Augustin Liu, Rohit Jagga, Sampanna Bhatta
May 4th, 2022
NETS-150
Dr. Swapneel Sheth

M.A.P. (Mental Health & Academics at Penn)

I. Introduction & Hypotheses

In this report, we wanted to analyze the relationship between CAPS/mental health and academics, as represented by Reddit posts in the r/UPenn subreddit[1]. Our motivation for this is that given final exams are currently underway, we suspected that there might be some relationship between CAPS-related posts and academic-related posts that we could find evidence of in the subreddit. To investigate this relationship, we considered the following hypotheses:

1. We hypothesize that the posts in the "Mental Health" flair category of Penn subreddit will be the most similar to a query list of commonly referenced words that represent sadness and depression, followed by the "Academic" flair of Penn subreddit category, followed by the general Penn subreddit.
    - Note: the general Penn subreddit is represented by the "Hot" section on Penn Reddit's homepage), followed by Reddit in general (represented by the "Popular" section on Reddit's homepage).
2. We hypothesize that academic posts should have more points and higher upvote ratio on average than CAPS-related posts since we are currently in the final examination season, which we believe may garner more attention to academic posts.
3. We hypothesize that posts with positive sentiment will have more points and higher upvote ratios than those with neutral sentiment. In addition, we hypothesize that posts with neutral sentiment would garner more points and higher upvote ratios than posts with negative sentiment. The justification for these hypotheses is that we believe people would like and tend to upvote posts of positive nature.
4. We hypothesize that in the affiliation network graph consisting of users and post categories (CAPS, academic), the network exhibits membership closure in that the users who have made a CAPS/mental health-related post will be more likely to also post about academics, and vice versa.

II. Dataset

Data Collection: In order to create the dataset of Reddit posts, we first used the jsoup library to perform a Document Search over the post pages in the r/UPenn subreddit. The CAPS-related posts were obtained by parsing the post title + post body, for every post that came up when we searched the query "CAPS" in the subreddit's search bar. The academic-related posts were obtained by parsing the post title + post body, for every post tagged under the

[1] r/UPenn subreddit, https://www.reddit.com/r/UPenn/

"Academic" flair category. This document search process for our data collection was performed using the getCapsPosts() and getAcademicPosts() functions (see **Exhibit 1**).

Also, we ran into issues retrieving all the posts due to Reddit's lazy loading (more posts load only as you scroll to the bottom), but we worked around this by opting to parse pages from old.reddit.com (which uses multiple pages of posts instead of lazy loading), where we were able to iterate through all the pages by selecting the next button via jsoup and traveling to the next page of posts.

Then, the getPostsText() and getPostPointsAndUpvoteRatio() functions were used to extract relevant information from each of the CAPS-related and academic-related posts, which were subsequently stored in separate arrays to be used in our data analysis.

III. Methods & Analysis

Cosine Similarity: In addition, we will find the cosine similarity between a query list of commonly referenced words that represent sadness/depression, and academic and CAPS posts. The CAPS-related posts were obtained by parsing the post title + post body, for every post that came up when we searched the query "CAPS" in the subreddit's search bar. The academic-related posts were obtained by parsing the post title + post body, for every post tagged under the "Academic" flair category. The general Penn subreddit posts were obtained by parsing the post title + post body, for every post that came up when we selected the "Hot" section on Penn Reddit's homepage, and the general Reddit subreddit posts were obtained by parsing the post title + post body, for every post that came up when we selected the "Popular" section on Reddit's homepage. All of these were then compared with a random list of sad and depressive words in a text file. The cosine similarity was calculated through the Vector Space Model Java project from Homework 3.

Points/UpvoteRatio: For each post, we collected the number of points (i.e. the net upvote-downvote sum), the upvote ratio (i.e. percent upvoted), and the sentiment (e.g. positive, neutral, or negative) using the getPostPointsAndUpvoteRatio() method. Once we collected this data for each post, we created visualizations (see **Exhibit 6**) so that we could analyze the data and extract insights.

Sentiment Analysis: We also thought it would be insightful to analyze the sentiment (positive/negative) of the posts' title & body, and see how that corresponded to the post's level of engagement (i.e. through the points/upvotes it received). This was done by first using document search via jsoup to extract each post's text content (see **Exhibit 2**), then the text content was cleaned and tokenized in Python via Google Colab. After removing stopwords and computing the nltk.polarity() and nltk.subjectivity() scores from the nltk (a Python package for natural language processing), we used compared these sentiment analysis metrics across the two groups of posts to compare differences among positive/neutral/negative posts (see **Exhibit 3**). Note that the nltk polarity metric ranges from -1 (negative) to 1 (positive), and the nltk subjectivity metric ranges from 0 (objective) to 1 (subjective), as described in the documentation[2].

[2] Polarity/subjectivity docs, https://textblob.readthedocs.io/en/dev/quickstart.html

Affiliation Network Graph: Additionally, in order to incorporate the social networks & graphs components into our project, we decided to construct a bipartite affiliation network using two groups of vertices – user nodes representing each Reddit user, and post category nodes representing the CAPS-related and academic-related post types. To do this, we again utilized document search via jsoup to extract the Reddit usernames for each post in our dataset; then we took the set of distinct users and created the affiliation network by adding an edge from the user to each post category they created a post in (see **Exhibit 4**). This allowed us to visualize the network of posts & users through the networkx Python package, as shown in the graph depicted by **Exhibit 5**. The evidence of triadic closure used for testing our hypothesis was computed by looking at the frequencies of edges between each of the post groups, and the proportion of users who went on to create posts in both categories (CAPS-related and academic-related).

IV. Results

Cosine similarity scores computed in reference to our query:

1. Comparing to academic_posts.txt
    a. 0.001256527340253206
2. Comparing to mental-health.txt
    a. 0.0032731570286003795
3. Comparing to general-penn-subreddit.txt
    a. 0.001212112332057713
4. Comparing to general-reddit.txt
    a. 4.286664095883971E-4

Similarity to our query words denoting depression (cosine similarities):
Mental health > Academic posts > General Penn Subreddit > General Reddit

Mental health was shown to have the greatest cosine similarity with the query words of depressive words, followed by academic posts, followed by the general Penn subreddit, and with the general reddit with the least levels of similarity (depicted in Figure 1 below).

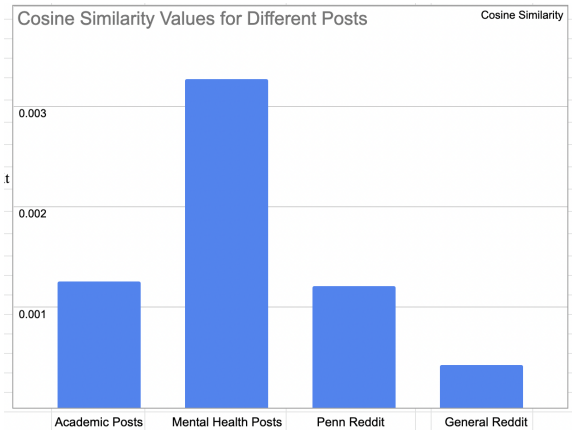


Figure 1

After creating our data visualizations (see **Exhibit 6**), we find many interesting trends in each of our graphs:

1. Points for Caps Posts vs. Points for Academic Posts:
    a. Average number of points for CAPS-related posts = 30.77 points
    b. Average number of points for academic posts = 7.75 points
2. Upvote Ratio for Caps Posts vs. Upvote Ratio for Academic Posts:
    a. Average upvote ratio for CAPS-related posts = 88.9%
    b. Average upvote ratio for academic posts = 90.4%
3. Points per Post vs. Post Sentiment:
    a. Average number of points for positive-sentiment posts = 29.07 points
    b. Average number of points for neutral-sentiment posts = 25.29 points
    c. Average number of points for negative-sentiment posts = 46.94 points
4. Upvote Ratio per Post vs. Post Sentiment:
    a. Average upvote ratio for positive-sentiment posts = 88.0%
    b. Average upvote ratio for neutral-sentiment posts = 86.8%
    c. Average upvote ratio for negative-sentiment posts = 95.8%

From the results above, we can observe:

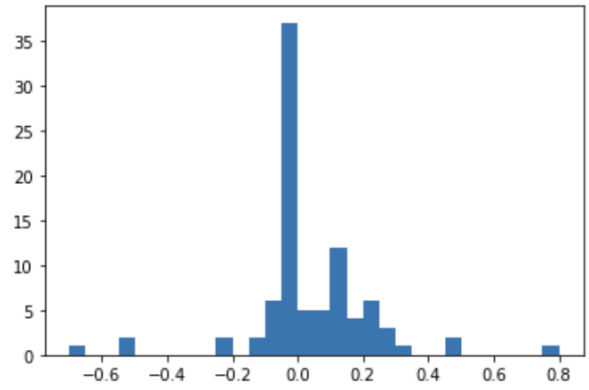
Negative-sentiment posts > neutral-sentiment posts > positive-sentiment post

---

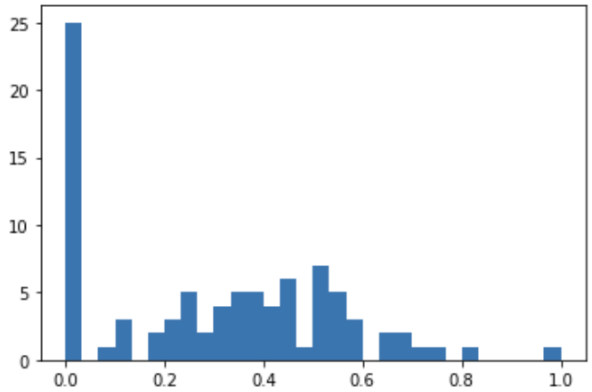
After constructing the affiliation network graph, we computed the frequencies of the various edge types, and we found that of the 325 total distinct users in the dataset, 77.54% made a CAPS-related post, 25.8% made an academic-related post, yet there were only 2.46% of the total users that ended up posting in both the CAPS and academic categories.

Displayed below is a side-by-side comparison of the polarity and subjectivity sentiment analysis scores computed for CAPS-related posts and academic-related posts. Observe that the majority of posts in both groups are clustered near 0 (indicating neutral polarity), although there is a significant number of academic posts in the -0.30 to 0.00 polarity range, indicating a relatively stronger skew towards negative sentiment in academic posts vs. CAPS posts. Looking at the subjectivity scores, the distributions are very similar across both groups, and may be a general indication of the overall subreddit's distribution of subjectivity scores.

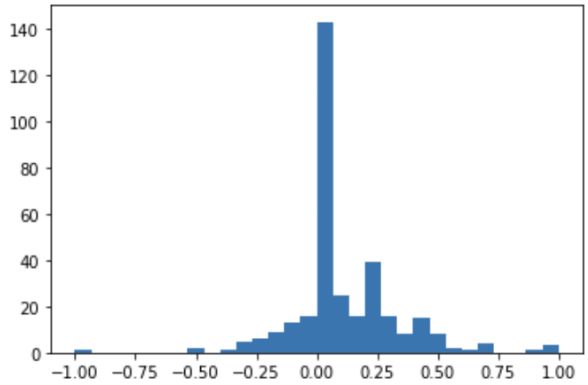
CAPS polarity scores:



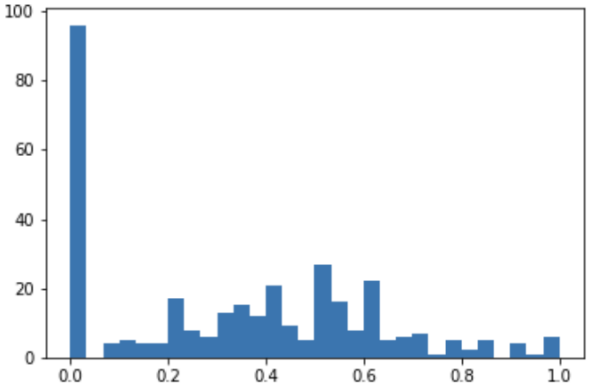
CAPS subjectivity scores:



Academic polarity scores:



Academic subjectivity scores:



V. Conclusion

Our results proved that our first hypothesis was close to the actual outcome, however, we were surprised to find out how much higher the mental health subreddit reflected the words of sadness and depression (even higher than expected). It was also interesting to see how low the general reddit score was when compared to our query list of words reflecting emotions of sadness and depression.

Pondering over this, Reddit overall seems to be happier (at least, in terms of their word usage) than the r/UPenn subreddit. However, inside the Penn subreddit, the mental health flair category seems to be the one most afflicted with sadness (in terms of word choice), which is obvious, but definitely something important to note.

Overall, it was an interesting find to observe the lack of depressive word choice in Reddit in general, and the abundance of them in the Penn reddit, especially when coupled with the mental health flair.

After looking at our data visualization, we can disprove our hypothesis that academic posts should have more points and higher upvote ratio on average than CAPS-related posts. We found that CAPS-related posts get more points than academic posts (Average number of points for CAPS-related posts = 30.77 points > Average number of points for academic posts = 7.75

points). Though the averages for the upvote ratios seem to be similar between academic and CAPS posts, our plot shows that CAPS posts have more consistently high upvotes (higher density of points at the top).

We also can disprove our hypothesis that posts with positive sentiment will have more points and higher upvote ratios than those with neutral sentiment and our hypothesis that posts with neutral sentiment would garner more points and higher upvote ratios than posts with negative sentiment. Our findings actually show that negative posts get significantly both more points and higher upvote ratios than positive posts. This is a surprising finding, but it may indicate that the Penn community is supportive of those going through tough situations (assuming that those we post negatively are the same as those we are in tough situations).

For our last hypothesis that the affiliation network exhibits membership closure such that the users who have made a CAPS/mental health-related post will be more likely to also post about academics (and vice versa), our resulting data didn't quite support our hypothesis as only 2.46% of the total users represented ended up posting in both CAPS-related and academic-related categories across the r/UPenn subreddit. As this is a relatively small proportion compared to the makeup of the post categories (78% CAPS, 26% academic), we conclude that there is not enough evidence to support that triadic closure exists amongst the post categories in the way that we described.

VI. Appendix

Exhibit 1. Code snippet for the getCapsPosts() function. Similar approach was used for the getAcademicPosts() function.

```
public void getCapsPosts() {
    String nextPageLink = "https://old.reddit.com/r/UPenn/search?q=caps&restrict_sr=on&include_over_18=on&sort=relevance&t=all";
    boolean firstTime = true;
    while (nextPageLink != null) {
        try {
            this.currentDoc = Jsoup.connect(nextPageLink).get();
        } catch (IOException e) {
            return;
        }
        Elements postElems = this.currentDoc.select( cssQuery: "a[href*=https://old.reddit.com/r/UPenn/comments/]");
        for (Element post : postElems) {
            String postURL = post.attr( attributeKey: "href");
            if (this.capsPostLinks.size() == 0) {
                this.capsPostLinks.add(postURL);
            } else if (!this.capsPostLinks.getLast().equals(postURL)) {
                this.capsPostLinks.add(postURL);
            }
        }
        Elements nextButton = this.currentDoc.getElementsByClass( className: "nextprev");
        if (nextButton.size() == 0) {
            nextPageLink = null;
        } else {
            Elements buttonATag = nextButton.first().select( cssQuery: "a");
            if (buttonATag.size() == 2) {
                nextPageLink = buttonATag.get(buttonATag.size()-1).attr( attributeKey: "href");
            } else if (firstTime) {
                nextPageLink = buttonATag.get(0).attr( attributeKey: "href");
                firstTime = false;
            } else {
                nextPageLink = null;
            }
        }
    }
}
```

Exhibit 2. Code snippet for the getPostsText() function, which was primarily used for the sentiment analysis & affiliation network construction.

```
public ArrayList<String> getPostsText(boolean caps) {
    LinkedList<String> postLinks = new LinkedList<String>();
    ArrayList<String> postText = new ArrayList<String>();
    if (caps) {
        postLinks = this.capsPostLinks;
    } else {
        postLinks = this.academicPostLinks;
    }
    for (String postLink : postLinks) {
        try {
            this.currentDoc = Jsoup.connect(postLink).get();
        } catch (IOException e) {
            System.out.println("failed to connect to post page");
            return null;
        }
        Element titleTag = this.currentDoc.select( cssQuery: "a[data-event-action='title']").first();
        Element bodyDivTag = this.currentDoc.getElementsByClass( className: "expando").first();
        if (titleTag == null) {
            postText.add("||SEPARATOR||");
            continue;
        }
        String currPostText = "||SEPARATOR||" + titleTag.text().replaceAll( regex: "\"\'", replacement: "");
        if (bodyDivTag != null) {
            Element formTag = bodyDivTag.select( cssQuery: "form").first();
            if (formTag == null) { postText.add("||SEPARATOR||" + currPostText); continue; }
            Element div1Tag = formTag.select( cssQuery: "div").first();
            if (div1Tag == null) { postText.add("||SEPARATOR||" + currPostText); continue; }
            Element div2Tag = div1Tag.select( cssQuery: "div").first();
            if (div2Tag == null) { postText.add("||SEPARATOR||" + currPostText); continue; }
            Element pTag = div2Tag.select( cssQuery: "p").first();
            if (pTag == null) { postText.add("||SEPARATOR||" + currPostText); continue; }
            currPostText = currPostText + " " + pTag.text().replaceAll( regex: "\"\'", replacement: "");
        }
        postText.add(currPostText);
    }
    return postText;
}
```

Exhibit 3. Sentiment analysis using the nltk package in Python.

```
stopwords = set(stopwords.words('english'))

cleaned_caps_posts = []
caps_posts = capsText.split("||SEPARATOR||")
for text in caps_posts:
  cleaned_post = clean_post(text)
  tokens = nltk.word_tokenize(cleaned_post)
  cleaned_tokens = [token.lower() for token in tokens if token.lower() not in stopwords]
  cleaned_caps_posts += [cleaned_tokens]


cleaned_academic_posts = []
academic_posts = academicText.split("||SEPARATOR||")
for text in academic_posts:
  cleaned_post = clean_post(text)
  tokens = nltk.word_tokenize(cleaned_post)
  cleaned_tokens = [token.lower() for token in tokens if token.lower() not in stopwords]
  cleaned_academic_posts += [cleaned_tokens]
```

Exhibit 4. Code snippet for the affiliation network graph setup via the networkx package in Python.

```
import networkx as nx
import random
G = nx.Graph()
color_map = []

G.add_node("both", size=100, type_='post_category') # both
G.add_node("CAPS", size=100, type_='post_category') # CAPS
G.add_node("academic", size=100, type_='post_category') # academic
color_map.append('tab:red')
color_map.append('tab:blue')
color_map.append('tab:green')

for i in range(len(user_categories)):
  G.add_node(i, size=1)
  color_map.append('tab:gray')
  if user_categories[i] == 3:
    G.add_edge(i, "both", color='tab:red')
  elif user_categories[i] == 2:
    G.add_edge(i, "CAPS", color='tab:blue')
  elif user_categories[i] == 1:
    G.add_edge(i, "academic", color='tab:green')

pos = {
    n: (
        0,
        random.random()
    )
    for n in G.nodes
}
pos["both"] = (1, 0)
pos["academic"] = (1, 0.5)
pos["CAPS"] = (1, 1)
# pos = nx.spring_layout(G)
colors = [G[u][v]['color'] for u,v in G.edges()]

nx.draw(G,
        pos=pos,
        node_color=color_map,
        edge_color=colors,
        with_labels=[True])

plt.show()
```

Exhibit 5. Visualization of the affiliation network graph. The left-hand side group of nodes is the ~400 different Reddit users in our posts dataset, who are connected by edges representing the posts they have made on the r/UPenn subreddit.

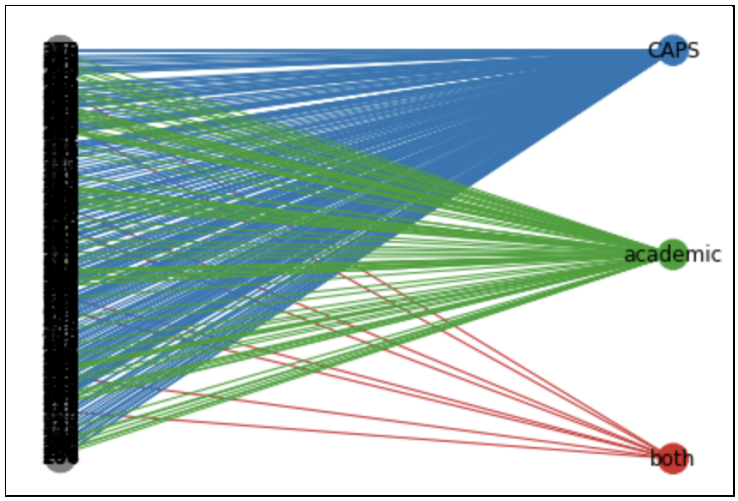


Exhibit 6. Data Visualization of points, upvote ratios, and sentiments between posts of positive, neutral, and negative sentiment through box-and-whisker plots. The upper left graph is a comparison between the number of points of CAPS-related posts and of academic posts. The upper right graph is a comparison between the upvote ratios of CAPS-related posts and of academic posts. The bottom left graph is a comparison between the number of points of posts of varying sentiments. The bottom right graph is a comparison between the upvote ratios of posts of varying sentiments.