May 2, 2021

Kelly Chen, Sara Kenefick, Rebekah Varghese
Homework 5–Final Project

Hypotheses

1. As time progresses, the frequency of articles that address mental health will increase. In other words, more recent years will have a greater number of articles pertaining to mental health.
2. Considering the time scale of semesters, the frequency of articles that address mental health will be highest around mid-semester, when classes are most busy.
3. Comparing the content of the articles, the articles that will be the most similar are articles written by the same person, since they will likely use similar language and address similar topics.

Data Set

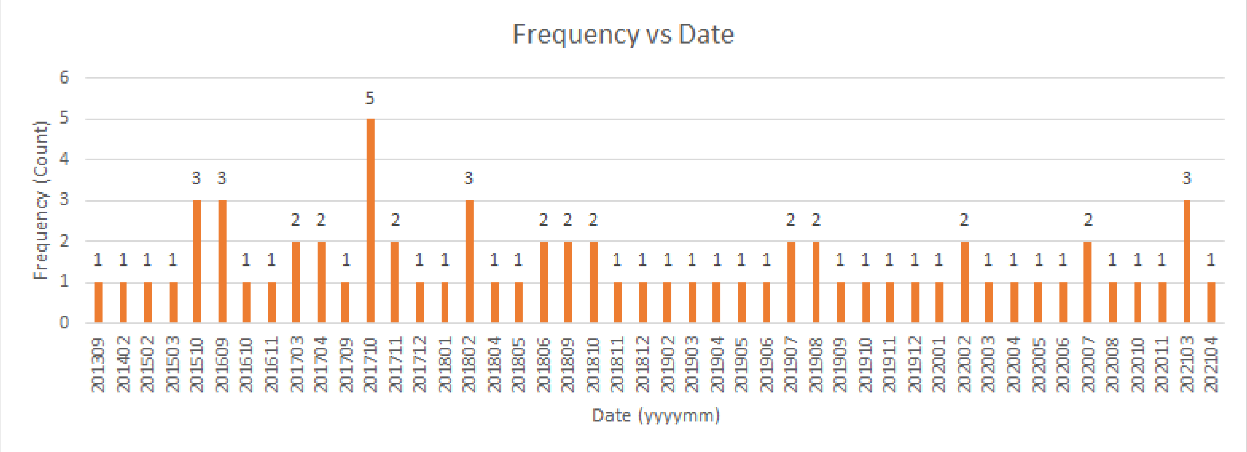
The Daily Pennsylvanian (1500 most recent Opinion Column articles):
https://www.thedp.com/section/columns?page=1&per_page=1500

Table 1: Cosine similarity values of top 10 most similar article pairings

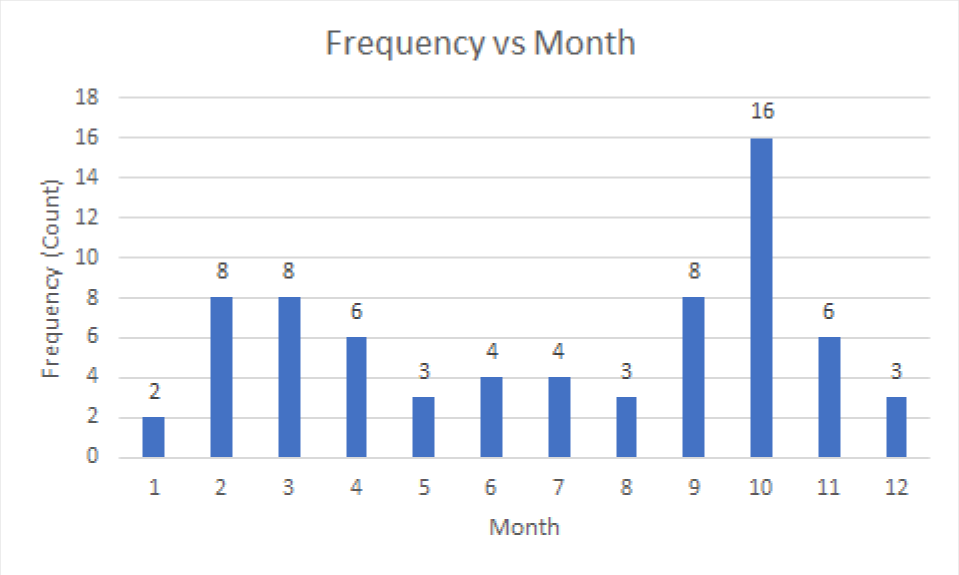
Cosine Similarity	**Article 1**	**Article 2**
0.3276	*The need to destigmatize therapy* by Calvary Rogers, 10/2017	*I went to CAPS and you should too* by Emilia Onuonga, 2/2020
0.3075	*Penn students don't just want longer breaks. They need them.* by Carlos Arias Vivas, 12/2017	*Penn must give us more breaks* by Alfredo Praticò, 11/2019
0.2934	*Addressing the anxieties of the transition to college* by Isabella Simonetti, 10/2017	*What I want from the 'Campus Conversation'* by Isabella Simonetti, 10/2017
0.2732	*The 'work hard, play harder' mentality is deeply flawed* by Emilia Onuonga, 12/2019	*Stop using mental health as an excuse to party* by Bridget Yu, 3/2021
0.2715	*The need to destigmatize therapy* by Calvary Rogers 10/2017	*Take advantage of the support groups that CAPS has to offer* by Bridget Yu, 7/2019
0.2607	*Take time to understand eating disorders this week* by Sophia DuRose, 2/2020	*The reality of eating disorders during quarantine* by Bridget Yu, 7/2020
0.2384	*There's nothing wrong with prioritizing our mental health* by Bridget Yu, 6/2019	*OCD is not a personality quirk* by Bridget Yu, 8/2020

0.2376	*Take advantage of the support groups that CAPS has to offer* by Bridget Yu, 7/2019	*I went to CAPS and you should too* by Emilia Onuonga, 2/2020
0.2267	*Pro-NSO; Pro-ANA* by Harrison Glicklich, 9/2016	*The reality of eating disorders during quarantine* by Bridget Yu, 7/2020
0.2228	*Closing Huntsman early is not a ridiculous idea* by Jay Shah, 9/2018	*How Penn can make midterms and finals less stressful* by Jay Shah, 10/2018

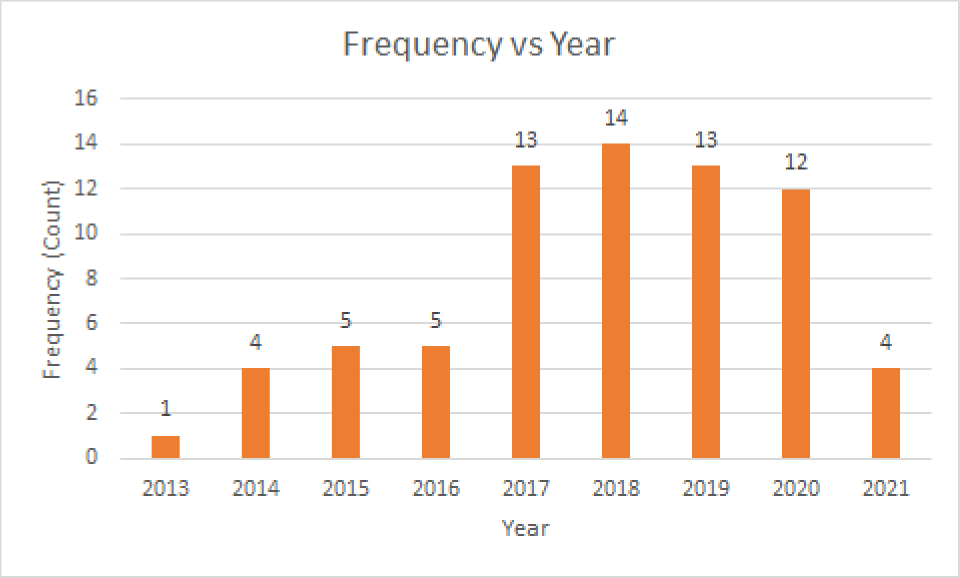
Graph 1: Graph showing all of the frequencies of each month/year in the yyyymm format



Graph 2: Graph showing all of the frequencies of each month



Graph 3: Graph showing all of the frequencies per year



Analysis

The motivation behind this research project was to examine the conversation around mental health at Penn. As the opinion columns of the Daily Pennsylvanian provided a solid representation of the student body's interests, preferences, and concerns, from this data we hoped to glean insight regarding Penn's openness on the subject. We reasoned that the more frequently words such as *anxiety*, *depression*, and *wellness* appeared in the headlines, the more mental health was a topic of discussion. Increasing frequency would indicate a diminishing stigma around mental disorders and increased awareness of the students' well-being.

Regarding hypothesis 1, there was a clear increase in the frequency of articles containing mental health-related words from 2014 to 2020. Our hypothesis is therefore supported by the data collected although not in the way we originally expected. Instead of a steady increase in the number of articles relating to mental health, 2016-2017 marked a turning point, where the article count spiked. After more than doubling, the frequency remains constant from 2017 to the present. This trend could be explained by the unrest caused by the election or tragedies that took place closer to home, including seven suicides on Penn's campus between 2013-2015. Our results suggest that these events brought to light the massive problem of mental health among Penn students, which we can only hope is being addressed. This data offers encouraging evidence that Penn students are talking more about mental health, dismantling "Penn Face," and expanding resources to help people struggling with mental health issues.

For hypothesis 2, there is a greater frequency of articles containing mental health-related words around September-November during the fall semester and around February-April during the spring semester. We originally hypothesized that the greatest frequency would be found mid-semester, as it allows students to settle into the semester a little bit, after the first round of midterms. Therefore, our data supports this hypothesis as the months of February through April have either 6 or 8 articles, compared to 4 or less for the spring semester. In addition, the months of September through November have greater than 6 articles each, compared to 3 or less for the rest of the fall semester. It is also interesting to note that October has a significantly greater frequency of 16 than the other months. One potential reason for this may be that October 2017 was a very heavy time in terms of mental health-related topics as at least 3 students died by October of that semester. This is also highlighted in Graph 1, where 201710 has the greatest frequency of 5. This may have boosted the total for October as a whole. In general, this data highlights the key areas of the semester where more mental health awareness is needed, leading the DP to cover it at a greater frequency. We can see that this is very cyclical and is a recurring theme throughout the years. This is encouraging as the DP is providing support and awareness at times where it is likely most beneficial and can have a greater impact.

For hypothesis 3, every relevant article that was collected from the DP columns page was compared using the vector space model and cosine similarity. We hypothesized that articles written by the same author would have the highest cosine similarity, as each writer has their own particular style, and will usually write about similar topics. The ten pairs of articles that had the highest cosine similarity are shown in Table 1 above. As displayed in the table, our initial hypothesis was not supported by the data, as for the most part, the most similar articles were not written by the same columnist. Instead, the trend that we found was that similar articles were centered around closely related topics, which in turn would require similar vocabulary. For example, the most similar articles, *The need to destigmatize therapy* and *I went to CAPS and you should too*, both advocate for students to seek therapy and help for their mental health issues if they need it. Similarly, the next two articles on the list both tackle Penn's lack of sufficient break days for students to rest and recover from the toll of a semester. There are some pairings that share the same author, but these articles cover related topics, in addition to being written by the same person. These results of the top 10 most similar pairings are also quite telling of what topics Penn students are passionate about when it comes to mental health—many of them emphasize the importance of destigmatizing therapy and encouraging students to seek the help they need. The topic of eating disorders is also commonly written about, both from back in 2016 to more recently in 2020, when talking about how quarantine has affected people's perception of themselves. We can see that many of the same topics that were important several years ago are still priorities for students now, and that campus conversation about mental health has expanded and built upon conversations sparked in the past.

Summary.txt File

Names: Kelly Chen, Sara Kenefick, Rebekah Varghese

Description: Scaping the headlines and subheadings of opinion columns from The Daily Pennsylvanian, we counted the frequency of articles containing words related to mental health published each year from 2014-2020. From this data we hoped to conclude that the discussion around mental health and mental disorders at Penn has increased in the recent years. Additional analysis pointed to more conversation about this subject during high-stress time frames (i.e. midterms/finals). Finally, comparing the cosine similarities between pairs of articles indicated similar topics rather than the same author contributed to higher cosine similarity.

Categories: physical networks, information networks, document search

Work Breakdown:
Sara- wrote the java program to retrieve the articles from The Daily Pennsylvanian. This step involved crawling the web page, scraping the headlines/subheadings, and extracting the desired information (i.e. title, date, caption, url). I implemented JSoup to parse the Document, and I used regex to only add the relevant articles to the array list.

Rebekah - Used articles collected by Sara to implement the Updates class, which created TreeMaps for each of the 3 categories (date, month, year) and updated their frequencies appropriately. Created all of the graphs in Excel and wrote the hypotheses section as well as hypothesis section 2 of the report.

Kelly - Used the articles collected by Sara and extracted the article text using JSoup/Regex, and wrote each of the articles to a .txt file to be used in the Vector Space Model. Altered the Vector Space Model in order to compare each of the written .txt files and rank their cosine similarities.
Wrote the hypothesis 3 section of the report and provided the data for Table 1.

Running the code:
Our Main method contains all three of our parts: (1) extracting the relevant articles, (2) updating the frequencies for the graphs, and (3) using the vector space model. The print statements for (1) and (3) are currently commented out, but can be uncommented to view the data that they print.