ranks as the 10th most visited website in the United States and the 20th most visited website in the world [1]. All posts on Reddit are organized into boards called "subreddits". Each board is devoted to a particular topic and usually has moderation to enforce that posts remain on topic. Posts may include text-based discussion, images, and links to external sites.

Many subreddits exist for specific conspiracy theory topics, including r/pizzagate and r/GreatAwakening (devoted to the QAnon conspiracy theory). However, r/Conspiracy is the general hub of Reddit's conspiracy theory discussion. As of April 2020, more than 1.1 million accounts are subscribed to the community. Among Reddit communities, r/Conspiracy ranks 255th in subscribers. The board receives approximately 888 posts and 9479 comments per day [2].

Methods

Dataset Cleaning and Preparation

PushShift has become a leading dataset for studies of Reddit content. The data contains crawls of all Reddit submissions and comments. The PushShift API returns a JSON file of up to 500 Reddit submissions at a time matching the provided criteria. The text content of Reddit submissions is contained in the "selftext" attribute [3].

External links drive much of the engagement on Reddit. Because the purpose of this study was to analyze the theories held by Reddit users themselves, submissions that contained solely links were filtered out. This was accomplished by removing all posts in which the selftext attribute was empty. Further filtration removed any links that were embedded within text posts. Posts that were deleted before being accessed by the PushShift web crawler are marked by the text "[removed]". All deleted posts were also filtered out.

All non-alphanumeric characters were stripped from the documents. It is a common practice on Reddit to post an image, with only a few words in the selftext field stating one's agreement with the content of the image. Therefore, all posts with less than three words were filtered out.

Natural Language Toolkit, or NLTK, is a Python library suite that offers packages for analysis of human language. Lemmatization is an advanced word stemming technique that analyzes the context of a word to find its root word. Some of its capabilities include reducing plural nouns to their singular form and changing all conjugated verb forms to the verb's base form. This process allows the analysis to consider a word's many inflected forms as having the same meaning. Lemmatization can be improved with by knowing the part of speech words are associated with. The NLTK stem package was used for lemmatization, and the NLTK corpus and tag packages were used for approximating each word's part of speech [4].

Stop words are words which are commonly used and reveal little to no information about a sentence's topics. Stop words frequently include articles, conjunctions, and prepositions. The NLTK corpus package provides a list of common English-language stop words which can be removed from the dataset with the stopwords() function [4]. During successive trials of my clustering analysis, I added several words to this list that I noticed to be particularly common amongst these posts.

1000 posts were sampled from each month from May 2019 to April 2020, representing one full year before the writing of this analysis. Of these 1000, 500 posts were taken from the beginning of each month, and 500 were taken from the middle of the month. After all filtration processes were complete, 9065 posts remained. These posts contained a total of 40504 unique vocabulary words.

Because machine learning algorithms expect numerical feature vectors rather than text, the text from the posts were vectorized in Term Frequency, Inverse Document Frequency (TF-IDF) representation. This process begins by assigning a numerical ID to each word. Each number is then weighted according to its frequency in each given document and in the document set. Words that are common across the dataset offer little insight into the differences between documents and are therefore weighted less than words that appear rarely. I vectorized the processed text with the TfidfVectorizer() function from the Scikit-learn machine learning library [5]. During vectorization, words which were found in more than 70% of documents were removed because their high prevalence provides little insight into the document's distinguishing features. Words which were found less than 10 times in the data were removed to account for factors like spelling errors. The resulting vectors has a dimensionality of (9065, 7417), representing 7417 words used across 9065 documents.

Clustering

K-means clustering was applied first to the vectorized data. To estimate the optimal number of clusters, K-means was run repeatedly, using a range of clusters from 2 to 20. Silhouette analysis was performed for each number of clusters to determine the degree of separation between clusters. The number which produced the best silhouette score was selected as K. As further validation, the inertia of each run was plotted over the number of clusters used.

The point at which the rate of decrease in inertia changed most sharply, commonly called the “elbow point”, is another indicator of a strong K value.

With the K-value chosen, the clusters were calculated once more using the K-Means function from the Scikit-learn library. The K-means algorithm used the Euclidian distance metric. The vectorized data was in a high-dimensional structure inconvenient for visualization purposes. The PCA() function from the Scikit-learn library was used for performing dimensionality reduction on the data, allowing for plotting in 2- and 3-dimensional space [5].

Hierarchical clustering was applied next to the vectorized data. The SciPy library for scientific and mathematical operations contains the linkage() method, which generates a linkage matrix from observation vectors [6]. The Ward variance minimization algorithm was used to calculate the distances between clusters, using the Euclidian distance metric.

The vast size of this dataset prohibited the construction of a dendrogram of all posts in the data. Therefore, the size of the dendrogram was truncated with the same number of classes as the number used in the K-means clustering. The fcluster() function from the SciPy library used the linkage matrix to assign each post in the data to clusters [6].

Results

Analyzing the clusters through the elbow method proved to be difficult. There were no points within the range of 2-20 clusters which saw a significant difference in the rate of change in inertia. The findings from the silhouette method showed that the most distinct clusters at 18. Therefore, the data were grouped into 18 clusters by K-means and agglomerative clustering.

The clusters produced by the K-means algorithm overlapped each other to a significant degree. The silhouette analysis of the final 18-cluster run returned a value of 0.00588. The proximity to 0 suggests that the majority of points were extremely close to the boundary of a different cluster than they were assigned to.

The clusters produced by the hierarchical clustering algorithm fared worse, with a silhouette analysis of -0.00290. There was a large disparity in the sizes of clusters produced by the hierarchical clustering algorithm. One cluster contained 5094 documents, 56% of the total number of documents.

Discussion

Through clustering, I encountered interesting and unexpected trends within the data. Some of the clusters were not surprising, featuring topics that have dominated the news cycle of the past year for in controversial manners. Such clusters include the ones devoted to United States President Trump, the apparent suicide of Jeffrey Epstein, and the Coronavirus pandemic. Other clusters were more surprising. I hardly expected, for example, the 2020 Sonic movie to be a significant enough topic to warrant its cluster.

I was disappointed in the lack of a clear distinction between clusters. Both algorithms produced a clusters that had distinct topics, but they also produced clusters which lacked any

distinguishable subject. I expected such a result from the outset; after all, most people combine multiple conspiracy topics to form their own, hybridized theories (a recent notable example being the perceived connection between 5G towers and COVID-19). With such a great deal of mixing between topics, it is little surprise that the algorithms developed some clusters that had no distinguishable main topic.

This was certainly aggravated by the lack of purity within this dataset. Preparing natural language text data can be complicated even when using text that follows the standard rules of language. However, that task becomes substantially more difficult when dealing with posts from online forums. The texts I collected were rife with spelling and grammatical errors, abbreviations, and slang terms unique to the internet or to specific conspiracy theory communities. Such irregularities artificially increase the number of unique words in the dataset while decreasing the frequencies of the actual keywords they are meant to represent.

Research suggests that Euclidian distance may be an ill fit for document analysis. In the future, the first priority should be examining better methods for calculating distance between documents. Better techniques for text preparation and cleaning might be employed to better control for the unusual language features prevalent in online discussion. It may also help to run the clustering analysis with a higher k-value, perhaps searching for an ideal value in the range of 0 to 40.

As an extension of this project, it could be fascinating to introduce the date of posts as a variable, analyzing how certain conspiracies may rise and fall in popularity in accordance with the world news cycle.

References

[1] *Top Websites Ranking.* SimilarWeb LTD, 2020. Accessed on: Apr. 13, 2020. [Online]. Available: https://www.similarweb.com/top-websites

[2] *Subreddit Stats.* Accessed on: Apr. 13, 2020. [Online]. Available: https://subredditstats.com/r/Conspiracy

[3] *API Documentation*. pushshift.io. Accessed on: Apr 7, 2020. [Online]. Available: https://pushshift.io/

[4] Bird, Steven, Ewan Klein, and Edward Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009. Accessed on: Apr. 10, 2020. [Online]. Available: https://www.nltk.org/book/

[5] Pedregosa et al., *Scikit-learn: Machine Learning in Python.* JMLR 12, pp. 2825-2830, 2011. Accessed on: Apr. 9, 2020. [Online]. Available: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[6] P. Virtanen et al., *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*. Nature Methods, 2020. Accessed on: April 13, 2020. [Online]. Available: https://www.scipy.org/scipylib/index.html

Tables

Table 1. K-means Clustering Results

Cluster	Words Nearest to Center	Most Frequent Words	Cluster Size
0	guy, woman, talk, look, want	guy, go, look, want, say	322
1	moon, alien, 51, area, UFO	alien, moon, area, 51, time	161
2	conspiracy, theory, believe, theorist, sub	conspiracy, theory, believe, want, look	454
3	human, god, world, life, time	time, world, human, life, come	676
4	Epstein, Jeffrey, suicide, kill, Clinton	Epstein, Jeffrey, go, case, time	310
5	white, black, American, race, woman	white, black, American, race, call	137
6	look, thought, time, use, year	go, time, use, look, make	3527
7	government, world, want, country, money	government, time, use, want, year	1082
8	post, Reddit, comment, sub, mod	post, Reddit, comment, sub, r/	451
9	war, Iran, Israel, attack, world	war, Iran, world, Israel, go	243
10	meditation, Merkaba, peace, link, heal	mass, awareness, field, mother, great	17
11	video, YouTube, watch, find, channel	video, YouTube, watch, post, find	384
12	virus, China, corona, Chinese, coronavirus	virus, China, world, make, spread	353
13	gt, lt, year, state, new	gt, year, say, time, use	102
14	search, Google, find, result, engine	search, Google, find, use, result	131
15	big, happen, deal, time, day	big, go, time, come, happen	159
16	Trump, president, Donald, Democrat, Clinton	Trump, president, time, say, use	362
17	vaccine, news, flu, measles, medium	news, vaccine, time, medium, use	194

Table 2. Hierarchical Clustering Results

Cluster	Most Frequent Words	Cluster Size
0	Gt, 1, come, 2, year	42
1	another, yt, page, wish, 13	17
2	Kong, Hong, China, government, happen	69
3	virus, China, Chinese, make, world	139
4	video, YouTube, watch, find, Conspiracy	167
5	post, sub, comment, Reddit, conspiracy	424
6	Epstein, Clinton, come, Jeffrey, Trump	160
7	Biden, Joe, Trump, well, fire	31
8	Trump, time, say, state, use	454
9	vaccine, use, year, cause, time	60
10	moon, space, NASA, go, land	61
11	time, world, go, come, want	637
12	strange, search, nfrealmusic, admit, target	8
13	white, Israel, Jew, year, state	157
14	movie, Sonic, trailer, make, look	18
15	time, go, use, look, year	5094
16	virus, Coronavirus, time, world. Well	303
17	want, go, time, world, use	1224

Table 3. Sampled Predictions

Reddit Post	K-Means Cluster	Hierarchical Cluster
Can we just start looking into shill accounts. Like list the reasoning and the account name in the comments and our super sleuths can do some looking than mods can ban because this sub is literally under seige.	8 (post, Reddit, comment, sub, mod)	5 (post, sub, comment, Reddit, conspiracy)
I'm not just talking about at least trying to save the level of discourse online from the influence of the elites or always being right when it comes to researching the topics related to the corruption going on in the government but also of being reliable to others from a moral point of view when it comes to dealing with all the degeneracy and insanity around us...	7 (government, world, want, country, money)	15 (time, go, use, look, year)
... Senator Lindsay Graham on Donald Trump in 2015: "Donald Trump is a racist, xenophobic bigot." Something happened to change Lindsey's scope on Trump and impeachment? What could it be? Does the presidents's formost cheerleader eat at Comet Pizza?	16 (Trump, president, Donald, Democrat, Clinton)	8 (Trump, time, say, state, use)
Hi all. Can anybody suggest a search engine that performs **exactly** like google did in the early days? Proper page ranked results, proper results as per an advanced search string query (like intext, intitle, inurl) If there isn't one that can work *precisely* in this way, can anybody give a good reason why not?	14 (search, Google, find, result, engine)	4 (video, YouTube, watch, find, conspiracy)
... Climate change has become the incubator for these 'Franken Viruses' so it will spread like wild fire which is ironic because "The Global Smoke Cloud" in the atmosphere will carry the Corona Virus from county to country and it will be an "invisible rain".	12 (virus, China, corona, Chinese, coronavirus)	3 (virus, China, Chinese, make, world)

* Note: posts which were several paragraphs long were cut for table conciseness

Figures

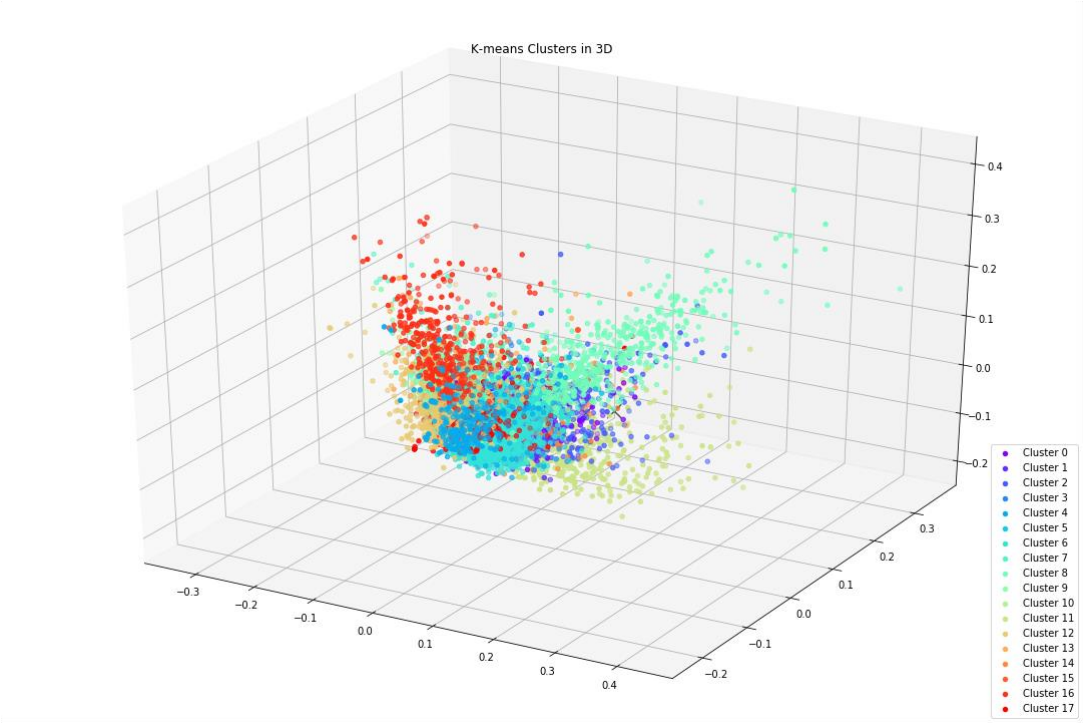

Figure 1. 3-dimensional scatter plot representation of K-means clusters.

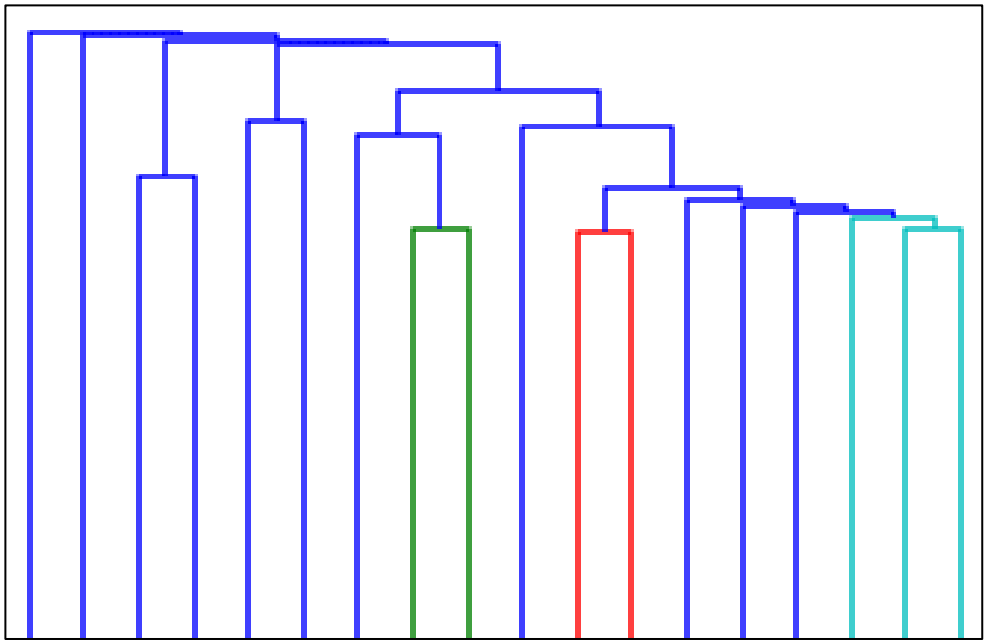

Figure 2. Dendrogram of hierarchical clusters.