Lim, Riana Mary Claire G.
171309
Hands-on Exam

The question I want to answer is whether or not the "distances" between the students determine the similarity or difference of the language they use. First of all, I can't really say for sure the distance between students has a causal relationship with the language they use, instead, I tried to look for any correlation between the two. I think that answering this question would be interesting because we would see how much influence students actually have on each other. This might be something teachers could look into, especially because education nowadays has been putting a lot of importance on collaborative learning. In the end; however, I found that there is no correlation between what I defined to be distance and the students' reflection papers.

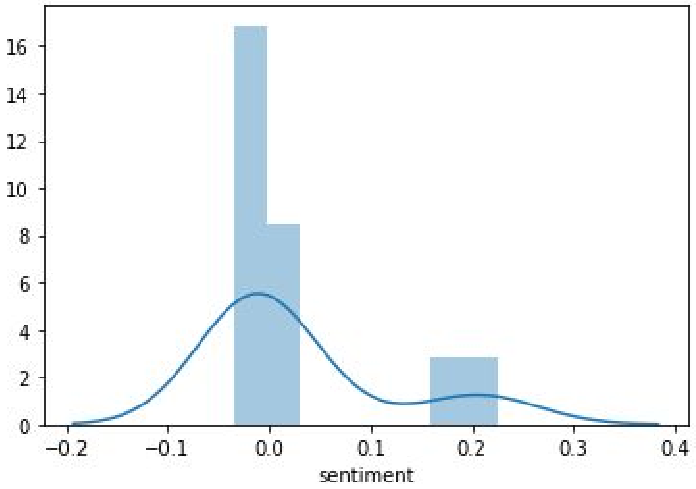
Text Preprocessing

First, I preprocessed the reflection papers as usual. This included removing special characters and converting the characters to lowercase for uniformity. I initially used the `clean_text()` function which was provided in one of the jupyter notebooks from class. However, when I started doing lemmatization, I noticed that words like "ive" and "dont" remained in the lemmatized text. This made me realize that I should expand the contractions first.

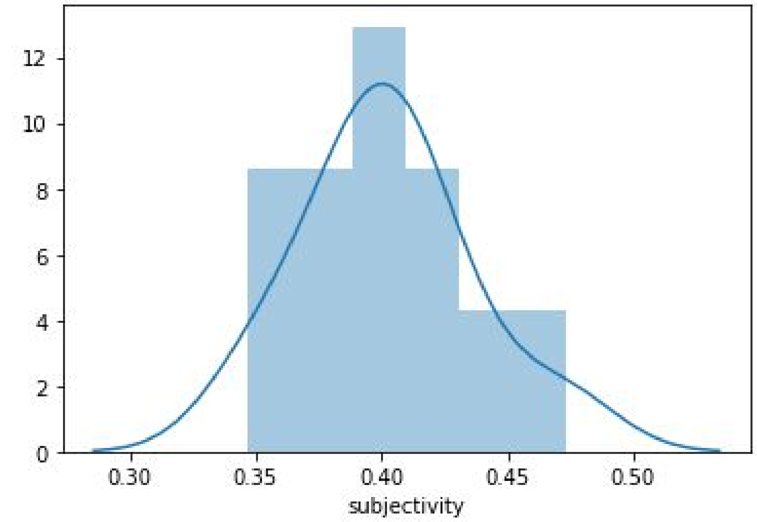
I was able to find a map of contractions and a corresponding function for using it. I noted that the process would not be perfect because some contractions have different expansion possibilities.

Sentiment Analysis

After preprocessing the text data, I then did sentiment analysis. I found that the general sentiment of the papers was neutral.

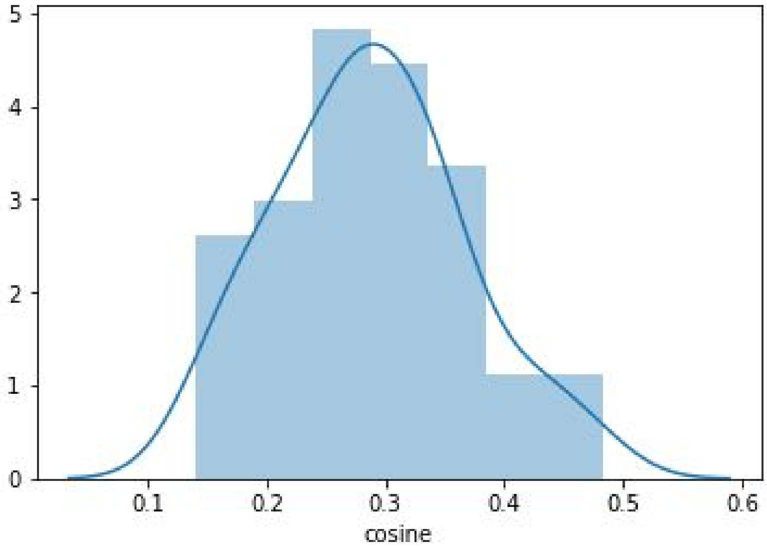


I also found that the general subjectivity of the papers is slightly positive.



Text Similarity

I used cosine similarity and term frequency-inverse document frequency to vectorize the reflection papers. I also tried euclidean similarity but decided on sticking with cosine similarity in the end.



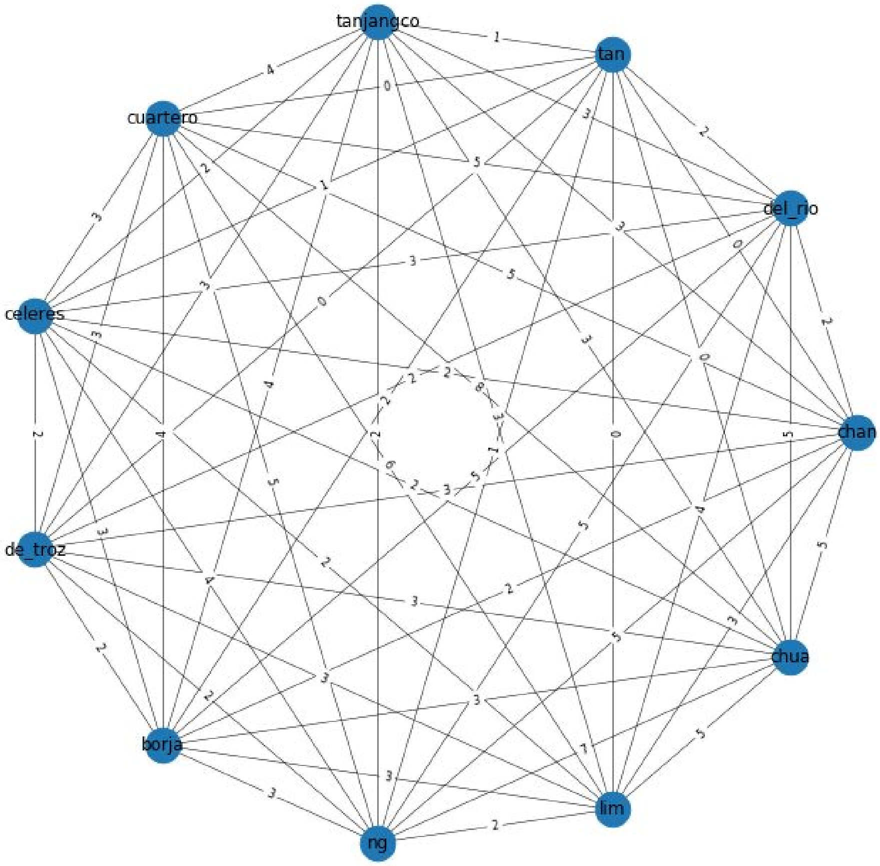
Generally, the papers are not very similar. I computed for the mean of the cosine similarity and found that it was 0.2878.

Distance

Our class was able to create a small dataset where we asked each other questions and answered them. From the entire dataset, I decided to stick use the following information: table number during the activity (a string of either `'one'` or `'two'`), organizations last semester (space-delimited), high school, classmates they knew before the class (space-delimited), and classmates who are among their top contacts on Facebook messenger (space-delimited). Among the features we were able to each other, I found those to be the most interesting. I felt that these features would give us an indication of how much the students are exposed to each other, and also how similar the different environments they are exposed to are.

For each feature, I used a document term frequency to make it easier to count the weights. The distance I used then became the weight of each students' relationship with the others. Using this, I was able to produce different data frames which gave me integer values based on how much they had in common.

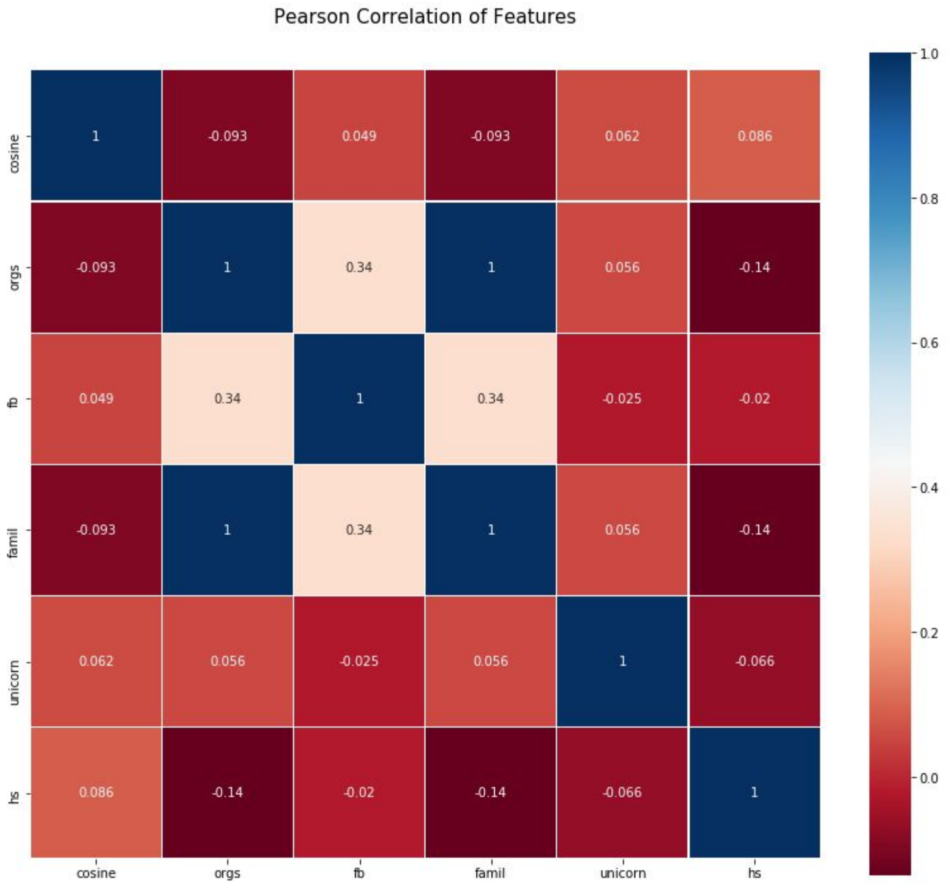
I also tried to graph the entire social network, considering all the sources of relationship weights. In the graph, we can see that some people are much closer to others. There are also some who don't have much association with the other students.



Correlation

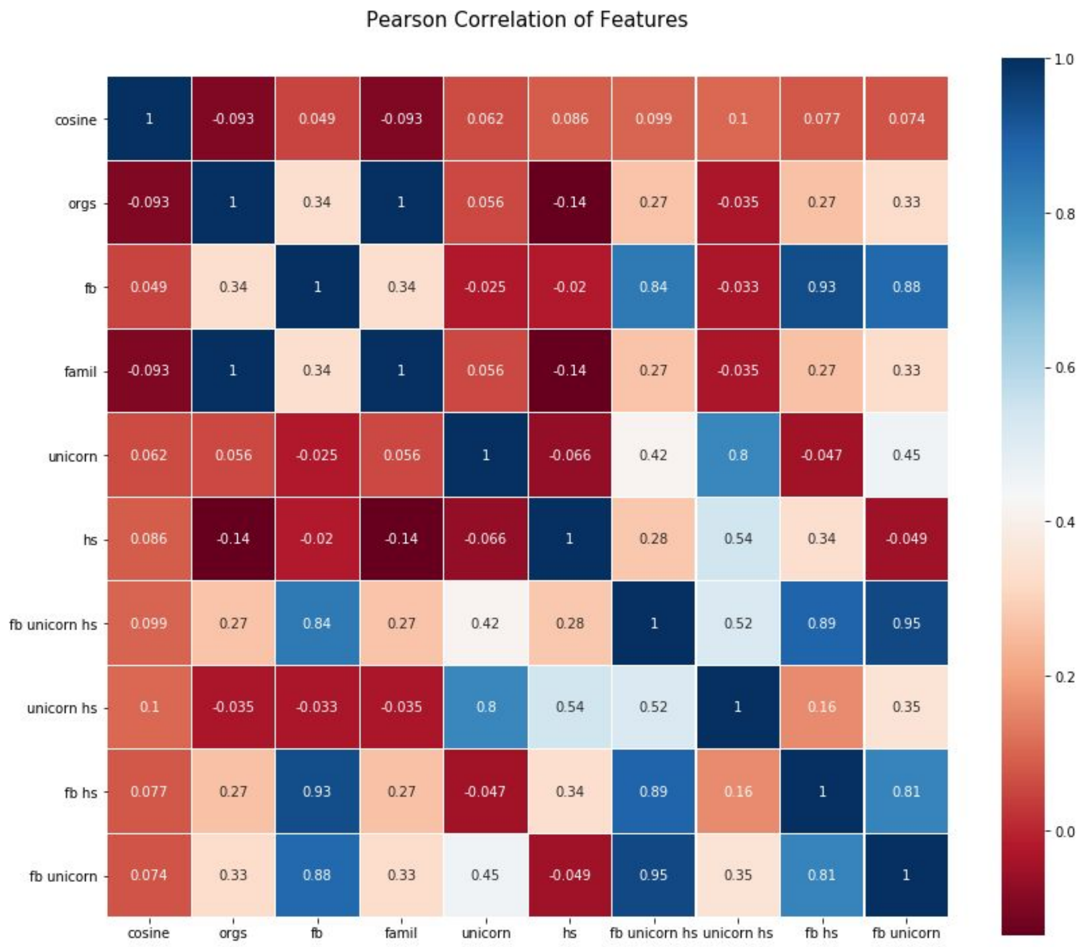
I wanted to use a heatmap to easily visualize the correlations of the different features to the cosine similarity of the text data.

First, I created a data frame which contained the student indices and the cosine similarity. I then added each of the five distance features individually to this data frame.



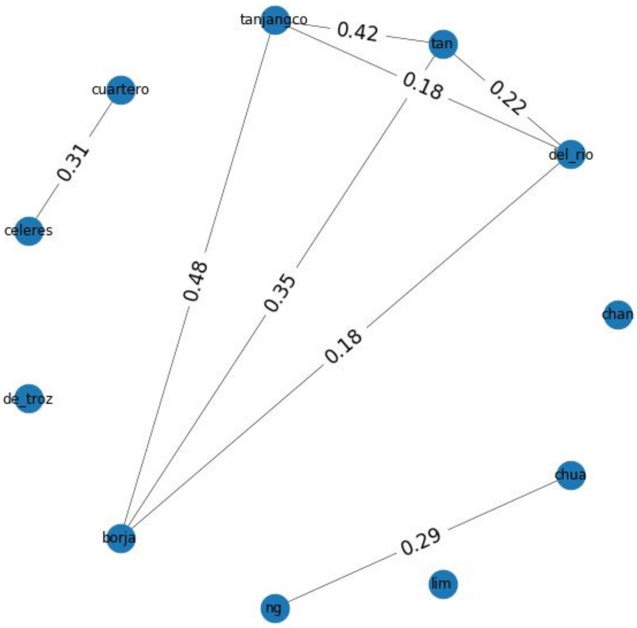
Right off the bat, we can see that there isn't much correlation between the cosine similarity and the different features. However, there are some features that perform slightly better than the rest which are the top Facebook contacts, Unstable Unicorns group, and the high school.

I added different combinations of these three features to the heatmap.



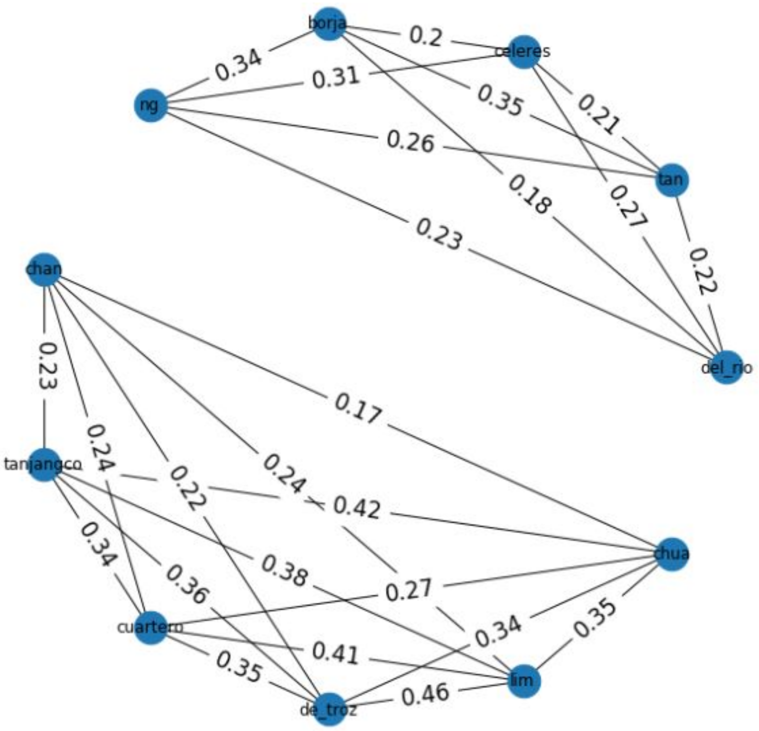
Out of all the distances, the best combination was when we considered only their Unstable Unicorn group and their high school of origin.

I then created a social network and grouped them according to their high schools.



As observed previously, there is generally not much similarity among the texts. However, it is interesting to note that those students who studied in the same schools wrote papers with slightly higher cosine similarities. This is especially evident among Borja, Tanjangco, and Tan who all studied in Ateneo High School and scored cosine similarities which are greater than the mean of 0.2878. The Xavierians and those from Chiang Kai Shek also scored a similarity which is higher than the mean.

I also tried to group them according to their Unstable Unicorns group.



There is still not much similarity, but the groups have similarities with each other within the 0.20 to 0.40 range.

Conclusions

There were certain features that I took notice of because I was working with the hypothesis that, the more the students are exposed to each other or to similar environments, the more similar they might think or type. In the end, I found that the features that had slightly more correlation to the similarities were the activity group and the high school.

For the activity group, they may have been talking about the same experiences during the game and thus using similar words.

The high school of origin is also very interesting. I think that they might have been exposed to the same kinds of teachers and teaching styles, and because of this, similar work might have been expected of them. I think that the years of influence from the same high school might have caused them to write a bit similar to each other.

In the end, there is not much correlation between their style of writing and the weight of their relationships.