Eric Chen and Jordan Hall
Math 22
Professor Rockmore
November 18, 2014

Professor Text Comparison Project

I. Overview/Project Goals

At the beginning, we were interested in analyzing the writings of Dartmouth professors from various departments. Our initial idea was to compare user writings to the writings of professors and develop a measure for how "similar" the two sources were. This idea evolved into a project to analyze and compare professor papers across various departments, with the goal of determining whether or not there were legitimate stylistic differences between disciplines. To do this, we used the fifty most common function words in the English language and recorded their usage across the papers. Examples of words that we used are 'the', 'it', and 'for'. We first calculated the Euclidean distance between each vector, and calculated the average distance between each subject pair. We then used Principal Components Analysis and a handful of other calculations to create a visual representation of the data. From our results, we were able to come to a conclusion regarding the stylistic differences across departments at Dartmouth.

In terms of our learning goals, we undertook the task of the project hoping to learn more about some of the "real life" applications of the course material. Of course, text analysis is a very relevant application of linear algebra, and because of that, we learned a lot about analyzing text in a manner relevant to what we learned in class.

II. Implementation

We first had to decide which technologies to use. We used Python for file reading and vector construction, and Matlab for performing Principal Components Analysis and creating the visualization of the data. We chose these two languages because of the breadth of their built-in libraries.

For our data, we selected a total of thirty-one papers from professors across eight different departments: Anthropology, Biology, Math, Computer Science, History, Chemistry, Economics, and Government. We converted all the papers to text files, and removed graphs, images, and references. We created a vector for each paper, which contained the number of times each style-word appeared in that paper. We standardized the vectors in order to account for varying paper lengths. We standardized the vectors by dividing each entry of a given vector by the total number of style-words in that paper, which is the sum of all the entries in that vector.

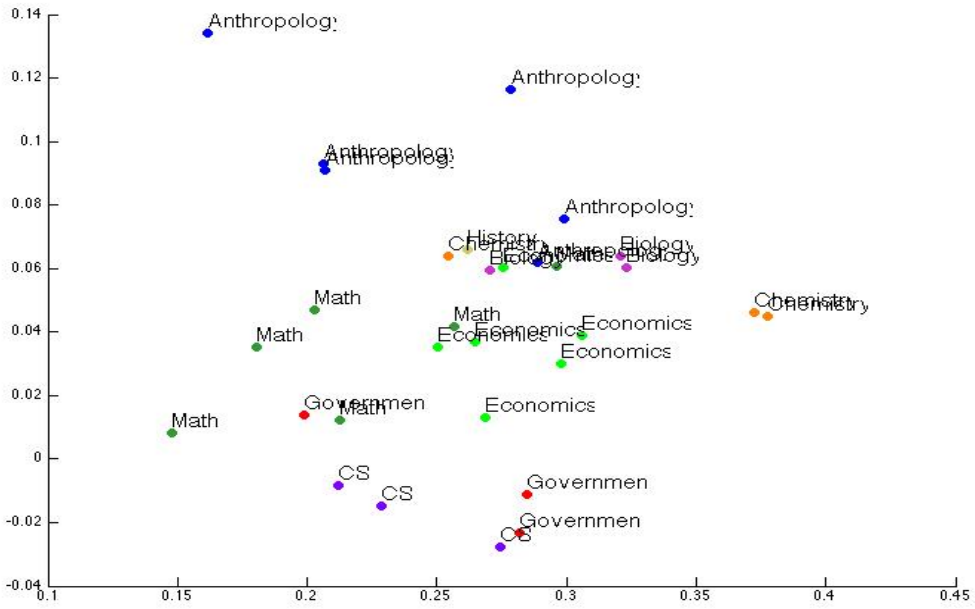
We found the Euclidean distance between each pair of papers. We also found the average distance between each pair of subjects. We used the Matlab pca command to perform Principal Components Analysis on our data set. We then projected each vector onto the first, second, and third principal components. This gave us the amount that each of these principal components explained for each vector, which can be used for assessing the similarities of different papers. We graphed the projections in two- and three-dimensions to visualize the data.

III. Results

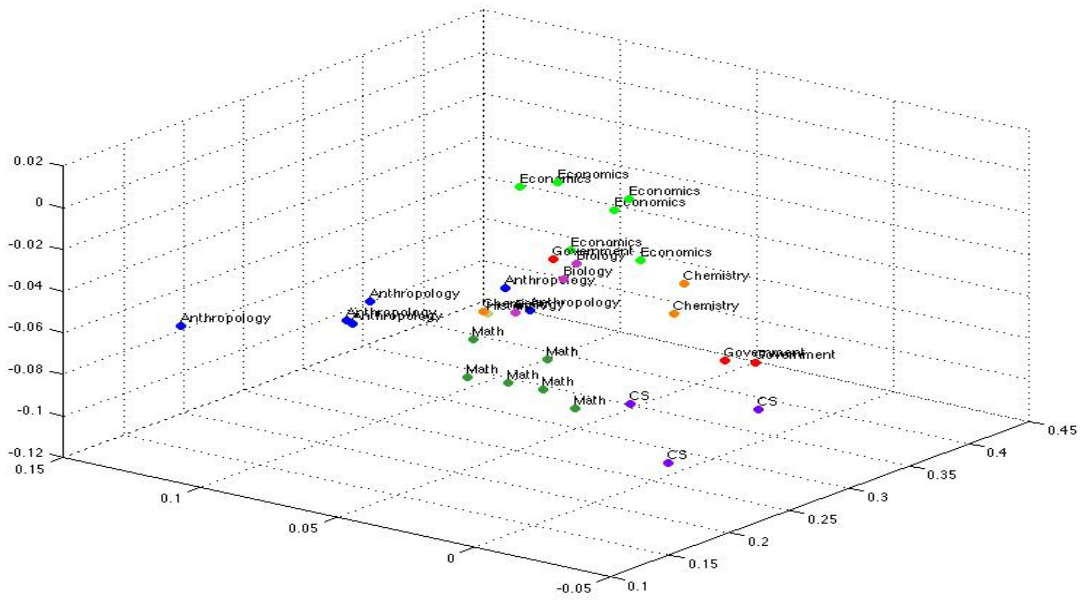
Based on the final visualization that we developed, it is clear that there is significant stylistic variation across disciplines. We see that each department clusters

together, and that certain departments cluster closer to other departments. Specifically, math and computer science clustered close together, biology, chemistry, and anthropology clustered close together, and government and economics also clustered close together. This observation is not quite strong enough to make a strong conclusion, but it is certainly notable. To further validate this observation, we would want to consider more papers, more departments, and more authors.

The eigenvalues for the covariance matrix in descending order were: .0030, .0015, .0008, .0006, .0003, .0003, .0002, .0001, .0001, .0001, .0001, .0001, and the rest had trivial magnitudes. These eigenvalues sum to .0072. This means that the first principal component explains (.0030/.0072) ≈ 42% of the variance in the data, the second principal component explains (.0015/.0072) ≈ 21%, and the third principal component explains (.0008/.0072) ≈ 11%. In total, the first three principal components explain 74% of the variance in the data.

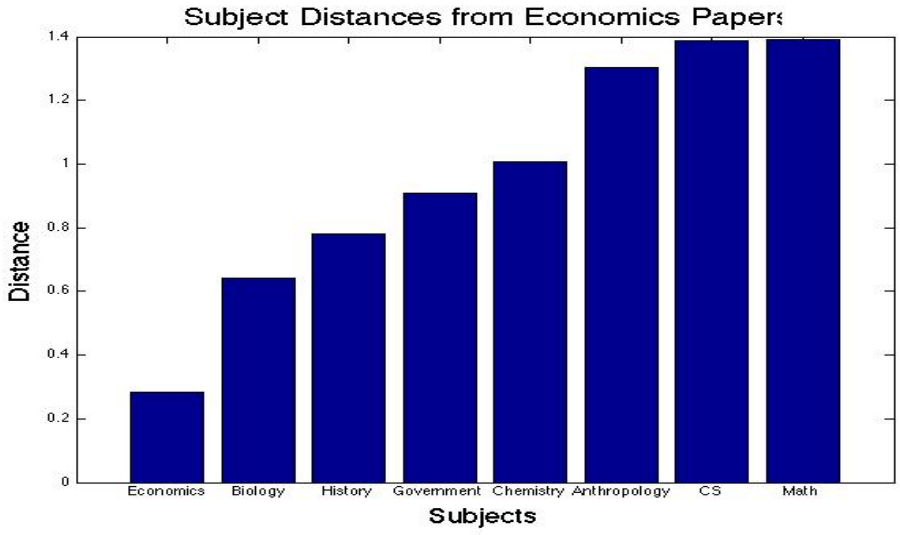


2d PCA visualization



3d PCA Visualization

Our calculations of Euclidean distance also found discrepancies between subjects, though not as strongly as the results we found the PCA. The average distance between papers of the same subject were usually smaller than the distance between different subjects, and we did see some instances where “similar” subjects had smaller distances that subjects that were not similar.



IV. Final Thoughts

Overall, we were very pleased with the quality of the final visualizations that we produced and the conclusions about style that we were able to reach. However, ideally, there would be several other elements of the project that we would include. We wanted to implement Singular Value Decomposition but time simply ran out. In order to do that, we needed data representing the content of papers. This would involve a matrix consisting of standardized word counts for all words in every paper. After that we could simply use Matlab's svd command, similar to what we did for PCA. The inclusion of more papers would also be worthwhile with more time, to reduce the amount of potential bias in our data set. We also hoped to implement PCA and SVD from scratch, instead of relying on Matlab.

Ultimately, our project was a good learning experience that reinforced concepts relevant to the course. We worked with vectors, distance between vectors, and Principal Components Analysis. We also caught a glimpse of some research challenges, such as collecting data and reducing bias.