COVID, Conflict, and Reddit

A 5-year retrospective analysis of r/AmITheAsshole

Annabelle Strong
as8@wellesley.edu
Wellesley College
Wellesley, Massachusetts, USA

ABSTRACT

Reddit's r/AmITheAsshole community is a popular and divisive forum in which individuals post descriptions of personal moral dilemmas or accusations for other users to judge. While the judicial dynamics of subreddit are well-documented, few studies have examined the implicit sociological data embedded in its posts. In response to this blind spot, this paper aims to track the common topics behind those submissions over a five-year period (between June 2016 and June 2021, inclusive) to see if and how the COVID-19 pandemic and subsequent social shift impacted the subjects, prevalence, and nature of interpersonal conflict. Through topic modeling analysis, we find that no such shift is evident in the data.

KEYWORDS

Reddit, topic modeling, LDA, conflict

ACM Reference Format:
Annabelle Strong. 2022. COVID, Conflict, and Reddit: A 5-year retrospective analysis of r/AmITheAsshole. In *Proceedings of Data and Text Mining (CS 315).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Background and Motivation

The boom of the social internet over the last two decades has ushered in a new era of complex public interactions. With a global network available to anyone with a computer, social dynamics are becoming ever more visible and quantifiable through social media sites like Reddit. This study aims to examine a subset of such data, Reddit's r/AmITheAsshole (hereafter referred to as *AITA*).AITA is a subreddit in which individuals who may be in the wrong write about their situation and get an assessment of blame from other users. Given the acute popularity of AITA and the crowdsourced moral code it produces (see [5],[2],[4],[9] for deeper analysis of the patterns and dynamics of such judgements), the subreddit provides a uniquely accessible and authentic insight into the types of conflict experienced by users. This paper aims to detect and quantify patterns in these categories over the last five years, if any exist, to better understand the impact of worldwide social shifts necessitated by the COVID-19 crisis on social discord.

To accomplish this goal, this study parses submission dumps for each month between and including June 2016 and June 2021 from the PushShift archives. After extracting the posts of interest—namely AITA posts with the original submission text intact—we will apply topic modeling techniques to identify common sets of language both across and within the months and years in question. These topic sets will then be compared with respect to time, allowing us to discover and examine any latent patterns.

1.2 Research Questions

(1) What were the most common topics in the dataset?
(2) How similar are these topics across temporal groups?
(3) What, if any, significant shifts in these topics occurred in early 2020?

RQ1 and **RQ2** aim to establish the variety and consistency of topics present in AITA posts irrespective of larger social and temporal context. **RQ3** examines this data longitudinally and holistically to detect and understand any notable shifts that may have occurred.

2 LITERATURE REVIEW

2.1 Subreddit Demographics

2.1.1 "Don't Downvote A$$$$$$s!!": An Exploration of Reddit's Advice Communities. Cannon et al. (2021) analyzed posters' demographic information with respect to their post's linguistic features in AITA and the related r/relationships subreddit. Among other findings, they determined that the majority of AITA posters who disclosed their gender identity were women, and that the most heavily represented age group is 18-24, followed by 25-34. They also found that self-disclosed male posters were more likely to elicit a negative community ruling than their female counterparts, although they acknowledged that the rarity of strongly negative rulings limit the statistical power of this finding.[3]

2.2 Social Effects of COVID-19

2.2.1 Day-to-day relational life during the COVID-19 pandemic: Linking mental health, daily relational experiences, and end-of-day outlook. Merolla et al. (2021) found that participants on generally reported minimal daily interpersonal conflict during the pandemic—on a scale from 1("None") to 7("A great deal"), the mean was 1.46 with a standard deviation of .68. Other measures of social difficulty were also found to be relatively low. When asked to rate their agreement with the statement "Maintaining relationships was difficult and frustrating today" on an increasing scale from 1 to 7, participants reported an average of 2.98 (SD = 1.62), which

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CS 315, Spring 2022, Wellesley College
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

corresponds to slight disagreement. Surprisingly, when asked to do the same to indicate their degree of satisfaction with their social interactions, the participants reported an average of 3.56 with a standard deviation of 1.02—just slightly on the negative side of neutral. This study was conducted on 120 students at a large American university, with an average age of 21. As such, it captures the social experiences during the pandemic of individuals in the age group most heavily represented in AITA . That being said, the small size of the survey limits the conclusions that can be drawn from the data.[6]

2.2.2 A Study on the Psychological Wound of COVID-19 in University Students. A similar study done by Padrón et al. in 2021 examined the sources of stress in 932 Spanish university students (most aged 18-26) and their coping strategies. The authors found that changes in interpersonal conflict was ranked lowest of the five stressors they investigated—below academic, social distancing, pandemic, and general overload respectively. These factors were all moderately and significantly correlated, with correlation coefficients ranging from .29 to .53. This general increase in each source of stress corresponds to an increase in reported mental health issues. Most participants felt that their mental health changed during the COVID-19 crisis, 57.5% labeling it as "worse" and 14.7% as "much worse". This decrease was more pronounced in women than in men, though it should be noted that women were more heavily represented in the participant group (75.5%). [7]

2.2.3 COVID-19 and Mental Health/Substance Use Disorders on Reddit: A Longitudinal Study. Alambo et al. established correlations between BERT-generated topics for r/Coronavirus and the subreddits r/anxiety, r/depression, r/Suicidewatch, r/opiates, r/OpiatesRecovery, and r/addiction for each month between January and October (inclusive) of 2020. They found that the correlation for r/anxiety spiked early in the crisis, with a peak correlate value of .8 in May of that year, up from around .4 in February. r/depression was more variably correlated, with peaks and lows cycling every three months or so. r/Suicidewatch, however, steadily increased in correlation, with its peak (roughly .6) in August. Together, these three trends suggest that Reddit users generally experienced increased anxiety at the beginning of the pandemic, followed by periodic depression and increasing suicidal thoughts in the subsequent months. A similar story can be seen in the three substance-addition subreddits. r/opiates roughly increased in topic correlation with r/Coronavirus before peaking in July and beginning to decrease, while r/OpiatesRecovery spiked in July and continued to have relatively higher correlation coefficients (over .3 in September compared to just over .1 in June) for the rest of the months studied. Correspondingly, r/addiction increased steadily throughout the year, though it slightly decreased after August. Taken together, the data implies that some users began or increased their opiate usage as the pandemic ramped up before turning to substance management and recovery communities in the fall. Combined with the findings in Merolla et al. and Padrón et al., it seems that the effects of the pandemic were largely internal (relating to mental health, substances, academics, etc.) rather than interpersonal in nature. If this were the case, we would expect to see little change in AITA topics during the onset of the pandemic, as submissions typically pertain to social issues. [1]

2.3 Longitudinal Topic Modeling

2.3.1 A Geometry-Driven Longitudinal Topic Model. A 2021 paper by Wang et al. developed and analyzed a methodology for calculating and tracking LDA-generated topics over time in social media data. Using tweets collected between February 15 and May 15th, 2020, the authors first temporally smoothed the corpus by combining temporally-close documents according to an exponential formula. More specifically, for each set of documents d_{t_1} associated with time t, a proportion of each other set of documents d_{t_0} was calculated using the formula

$$w(t_0, t_1) = \gamma^{|t_1 - t_0|} \tag{1}$$

with coefficient $\gamma = .75$. Applied to every pairwise combination between one time slice (e.g. one particular day) x and its n peers (self-inclusive), its set of weights W_x is calculated to be

$$W_x = \{w(t, x), 0 \leq t \leq n\} \tag{2}$$

$$= \{.75^{|t-x|}, 0 \leq t \leq n\} \tag{3}$$

For $t = 2$, for example, the corpus would contain 100% of d_2, 75% of d_1 and d_3, and so on:

$$W_2 = \{.5625, .75, 1, .75, .5625, .4219, ..., .75^{|n-2|}\} \tag{4}$$

This method ensured that each set of documents was representative of a soft range of time rather than a single point t without biasing them with text too far removed in time. The authors then applied Latent Dirichlet Allocation (LDA) to each corpus to extract a set of topics for each. These topics consist of a set of words that are likely to occur together in the given corpus, and therefore capture cohesive patterns in its texts. These topics were then compared pairwise using the Hellinger distance formula (Equation 5) to establish geometric similarity, which the authors noted is more effective for probability distributions than Euclidean distance. The resulting

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=0}^{n} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{5}$$

Figure 1: The Hellinger Distance formula

$K \times T$ matrix represents quantified similarity between ever topic K associated with each time T, with higher values indicating higher dissimilarity. The authors used these values to find topic trends using shortest-path techniques (such as Dijkstra's algorithm) on a 10-nearest neighbor graph derived from the Hellinger matrix. They separately graphed the distances represented by the Hellinger matrix using a variety of multidimensional scaling methods including t-distributed stochastic neighbor embedding (t-SNE), which is our chosen method for scaling the dimensionality of our own data. This paper relies heavily on these techniques proposed by Wang et. al, and will employ highly similar methodology to find topic trends in the posts we collect.[10]

3 DATA AND METHODS

3.1 Collection Strategy

The data required to build on these findings will be collected from PushShift API, a public database of Reddit posts and comments

Figure 2: Histogram of collected submissions over time (divided at March 15th, 2020, roughly when US lockdowns started)

created by Jason Michael Baumgartner, Eddy Lazzarin, and Alexander Seiler. This method was chosen over the alternatives (such as the Reddit API or web crawling)[8] because of its cohesive and complete collection of submissions by month, which simplified the collection and organization process.

From its archives we downloaded compressed files of Reddit submissions, each containing every post for a particular month. As we were only concerned with the topics brought up by each post but not in their responses, we did not download the associated comment data. In total, we collected 61 such files for June 2016 through June 2021 (inclusive). From these we extracted all posts from AITA, filtering out those written by moderators and those without the post body intact. All in all, the raw dataset contained close to 300,000 unique posts (see Figure 3), each accompanied with the post's body text and details about its author's username, posting date and time, popularity (via comment and crosspost counts), etc., although for our purposes only the body text is of interest. The only modification made to this dataset was the addition of year, month, and day columns using data extracted from each submission's UTC string indicating its time of posting.

	num_comments	subreddit_subscribers	year	month	day
count	297442.000000	2.949030e+05	297442.000000	297442.000000	297442.000000
mean	72.988862	1.894884e+06	2019.570407	6.401288	15.850099
std	295.574866	8.095511e+05	0.770387	3.369811	8.835055
min	0.000000	5.045700e+04	2016.000000	1.000000	1.000000
25%	12.000000	1.202962e+06	2019.000000	4.000000	8.000000
50%	21.000000	1.874987e+06	2020.000000	6.000000	16.000000
75%	41.000000	2.744655e+06	2020.000000	9.000000	24.000000
max	99993.000000	2.891768e+06	2021.000000	12.000000	31.000000

Figure 3: Ranges and averages of quantitative columns

3.2 Data Summary

The final dataset contains 297,442 submissions posted by 243,728 different users on 1,794 individual days during the 5-year period of interest. The vast majority of these posts are from the latter half of this window, with 2019 marking the 25th percentile for year of submission. This corresponds with subscriber count growth over the years, which steadily increased from about 50,000 to almost 3 million. Comment counts also increased, but far less dramatically—most posts had fewer than 100 comments regardless of year posted. The biggest outlier is an open forum from June 1st, 2020 that garnered nearly 100,000 comments. As such, that submission will be excluded from analysis. Post length, on the other hand was fairly consistent: as Figure 4 shows, most posts had around 10^5 to 10^6 words in their body text, with a negative skew on the distribution and a maximum of roughly 10^8.

Figure 4: Histogram of logarithmically-scaled post word count

3.3 Methodology

To extract and interpret topics embedded in this corpus, we will employ the following strategy motivated by the methods proposed in Wang et al. [10]:

(1) Extract vocabulary from dataset and generate topics ($K = 10$) with LDA (**RQ1**)
(2) Group and temporally smooth the collected documents by month
(3) Apply LDA with the extracted collective vocabulary to each resulting corpus independently to extract a total of 610 topics ($K = 10$ for each $0 \leq t \leq 60$)
(4) Compute the perplexity values of each pairwise combination of trained LDA model and vectorized document
(5) Generate the similarity matrix $K \times T$ by calculating the Hellinger distance metric for each pair of topics (**RQ2**)
(6) Plot individual token frequencies over time and compare with COVID-specific language (**RQ3**)

Note that the choice to extract $K = 10$ topics from the corpus and its subsets is not necessarily optimal. Most previous work in topic modeling calculates an optimal value for K by comparing performance scores (such as perplexity and coherence) between models trained to extract different numbers of topics from the same data. While we encourage future work to do the same with respect to the data at hand, time and computing constraints restricted our K value to an integer large enough to capture topic variety, but small enough that pairwise calculations (steps 4 and 5 above) were able to be completed in a reasonable amount of time (though the operations required for step 5, described in 4.3.2, still took roughly nine hours total). We thus arbitrarily chose 10 for this value.

4 RESULTS

4.1 Vocabulary Extraction and Collective Topics

Before splitting the corpus into bins by month, we analyze the textual features of the entirety of the dataset to get a sense of overarching patterns. First, we vectorize the documents using term frequency-inverse document frequency (TF-IDF) as the quantifier. This method assigns an importance score to each word in the corpus that is directly proportional to the frequency of that word within each document and inversely proportional to its frequency across the corpus. Using the resulting document-topic matrix (DTM), we train an LDA model to extract $K = 10$ topics as defined by correlated probabilities between words. After plotting the result using pyLDAvis with TSNE scaling, we manually label the resulting topics based on observed commonalities in the words within each set (see Figure 5). The result answers our first research question **RQ1**, revealing that the top topics in AITA are *Relatives*, *Social*, *Home Life*, *Work*, *Logistics*, *Food*, *Neighborhood*, *Pets*, *Video Games*, and *Individuals* respectively.

4.2 Temporal Modeling

To get a more granular view, we group the data by month and apply the temporal smoothing technique described in Section 2.3.1. To summarize, we modify each group to include a exponentially decaying portion of each other group's documents based on temporal distance between the two sets. Figure 6 shows the number

Figure 5: Topics for the full corpus

of documents from each month (y-axis) in the resulting 61 group compilations (x-axis). Having compiled these groups, we vectorize each corpus as we did the collective in Section 4.1, using the vocabulary generated in that section as a base to ensure consistent vectors. The resulting DTMs are then used to train individual LDA models.

Figure 6: Number of documents in each group within the temporally-smoothed corpus

4.3 Similarity Analysis

4.3.1 Perplexity. Perplexity is a measure of how well a model predicts a sample (see [**?**] for more detail)(*error with auto-citation; intended reference is [4]*). To apply it here, we feed models trained on each month's data with the document-topic matrices of each other month's data to quantify the similarity between models. The result can be seen in Figure 7a. Bright colors indicate that the the model trained on the month indicated on the y-axis is not very predictive of data from the month on the x-axis. More specifically, the bright patch in the upper-right corner shows that models trained on 2016-2017 data were "surprised" by submissions from later years.

(a) Perplexity between each pair of months' models

(b) Hellinger distances between topic vectors

Figure 7: Topic similarity metrics

This implies that some kind of shift occurred around late 2018, causing the later datasets to contain words that earlier models didn't encounter in their training.

4.3.2 Hellinger Distance. To more closely examine this shift, we compare the topics generated by each of these models. As discussed in Wang et al. [10], this is done by calculating the Hellinger distance between the topics' probability distributions. We first extract the topic vectors and compile them into a 610×610 matrix, each row and column representing one of ten topics for each of the 61 months in the dataset. We then apply the Hellinger equation (see Figure 5) pairwise to this topic matrix to obtain an array of distance coefficients (as visualized in Figure 7b). The resulting striped heatmap shows highly consistent (though not constant) Hellinger distances between each topic and the others. Topics from the latter half of the dataset are more frequently dissimilar, though those that have high distance values to the majority of the nearly always have low distance values to each other. This suggests a split in topics, where the topics that appear in blue on the heatmap represent one set of cohesive topics and those that appear in orange or yellow represent another. It's noteworthy that topics in the second category span several years, including but not limited to the months since COVID-19 first became a global issue. This implies that such topics are not specific to the pandemic, instead representing a different subject of discussion that first becomes relevant around late 2018. That being said, the number of posts in AITA dramatically increased around the same time, so it's possible (if not probable) that document count is responsible for some degree of the apparent split in topics.

Figure 8: Figure 7b magnified between the coordinates ((2018, 10), 6) and ((2019, 2), 3)

From these results we also noticed consistently low Hellinger distances for a given topic number over time. As shown in Figure 8, the intersections of two bright bands representing the same topic number (in this case, most representing Topic 7) between months were noticeably darker than their surroundings. This implies that some degree of consistency might exist for each numbered topic.

To further investigate this possible trend across time, we filter each row of the Hellinger matrix to only include values where the constituent topics were of the same number (e.g. the values comparing Topic 3 from month 0 with Topic 3 for every other month). We flatten these lists by taking the median of each, which results in a 61×10 matrix where each cell contains the median dissimilarity between topic y for a given month x and the topic y values of the other months. As seen in Figure 9a, most topics have

(a) Violin plot

(b) Heatmap

Figure 9: Hellinger distances for each topic number over time

fairly constant Hellinger distances between their members. The exceptions are topics 2 and 7. Upon further investigation, the two share sets of words: "just", "like", "don"('t), "time", "friend", "told", "said", "didn"('t), and "really", among a few others. These words are among the top 10 listed for Topic 2 in the months prior to April 2018, while they are among the the top 10 for Topic 7 instead after that point. This switch seems to be at least partly responsible for the inconsistencies within the two topics—looking at Figure 9b, the month of the switch is clearly marked with a distinct change of color in both rows.

4.4 Token Trends

Considering this apparent lack of COVID-related topic changes, we look closer at the frequencies of certain words over time to see if any individual trends emerge. Figure 10 shows topic frequencies throughout the corpus for the words "anxious", "depressed", and "covid" respectively. These specific tokens were of interest due to the findings of Padrón et al. [7] and Alambo et al. [1], both of which indicated increased anxiety in depression in the groups studied, with the latter specifically investigating Reddit posts. The consonance between these results suggested that if there were to be any COVID-related shift on AITA, it would likely entail increases in the frequencies of anxiety- and depression-related submissions. The graphs do not support such an increase—while the term "covid" predictably experienced a sharp increase in usage in early 2020, the other two tokens don't appear to do the same. The only notable pattern—a dramatic increase in late 2018—is the same one we have previously discussed as a consequence of increased user activity. While we encourage future work to examine normalized figures that quantify the relative popularity of each term with respect to the subreddit's popularity, the data shown clearly indicates that no COVID-specific shift in these terms occurred.

That being said, there are a number of other terms that we did not manually track—such as "domestic violence" or similar—that may have been more volatile. Thus, our findings in this section should not be taken as a generalized conclusion about all potential COVID-related terms being uncorrelated with the onset of the pandemic; rather, just those explicitly discussed.

Figure 10: Token frequency over time: "anxious" (top), "depressed" (middle), and "covid" (bottom)

5 DISCUSSION

5.1 Research Question Results

5.1.1 RQ1: What were the most common topics in the dataset? Unsurprisingly, the top topics—*Relatives*, *Social*, *Home Life*, *Work*, *Logistics*, *Food*, *Neighborhood*, *Pets*, *Video Games*, and *Individuals* (see Section 4.1 and Figure 5)—are interpersonal in nature. *Relatives* contains words like "mom", "family", "dad", and "sister", while *Social*'s top words were "friend" and "friends". Many of the other topics are indirectly interpersonal as well—*Neighborhood*, for example, contains words related to noise and space complaints, while *Individuals* is simply a collection of names (primarily female). This observation is hardly unexpected; the objective of AITA is to determine fault in interpersonal struggles. That being said, these findings confirm such topic bias and give credence to the idea that any shift in the subreddit around the onset of COVID-19 is likely to be one of volume rather than of topics.

5.1.2 RQ2: How similar are these topics across temporal groups? As discussed in Section 4.3.2, the top 10 topics appear to be fairly consistent between months. While these topics don't neatly match those extracted from the full data—the former consists of probability distributions extracted from each month's raw model, while the latter is a set of manual labels based on a TSNE-scaled pyLDAvis visualization (Figure 5)—their quantitative similarities over time do show that submissions on AITA have neither significantly shifted over time, nor been influenced by changing sociological factors. The only notable inconsistencies seem to reflect one or more shifts in the assigned numeric labels of various topics rather than their qualitative contents.

5.1.3 RQ3: What, if any, significant shifts in these topics occurred in early 2020? The data does not suggest that any shift in topics occurred in response to the COVID-19 pandemic. As shown by the analyses done in Section 4.3, the topics were fairly consistent across time with minor exceptions. Many of the trends observed (see Figure 7) seem instead to be a result of increased submission volume beginning in late 2018, which in turn increased the number and frequency of new words compared to earlier months. Future work might consider investigating the cause of this dramatic growth in the subreddit; though the topics being discussed didn't significantly change, the spike in user activity still might be in response to societal factors (such as the discord surrounding the 2020 US election,

for example). With respect to the data at hand, however, none of the words examined correlated with with pandemic (Figure 10).

5.2 Validity

There are a few features of our study that could limit the validity of our findings. First of all, as noted in Section 3.3, our choice of $K = 10$ may not produce topics that accurately reflect subject trends within AITA. We also direct our analysis in Section 4.3.2 based on the noted observation of low distance values between topics with the same number. While this was an interesting point of investigation, it assumes that there is a consistent assignment of topic numbers between two similar models. While this assumption does not appear to be out of line, a more nuanced analysis, as in Wang et. al, would use shortest-distance formulas to find common topics across time groups instead of relying on potentially arbitrary topic numbers. We encourage future work to take such measures to generate a more accurate representation of topic trends in the dataset.

5.3 Ethical Considerations

Proferes et al., among others, has previously raised important considerations with respect to Reddit data analysis. They note that institutional review boards (IRB) tend to classify scraped Reddit data as "public" and therefore exempt from consent requirements [8]. It's unsuprising, then, that fewer that 15% of papers the authors reviewed in their meta-analysis even mentioned ethics reviews at all.That being said, data like quotes and media that would require attribution in other contexts are subject to competing needs to both protect the anonymity of the author (a key feature of Reddit) and give proper credit. Researchers most commonly resolve this issue by attributing such data to the authors' Reddit usernames, although identifying information is still a concern given the frequently highly personal nature of users' activity and contributions. Our data is no exception, but given the abstract nature of our analysis and the purpose of AITA—that is, publicly getting strangers' opinions on personal conflict—and the abstract nature of our analyses, we believe that our work and results pose few, if any, ethical issues.

6 CONCLUSION

Despite its well-documented impact on societal systems and norms over the past two year, COVID-19's impact on individuals is still a subject of much analysis. While previous literature has established negative mental health effects of the pandemic—namely, an increase in anxiety, depression, and addiction, among other struggles—we wondered about it interpersonal consequences. Did working from home reduce office conflict? Did quarantine increase familial tensions? As a popular forum that features discussion about highly personal conflicts, r/AmITheAsshole poses a uniquely raw insight into everyday disagreements. In this paper, we take advantage of this resource to search for trends in conflict over the course of the last five years: specifically, which topics were commonly discussed; how consistent these topics were over time; and if there was any shift in these topics around the time that COVID-19 became a worldwide crisis. We find that the most common topics, which were generally interpersonal in nature, were quantitatively consistent over the time period studied and did not seem to undergo any significant shifts in response to the pandemic. While we are confident in these findings, we note that they are limited by the analyses performed. More precisely, our analysis does not seek to label or understand the topics as calculated, and we do not attempt to connect these topics over time to get a quantitative sense of their changes (Wang et al. [10] accomplishes this with a shortest-paths algorithm, but we do not employ such methods in this study). We suggest that future work take these additional steps to better understand the nuances of AITA's topics beyond their quantitative similarities, and that it take changes in subreddit popularity into account to rule out activity-related confounds.

REFERENCES

[1] Amanuel Alambo, Swati Padhee, Tanvi Banerjee, and Krishnaprasad Thirunarayan. 2021. COVID-19 and Mental Health/Substance Use Disorders on Reddit: A Longitudinal Study. In *Pattern Recognition. ICPR International Workshops and Challenges (Lecture Notes in Computer Science)*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.). Springer International Publishing, Cham, 20–27. https://doi.org/10.1007/978-3-030-68790-8_2

[2] Nicholas Botzer, Shawn Gu, and Tim Weninger. 2021. Analysis of Moral Judgement on Reddit. (Jan. 2021). http://arxiv.org/abs/2101.07664

[3] Emily Cannon, Bianca Crouse, Souvick Ghosh, Nicholas Rihn, and Kristen Chua. 2021. 'Don't Downvote A*******s!!': An Exploration of Reddit's Advice Communities. (2021). http://arxiv.org/abs/2109.09044

[4] Ion Stagkos Efstathiadis, Guilherme Paulino-Passos, and Francesca Toni. 2022. Explainable Patterns for Distinction and Prediction of Moral Judgement on Reddit. *J. ACM* (Jan. 2022). http://arxiv.org/abs/2201.11155

[5] Ethan Haworth, Ted Grover, Justin Langston, Ankush Patel, Joseph West, and Alex C Williams. 2021. Classifying Reasonability in Retellings of Personal Events Shared on Social Media: A Preliminary Case Study with r/AmITheAsshole. (2021).

[6] Andy J. Merolla, Christopher Otmar, and Carlos Ruvalcaba Hernandez. 2021. Day-to-day relational life during the COVID-19 pandemic: Linking mental health, daily relational experiences, and end-of-day outlook. *Journal of Social and Personal Relationships* 38, 8 (Aug. 2021), 2350–2375. https://doi.org/10.1177/02654075211020137 Publisher: SAGE Publications Ltd.

[7] Isabel Padrón, Isabel Fraga, Lucía Vieitez, Carlos Montes, and Estrella Romero. 2021. A Study on the Psychological Wound of COVID-19 in University Students. *Frontiers in Psychology* 12 (2021). https://www.frontiersin.org/article/10.3389/fpsyg.2021.589927

[8] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society* 7, 2 (June 2021). https://doi.org/10.1177/20563051211019004

[9] Parvathy Sarat, Rohit Mujumdar, Prathik Kaundinya, and Sahith Dambekodi. 2020. Can Machines Detect if you're a Jerk? (2020).

[10] Yu Wang, Conrad Hougen, Brandon Oselio, Walter Dempsey, and Alfred Hero. 2021. A Geometry-Driven Longitudinal Topic Model. *Harvard Data Science Review* 3, 2 (June 2021). https://doi.org/10.1162/99608f92.b447c07e

Annabelle Strong

CS 315 Final Project

Part One: Data Collection and Parsing

Table of Contents

1. Setup and Helper Functions
   A. Packages
   B. File Processing
   C. Data Synthesis
2. Main

Setup and Helper Functions

Packages

In [1]:
```
! pip install zstandard
! pip install pandas
! pip install tqdm
```

```
Requirement already satisfied: zstandard in /Users/annabelle_strong/miniforge3/lib/python3.9/site-packages (0.17.0)
Requirement already satisfied: pandas in /Users/annabelle_strong/miniforge3/lib/python3.9/site-packages (1.4.1)
Requirement already satisfied: numpy>=1.20.0 in /Users/annabelle_strong/miniforge3/lib/python3.9/site-packages (from pandas) (1.21.5)
Requirement already satisfied: python-dateutil>=2.8.1 in /Users/annabelle_strong/miniforge3/lib/python3.9/site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /Users/annabelle_strong/miniforge3/lib/python3.9/site-packages (from pandas) (2021.3)
Requirement already satisfied: six>=1.5 in /Users/annabelle_strong/miniforge3/lib/python3.9/site-packages (from python-dateutil>=2.8.1->pandas) (1.16.0)
Requirement already satisfied: tqdm in /Users/annabelle_strong/miniforge3/lib/python3.9/site-packages (4.63.0)
```

In [3]:
```
import os # for file management
import zstandard as zstd # enables parsing of .zst files
import json # for parsed entries
import pandas as pd # for data handling and storage
from tqdm.notebook import tqdm_notebook # displays progress bar for iterative processes
import datetime # for UTC parsing
```

File Processing

.zst files collected from PushShift archives

Based on the following file structure:

- *this script*
- Data/
  - Dumps/
    - *.zst files* (e.g. "2017-03.zst")
  - Extracted/
    - *filtered/cleaned .csv versions of each .zst file*
  - *compiled .csv dataset*

In [4]:
```
def getPath(year, month, extracted=False):
    '''
        Returns path to data file for the given month and year, depending on extracted status
    '''
    if month < 10:
        month = f'0{month}'

    if extracted:
        folder = "Extracted"
        version = "_extracted"
        extension = "csv"
    else:
        folder = "Dumps"
        version = ""
        extension = "zst"

    return os.path.join(f"Data/{folder}",f"{year}-{month}{version}.{extension}")

getPath(2017, 3)
```

Out[4]:
```
'Data/Dumps/2017-03.zst'
```

the following function uses code modified from here

In [5]:
```
def extractZst(path, excludeFn = lambda x: False):
    '''
        Returns extracted entries in given .zst file as a pandas DataFrame
    '''
    print(f'decompressing {path}...')
    with open(path, 'rb') as fh:
        # iterate through zst contents via a filestream to minimize memory load
        dctx = zstd.ZstdDecompressor(max_window_size=2147483648)
        with dctx.stream_reader(fh) as reader:
            previous_line = ""
            entries = []
            while True:
                chunk = reader.read(2**24)
                if not chunk:
                    break
                else:
                    string_data = chunk.decode('utf-8')
                    lines = string_data.split("\n")

                    for i, line in enumerate(lines[:-1]):
                        if i == 0:
                            line = previous_line + line
                        entry = json.loads(line)

                        if excludeFn(entry):
                            pass
                        else:
                            entries.append(entry)
                            print(entry['title'])

                    previous_line = lines[-1]

    return pd.DataFrame(entries)
```

In [6]:
```
def excludeFn(entry):
    '''
        Returns whether or not an entry should be discarded based on a set of conditional checks
    '''
    valueof = lambda key: entry.get(key, "")
    return ((valueof('selftext') in ['', '[deleted]', '[removed]'])\
            or (valueof('subreddit').lower() not in ['amitheasshole'])\
            or (+valueof('stickied')))

def postsFromFile(year, month, save=True):
    '''
        Constructs a dataframe of relevant entries from the given timeframe, saving the result to csv
    '''
    # use extracted data if possible
    extracted_path = getPath(year, month, True)
    if os.path.exists(extracted_path):
        print(f'loading {extracted_path}...')
        return pd.read_csv(extracted_path, low_memory=False)
    else: # otherwise decompress zst
        zst_path = getPath(year, month, False)
        if os.path.exists(zst_path):
            df = extractZst(zst_path, excludeFn) # replace with extractZst(zst_path) to keep all entries
            if not df.empty and save:
                df.to_csv(extracted_path, index=False)

            return df
        else:
            print(f'file {zst_path} not found')
```

Data Synthesis

In [7]:
```
def cleanDF(input_df, index_col="id"):
    '''
        Cleans given dataframe by removing unecessary columns and duplicate rows, then parsing UTC data
    '''

    df = input_df.infer_objects()

    # remove duplicates by index value then set as df index
    df.drop_duplicates(subset=[index_col], inplace=True)
    df.set_index(index_col, inplace=True)

    # remove columns with null values in at at least half the rows
    df.dropna(axis='columns', thresh=len(df)/2, inplace=True)

    # remove columns with 2 or fewer unique values
    nunique = df.nunique()
    cols_to_drop = nunique[nunique <= 2].index
    df.drop(cols_to_drop, axis=1, inplace=True)

    # turn UTC data into year/month/day columns
    df['datetime'] = df['created_utc'].apply(datetime.date.fromtimestamp)
    for col in ["year", "month", "day"]:
        df[col] = df['datetime'].apply(lambda x: getattr(x,col))

    return df
```

In [8]:
```
def getPostDF(month = "", start="2016-06", end="2021-06", save=True):
    '''
        Return specified dataframe by either loading from existing file or generating fresh data
    '''
    path = f"Data/{start}_{end}.csv"
    if os.path.exists(path):
        print(f"loading dataframe from Data/{start}_{end}.csv")
        df = cleanDF(pd.read_csv(path, low_memory=False))
    else:
        getYear = lambda x: int(x[:4])
        getMonth = lambda x: int(x[5:])

        if month:
            start = month
            end = month

        startY, endY = getYear(start), getYear(end)
        months = []
        # iterate through and record each month in the given range
        for y in range(startY, endY + 1):
            startM, endM = 1, 12
            if y == startY:
                startM = getMonth(start)
            if y == endY:
                endM = getMonth(end)

            for m in range(startM, endM + 1):
                months.append((y, m))

        df = pd.DataFrame()
        # contstruct dataframe for each month (tqdm prints progress bar)
        for monthTuple in tqdm_notebook(months):
            data = postsFromFile(*monthTuple) # retrieve data from given month's file
            df = pd.concat([df, pd.DataFrame(data)])

        if not df.empty:
            df = cleanDF(df) # filter out extraneous data
            if save or not os.path.exists(path):
                df.to_csv(path, index=True)
                print(f"saved dataframe to Data/{start}_{end}.csv")

    return df
```

Main

In [9]:
```
postData = getPostDF(start="2016-06", end="2021-06")
```

```
loading dataframe from Data/2016-06_2021-06.csv
```

In [11]:
```
postData
```

Out[11]:

id	created_utc	edited	link_flair_css_class	permalink	score	url	selftext	author	num_comments	thumbnail	...	aut
4m2vs0	1464807402	1464818412	ass	/r/AmItheAsshole/comments/4m2vs0/aita_for_read...	28	https://www.reddit.com/r/AmItheAsshole/comment...	Me and my boyfriend have been together for 1.5...	Aperson434	35	self	...	
4m4wm4	1464833447	False	not	/r/AmItheAsshole/comments/4m4wm4/aita_for_not_...	38	https://www.reddit.com/r/AmItheAsshole/comment...	Okay, so I'm back home for the summer (health ...	Dollop_of_Cinnamon	16	self	...	
4mblyv	1464935483	False	not	/r/AmItheAsshole/comments/4mblyv/aita_my_boyfr...	58	https://www.reddit.com/r/AmItheAsshole/comment...	So here's the scoop.\n\nMy boyfriend rented a ...	acoffeebinge	28	self	...	
4mlvos	1465099957	1465133917	not	/r/AmItheAsshole/comments/4mlvos/aita_my_girlf...	46	https://www.reddit.com/r/AmItheAsshole/comment...	Hopefully you can leave party affiliations asi...	abductee92	34	self	...	
4mm6qv	1465106118	False	ass	/r/AmItheAsshole/comments/4mm6qv/aita_for_thes...	15	https://www.reddit.com/r/AmItheAsshole/comment...	[First comment.] (https://www.reddit.com/r/Fem...	ProjectileSpider	11	self	...	
...	...	...	...	...	...	...	...	...	...	...	...	
ob9bd3	1625095890	False	not	/r/AmItheAsshole/comments/ob9bd3/aita_for_not_...	3	https://www.reddit.com/r/AmItheAsshole/comment...	So I was hanging with some friends in Discord,...	Calamitus0	3	self	...	
ob9iph	1625096598	1625156298.0	not	/r/AmItheAsshole/comments/ob9iph/aita_for_back...	15	https://www.reddit.com/r/AmItheAsshole/comment...	I (31F) was invited by my friend (31F) to a di...	rtsen	38	self	...	
ob9n7y	1625097039	False	not	/r/AmItheAsshole/comments/ob9n7y/aita_for_excl...	25	https://www.reddit.com/r/AmItheAsshole/comment...	Please forgive formatting, posting from mobile...	TenacityFarrington	36	self	...	
ob9ohs	1625097165	False	not	/r/AmItheAsshole/comments/ob9ohs/aita_for_kick...	881	https://www.reddit.com/r/AmItheAsshole/comment...	Hay, My(23 M) sister(15) had to lived with me ...	ThrowRaway74	198	self	...	
ob9pzw	1625097315	False	not	/r/AmItheAsshole/comments/ob9pzw/wibta_if_i_do...	26	https://www.reddit.com/r/AmItheAsshole/comment...	At Uni I made friends with a girl (5) who live...	immachode	31	self	...	

297442 rows × 33 columns

back to top

Annabelle Strong

CS 315 Final Project

Part Two: Data Exploration

Table of Contents

- Descriptive Statistics
- Feature Extraction and Graphing
  - Post Distribution
  - COVID-19 Status
  - Post Length
  - Comment Counts
  - Subscriber Growth

Load DataFrame

In [1]:
```
import pandas as pd
import os

print("\n\nLoading\"Data\", sep='\n')

_DB_Store
Backup
Extracted
Analysis
Archived
Archived
2014-06_2021-06.csv
2014_2017-12.csv
Store

start = "2014-06"
end = "2021-06"

pandas = pd.read_csv("Data/{start}_{end}.csv", low_memory=False, set_index='id').infer_objects()
pandas.
```

Out[1]:

	created_utc	edited	link_flair_css_class	permalink	score	url	selftext	author	num_comments	thumbnail	...	eat
id												
4m2ced	1466879422	1466879912		aes	/r/AmItheAsshole/comments/4m2ced/aita_for_read...	28	https://www.reddit.com/r/AmItheAsshole/comment...	Ajaccuss434	35	self	...	
4m4aad	1466933447	False		nsf	/r/AmItheAsshole/comments/4m4aad/aita_for_not_...	39	https://www.reddit.com/r/AmItheAsshole/comment...	Dollop_of_Cinnamon	16	self	...	
4m8dqy	1466935483	False		nsf	/r/AmItheAsshole/comments/4m8dqy/aita_my_boyfr...	58	https://www.reddit.com/r/AmItheAsshole/comment...	acoffeedrinker	26	self	...	
4m6vea	1466939907	1466972397		nsf	/r/AmItheAsshole/comments/4m6vea/aita_my_girlf...	46	https://www.reddit.com/r/AmItheAsshole/comment...	abductee92	34	self	...	
4m4dqo	1466936718	False		aes	/r/AmItheAsshole/comments/4m4dqo/aita_for_thes...	15	https://www.reddit.com/r/AmItheAsshole/comment...	ProjectileSpider	11	self	...	
...	...	...	...	...	...	...	...	...	...	...	...	
o8d6a0	1623509360	False		nsf	/r/AmItheAsshole/comments/o8d6a0/aita_for_not_...	3	https://www.reddit.com/r/AmItheAsshole/comment...	Calentus0	3	self	...	
o8h3jn	1623509938	1623509235.0		nsf	/r/AmItheAsshole/comments/o8h3jn/aita_for_back...	15	https://www.reddit.com/r/AmItheAsshole/comment...	rhen	38	self	...	
o8ldcy	1623509739	False		nsf	/r/AmItheAsshole/comments/o8ldcy/aita_for_excl...	25	https://www.reddit.com/r/AmItheAsshole/comment...	Please forgive formatting, posting from mobile...	TenacityFarrington	36	self	...	
o8ldhk	1623509765	False		nsf	/r/AmItheAsshole/comments/o8ldhk/aita_for_kick...	891	https://www.reddit.com/r/AmItheAsshole/comment...	Hey, MILES WE KNEW!!!!!! had to meet with me ...	ThrowAway74	168	self	...	
o8lqpw	1623509795	False		nsf	/r/AmItheAsshole/comments/o8lqpw/aita_if_i_do...	26	https://www.reddit.com/r/AmItheAsshole/comment...	At uni I made friends with a girl (3) and low...	immachode	31	self	...	

297642 rows × 33 columns

In [2]:
```
pandas.columns
```

Out[2]:
```
Index(['created_utc', 'edited', 'link_flair_css_class', 'permalink', 'score',
       'url', 'selftext', 'author', 'num_comments', 'thumbnail',
       'retrieved_on', 'gilded', 'link_flair_text', 'title', 'suggested_sort',
       'distinguished', 'stickied', 'post_hint', 'is_self', 'is_crosspost',
       'subreddit_subscribers', 'link_flair_background_color',
       'link_flair_richtext', 'link_flair_template_id', 'ale',
       'total_awards_received', 'num_crossposts', 'is_awarded', 'ja',
       'total_awards_received', 'treatment_tag', 'datetime', 'year', 'month',
       'day'],
      dtype='object')
```

Descriptive Statistics

In [3]:
```
pandas[['num_comments', 'subreddit_subscribers', 'year', 'month', 'day'].describe()
```

Out[3]:

	num_comments	subreddit_subscribers	year	month	day
count	297642.000000	2.948610e+05	297642.000000	297642.000000	297642.000000
mean	72.980882	1.384884e+06	2019.576457	6.401288	15.850099
std	295.576968	6.560701e+05	0.775587	3.369811	8.830935
min	0.000000	5.040700e+04	2016.000000	1.000000	1.000000
25%	10.000000	1.010600e+06	2019.000000	4.000000	8.000000
50%	21.000000	1.874867e+06	2020.000000	6.000000	16.000000
75%	41.000000	2.174600e+06	2020.000000	9.000000	24.000000
max	99993.000000	2.897768e+06	2021.000000	12.000000	31.000000

In [4]:
```
pandas['author'].nunique()
```

Out[4]:
```
241728
```

In [5]:
```
pandas['datetime'].nunique()
```

Out[5]:
```
1794
```

Feature Extraction and Graphing

In [6]:
```
import numpy as np
import plotly.express as px
import plotly
plotly.offline.init_notebook_mode(connected=True)
```

Post Distribution

In [7]:
```
px.histogram(pandas, x="month", color="year", title="Post count over time")
```

Out[7]:

In [8]:
```
px.histogram(pandas, x="day", color="month", title="Post count over time")
```

Out[8]:

COVID-19 Status

In [9]:
```
def inforoutstate(year, month=1, day=1, df=pandas):
    inforoutstate = (df['year'] > year)
    inforout = (df['year'] == year)
    inforoutstate = inforoutstate | (inforout
    inforout = (df['month'] == month)
    inforoutstate = (df['month'] > month)
    inforoutstate = (df['day'] > day)

    return inforoutstate | (inforout & inforoutstate)

pandas['C-19'] = inforoutstate(2020, 3, 15)
pandas[['C-19', 'datetime']]
```

Out[9]:

	C-19	datetime
id		
4m2ced	True	2016-06-01
4m4aad	True	2016-06-01
4m8dqy	True	2016-06-03
4m6vea	True	2016-06-05
4m4dqo	True	2016-06-05
...	...	...
o8ldcy	False	2021-06-00
o8ldhk	False	2021-06-00
o8ldcy	False	2021-06-00
o8ldhk	False	2021-06-00
o8lqpw	False	2021-06-00

297642 rows × 2 columns

In [10]:
```
px.histogram(pandas, x="datetime", color="C-19", title="Post count over time")
```

Out[10]:

Post Length

In [11]:
```
pandas['char_count'] = pandas['selftext'].apply(lambda
pandas.sort_values(by='char_count', ascending=False)
```

Out[11]:

	created_utc	edited	link_flair_css_class	permalink	score	url	selftext	author	num_comments	thumbnail	...	gildings	all_awa
id													
aqz58k	1550358451	False		aes	/r/AmItheAsshole/comments/aqz58k/aita_for_dis...	1	https://www.reddit.com/r/AmItheAsshole/comment...	a6r58	31	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}	
ae28c4	1547405371	1547405555		nsf	/r/AmItheAsshole/comments/ae28c4/aita_for_blow...	1	https://www.reddit.com/r/AmItheAsshole/comment...	deepstoli_wasdisabled	15	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}	
8jcuka	1531629601	False		nsf	/r/AmItheAsshole/comments/8jcuka/aita_for_comp...	0	https://www.reddit.com/r/AmItheAsshole/comment...	hammer1ar12	18	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}	Nsfw
amuu8i	1546948857	1546949998		aes	/r/AmItheAsshole/comments/amuu8i/aita_for_gett...	1	https://www.reddit.com/r/AmItheAsshole/comment...	songsharking	35	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}	
5cvxgd	1543030649	1543304969		aes	/r/AmItheAsshole/comments/5cvxgd/aita_in_this_...	0	https://www.reddit.com/r/AmItheAsshole/comment...	**Text cut and Dynamic****Updated**: It's been a y...	NDKobaneProcessing	23	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}
...	...	...	...	...	...	...	...	...	...	...	...	...	...
j2euuf	1602372608	1602741243.0	Nsfw	/r/AmItheAsshole/comments/j2euuf/aita_for_dis...	0	https://www.reddit.com/r/AmItheAsshole/comment...	[deleted]	[deleted]	33	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}	
ohdqgc	1577705333	1577717872		nsf	/r/AmItheAsshole/comments/ohdqgc/aita_for_pow...	0	https://www.reddit.com/r/AmItheAsshole/comment...	Cake-Quake	14	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}	
dmdcqg	1572597664	1572126996		nsf	/r/AmItheAsshole/comments/dmdcqg/aita_for_brin...	0	https://www.reddit.com/r/AmItheAsshole/comment...	[deleted]	43	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}	
dtyl04	1573327667	1573538754		aes	/r/AmItheAsshole/comments/dtyl04/aita_for_not_...	0	https://www.reddit.com/r/AmItheAsshole/comment...	Throwaway1735553284	11	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}	
e2e6r1	1578923662	1620593870		nsf	/r/AmItheAsshole/comments/e2e6r1/aita_if_i_bo...	0	https://www.reddit.com/r/AmItheAsshole/comment...	9	lalalea123	26	self	...	{'gold': 0, 'silver': 0, 'platinum': 0}

297642 rows × 35 columns

In [12]:
```
px.histogram(pandas, x=pandas['char_count'].apply(np.log), title='Frequency of post lengths')
```

Out[12]:

In [13]:
```
import numpy as np
px.line(pandas, x="datetime", y=pandas['char_count'].apply(np.log), title="Post length over time")
```

Out[13]:

Comment Counts

In [14]:
```
pandas.sort_values(by='num_comments', ascending=False)
```

Out[14]:

	created_utc	edited	link_flair_css_class	permalink	score	url	selftext	author	num_comments	thumbnail	...	gildings	all_awardi
id													
gp07c	1591020054	False	Nsfw	/r/AmItheAsshole/comments/gp07c/introducing_m...	236	https://www.reddit.com/r/AmItheAsshole/comment...	Welcome to r/AmItheAsshole! It's open-forum. Post your stories here! ...	SnacsaysFast	99993	self	...	{}	
ddb8m	1569838334	1569879601		aes	/r/AmItheAsshole/comments/ddb8m/aita_for_refu...	26764	https://www.reddit.com/r/AmItheAsshole/comment...	farquaarthur	11989	self	...	{'gold': 5}	{'award_sub_id': '5L2G', 'award_type':
ed8cpw	1576843739	1576906260		aes	/r/AmItheAsshole/comments/ed8cpw/aita_for_feel...	23227	https://www.reddit.com/r/AmItheAsshole/comment...	Throwaway38240128	9163	self	...	{'gold': 7}	{'award_sub_id': 1S2W, 'award_type':
ebjcvf	1577607774	False		nsf	/r/AmItheAsshole/comments/ebjcvf/aita_for_upset...	39874	https://www.reddit.com/r/AmItheAsshole/comment...	HENRYDORN	9002	self	...	{'gold': 2, 'silver': 2}	{'award_sub_id': 1S2W, 'award_type':
cw4bxc	1566919725	False		nsf	/r/AmItheAsshole/comments/cw4bxc/aita_for_refu...	33899	https://www.reddit.com/r/AmItheAsshole/comment...	kanas	8858	self	...	{'gold': 1, 'silver': 1}	{'award_sub_id': 1S2W, 'award_type':
...	...	...	...	...	...	...	...	...	...	...	...	...	...
fu9qqy	1586050994	False	Nsfw	/r/AmItheAsshole/comments/fu9qqy/aita_for_refu...	1	https://www.reddit.com/r/AmItheAsshole/comment...	(Sorry if I do some mistakes in English, I am ...	Nemo00000	0	self	...	{}	
bwr9ec	1559667390	False	Nsfw	/r/AmItheAsshole/comments/bwr9ec/aita_for_sent...	1	https://www.reddit.com/r/AmItheAsshole/comment...	Hi everyone, a friend of mine asked me to post this for her. ...	thisisitthen	0	self	...	{}	
cxqoef	1567259508	False	Nsfw	/r/AmItheAsshole/comments/cxqoef/aita_for_refu...	1	https://www.reddit.com/r/AmItheAsshole/comment...	Posting here because I'm confused, confused and I have no ...	tryn00	0	self	...	{}	
euq842	1580664009	False	Nsfw	/r/AmItheAsshole/comments/euq842/aita_for_tell...	1	https://www.reddit.com/r/AmItheAsshole/comment...	Some min debief on Mydiet. Hello and thanks for ha...	captainsmurfy87	0	self	...	{}	
dbr5q	1569935391	False	Nsfw	/r/AmItheAsshole/comments/dbr5q/aita_for_not_...	1	https://www.reddit.com/r/AmItheAsshole/comment...	so some backstory: my dad moved away in a no ...	justonecharacterss	0	self	...	{}	

297642 rows × 35 columns

In [15]:
```
# Post "guilty" in large open forum and show data (has roughly 100k comments)
px.line(pandas.drop('guilty'), x="datetime", y="num_comments", title="Post count over time")
```

Out[15]:

Subscriber Growth

In [16]:
```
px.line(pandas, x="datetime", y="subreddit_subscribers", title="Growth in subscribers over time")
```

Out[16]:

go to top

Annabelle Erving

CS 375 Final Project

Part Three: Data Analysis

Table of Contents

- Setup
  - Install Packages
  - Load DataFrame
- Quantitative Analysis
  - Similarity Computation
  - Model Similarity LDA Pairwise
  - Model Generation
- Analyze Collective Corpus
  - Vocabulary Extraction and Topic Modeling
  - Data Grouping and Processing
  - Temporal Normalization
  - Grouping
  - Smoothing
- Analyze Individual Corpus
  - Vocabulary Extraction
  - LDA Topic Modeling
  - Data Grouping and Processing
  - 3-1 Temporal Normalization
  - 3-2 Grouping
  - 3-3 Smoothing
- Cross Corpus Analysis
- Stylometry Analysis
  - 4-1 Vocabulary Extraction
  - 4-2 LDA Topic Modeling
  - 4-3 Temporal Normalization
  - 4-4 Cross Corpus Comparisons
- Sentiment Analysis
  - 5-1 Sentiment Analysis
  - 5-2 Temporal Normalization
- Conclusion

Setup

Install Packages

In [1]: `%pip install -q -U tqdm numpyro[cuda]`

In [2]: `import os, sys, gc, pickle, datetime, glob, collections`
`import numpy as np, pandas as pd, matplotlib, matplotlib.pyplot as plt, seaborn as sns`

In [3]: `from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer # for vectorization`
`from sklearn.decomposition import LatentDirichletAllocation # for topic modeling`

In [4]: `# for topic coherence`
`import gensim, gensim.corpora, gensim.models`
`from gensim.models.coherencemodel import CoherenceModel`
`import numpy as np, scipy.stats, scipy.spatial.distance`

In [5]: `# for text preprocessing`
`import re, nltk, spacy, string, unicodedata, contractions, inflect, markupsafe`
`from nltk.corpus import stopwords, wordnet, gutenberg, brown, inaugural, webtext, nps_chat`
`from nltk.tokenize import word_tokenize, sent_tokenize, TreebankWordTokenizer, TweetTokenizer, MWETokenizer`
`from nltk.stem import PorterStemmer, LancasterStemmer, WordNetLemmatizer, SnowballStemmer`
`import bs4, lxml, pytz, pytz.gb, pytz.utc, dateutil, dateutil.parser, dateutil.tz, dateutil.relativedelta, tqdm, tqdm.auto, tqdm.notebook`

In [6]: `# for models`
`import torch, torchvision, torchaudio, numpyro, numpyro.distributions, numpyro.infer, numpyro.optim, numpyro.contrib.tfp, numpyro.contrib.callbacks, numpyro.ext.optimizers, numpyro.ext.vi, numpyro.ext.vae, numpyro.ext.gp, numpyro.ext.ar, numpyro.ext.rnn, numpyro.ext.dmm, numpyro.ext.iaf, numpyro.ext.nsf, numpyro.ext.flows, numpyro.ext.distributions, numpyro.ext.util, numpyro.ext.optim, numpyro.ext.infer, numpyro.ext.handlers, numpyro.ext.contrib, numpyro.ext.callbacks, numpyro.ext.ops, numpyro.ext.nn, numpyro.ext.stochastic, numpyro.ext.mcmc, numpyro.ext.vi, numpyro.ext.ar, numpyro.ext.rnn, numpyro.ext.dmm, numpyro.ext.iaf, numpyro.ext.nsf, numpyro.ext.flows, numpyro.ext.distributions, numpyro.ext.util, numpyro.ext.optim, numpyro.ext.infer, numpyro.ext.handlers, numpyro.ext.contrib, numpyro.ext.callbacks, numpyro.ext.ops, numpyro.ext.nn, numpyro.ext.stochastic, numpyro.ext.mcmc`

In [7]: `# for tokenizers`
`import tokenizers, transformers, sentencepiece, tiktoken`

In [8]: `# for text processing`
`import ftfy, unidecode, textwrap, textstat, textblob, pattern, spacy_langdetect, spacy_lookups_data, spacy_transformers, spacy_sentiws, spacy_wordnet, spacy_syllables, spacy_sentiment, spacy_spell, spacy_lemmatizer, spacy_stopwords, spacy_ngrams, spacy_phrases, spacy_pos_tagger, spacy_dep_parser, spacy_ner, spacy_tokenizer, spacy_vectors, spacy_similarity, spacy_matcher, spacy_phrase_matcher, spacy_entity_linker

```
Topic 191 is a 0.070*"pillow" + 0.070*"cushion" + 0.070*"cushions" + 0.070*"pillows" + 0.070*"throw" + 0.070*"decorative" + 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics"
Topic 192 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 193 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 194 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 195 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 196 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 197 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 198 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 199 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 200 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 201 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 202 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 203 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 204 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 205 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 206 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 207 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 208 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 209 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 210 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 211 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 212 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 213 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 214 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 215 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 216 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 217 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 218 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 219 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 220 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 221 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 222 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 223 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 224 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 225 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 226 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 227 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 228 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 229 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 230 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 231 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 232 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 233 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 234 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 235 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 236 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 237 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 238 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 239 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 240 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 241 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 242 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 243 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 244 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 245 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 246 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 247 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 248 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 249 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 250 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 251 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 252 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 253 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 254 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 255 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 256 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 257 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 258 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 259 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 260 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 261 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 262 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 263 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 264 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 265 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 266 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 267 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 268 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 269 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 270 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 271 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 272 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 273 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 274 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 275 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 276 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 277 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 278 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 279 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 280 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 281 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 282 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 283 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 284 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 285 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 286 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 287 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 288 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 289 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 290 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 291 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 292 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 293 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 294 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 295 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 296 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 297 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 298 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
Topic 299 is a 0.070*"cover" + 0.070*"covers" + 0.070*"fabric" + 0.070*"fabrics" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton" + 0.070*"cotton"
```

```
Topic 1 is a deprecated alias for the series "area". To silence this warning, use "area" by itself. Using "area" will default to a stacked area plot.
Topic 2 is a deprecated alias for the series "area". To silence this warning, use "area" by itself. Using "area" will default to a stacked area plot.
DeprecationWarning: Ax.text(...) The text() argument of matplotlib.offsetbox.TextArea() is deprecated. Use set_text() instead.
/tmp/ipykernel_1_0/tmp_0ln7k2ln.png:127:DeprecationWarning: Ax.text(...) The text() argument of matplotlib.offsetbox.TextArea() is deprecated. Use set_text() instead.
/tmp/ipykernel_1_0/tmp_0ln7k2ln.png:127:DeprecationWarning: Ax.text(...) The text() argument of matplotlib.offsetbox.TextArea() is deprecated. Use set_text() instead.
/tmp/ipykernel_1_0/tmp_0ln7k2ln.png:127:DeprecationWarning: Ax.text(...) The text() argument of matplotlib.offsetbox.TextArea() is deprecated. Use set_text() instead.
```

Intra-Topic Consistency

Calculating symmetrical cosine, sum, and standard deviations hellinger distance for each topic with respect to each month: essentially flattening the spots of color from the figure above and showing these with the same topic number as a group.

```
def plotsimpleheatmap(dictpatdfbyTS, quantfilby month, lbmonthdata):
    ''' Plot a simple heatmap of the hellinger distance between cosine topics and return firstmost
        list and of median results as a dictionary
    '''
    lbmonthdata = []
    for TS in range(0,len(dictpatdfbyTS.keys())):
        mn = dictpatdfbyTS[TS].groupby(quantfilby).mean()
        lbmonthdata.extend([mn[1]])
        firstmost.extend([TS,mn[1],mn[0]])
        print("Topic",TS,"is a",mn[0],"*",mn[1])
    print("firstmost",firstmost)
    print("lbmonthdata",lbmonthdata)

    ncols = lenset(firstmost[0:10])

    cm.rcparams['figure.dpi'] = 72
    plt.figure(figsize=(len(lbmonthdata),len(set(firstmost[0:10]))))
    ax = sns.heatmap(np.array([firstmost[0:10]*len(set(firstmost[0:10]))*len(lbmonthdata),
                               lbmonthdata,
                               np.ravel(np.transpose
```

```
(gdb) ls
1693\  ldap_free_sasldata(cred);
1694   ldap_msgfree(result);
1695   ldap_msgfree(msg);
1696   ldap_unbind(ld);
1697   ldap_memfree(base);
1698   ldap_memfree(filter);
1699   ldap_memfree(dn);
1700   ldap_memfree(pwattr);
1701   ldap_memfree(passwd);
1702   ldap_memfree(bindpw);
1703   ldap_memfree(binddn);
1704   ldap_memfree(uri);
1705   ldap_memfree(host);
1706   ldap_memfree(sasldata);
1707   ldap_memfree(sasldn);
1708   ldap_memfree(saslmech);
1709   ldap_memfree(ldapdb);
1710   ldap_memfree(ldapuri);
1711   ldap_memfree(extop);
1712   ldap_memfree(exvalue);
1713   ldap_memfree(rebindproc);
1714   ldap_memfree(matched);
1715   ldap_memfree(text);
1716   ldap_memfree(retoid);
1717   ldap_memfree(retdata);
1718   ldap_memfree(newhost);
1719   ldap_memfree(errbuf);
1720   ldap_memfree(outbuf);
1721   ldap_memfree(inbuf);
1722   ldap_memfree(lcup_ldapdn);
1723   ldap_memfree(lcup_ldapurl);
1724   ldap_memfree(lcup_ldapmech);
1725   ldap_memfree(lcup_ldapcreds);
1726   ldap_memfree(lcup_ldapbase);
1727   ldap_memfree(lcup_ldapfilter);
1728   ldap_memfree(lcup_ldapattrs);
1729   ldap_memfree(lcup_ldapscope);
1730   ldap_memfree(lcup_ldapsrchtmpl);
1731   ldap_memfree(lcup_ldapsrchattrs);
1732   ldap_memfree(lcup_ldapsrchscope);
1733   ldap_memfree(lcup_ldapsrchfilter);
1734   ldap_memfree(lcup_ldapsrchbase);
1735   ldap_memfree(lcup_ldapsrchdn);
1736   ldap_memfree(lcup_ldapsrchurl);
1737   ldap_memfree(lcup_ldapsrchmech);
1738   ldap_memfree(lcup_ldapsrchcreds);
1739   ldap_memfree(lcup_ldapsrchpwattr);
1740   ldap_memfree(lcup_ldapsrchpwdn);
1741   ldap_memfree(lcup_ldapsrchpwurl);
1742   ldap_memfree(lcup_ldapsrchpwmech);
1743   ldap_memfree(lcup_ldapsrchpwcreds);
1744   ldap_memfree(lcup_ldapsrchpwattr);
1745   ldap_memfree(lcup_ldapsrchpwdn);
1746   ldap_memfree(lcup_ldapsrchpwurl);
1747   ldap_memfree(lcup_ldapsrchpwmech);
1748   ldap_memfree(lcup_ldapsrchpwcreds);
1749   ldap_memfree(lcup_ldapsrchpwattr);
1750   ldap_memfree(lcup_ldapsrchpwdn);
1751   ldap_memfree(lcup_ldapsrchpwurl);
1752   ldap_memfree(lcup_ldapsrchpwmech);
1753   ldap_memfree(lcup_ldapsrchpwcreds);
1754   ldap_memfree(lcup_ldapsrchpwattr);
1755   ldap_memfree(lcup_ldapsrchpwdn);
1756   ldap_memfree(lcup_ldapsrchpwurl);
1757   ldap_memfree(lcup_ldapsrchpwmech);
1758   ldap_memfree(lcup_ldapsrchpwcreds);
```