Jeremy Mednik, Stephen Owen, Karim Naous

Project 4 Findings

Our objective was to create a Wiki dataset consisting of articles that were mentioned in the article, “List of Largest Companies in the United States by Revenue.” Using unsupervised learning, we were to make some interesting inferences in the data. We found a total of 781 Wiki article entries which consisted of cities, states and fortune 500 companies.

Firstly, we looked at which titles contained the most words in their respective articles. The highest worded articles were “Camden, NJ”, “Apple Inc.”, “New York City.” What was interesting was that the “United States” article was less worded than some of the cities. When we looked at the lowest worded articles, it consisted of companies we never heard of before. We then wanted to split the Wiki articles into 16 clusters. For the most part, each cluster had unique company industries grouped together, but clusters 5, 6, 7, 11 and 15 contained cities. We then wanted to create a PCA with different clusters and found the biggest differences between two clusters were clusters 6 and 11. We then created a UMAP and expected similar results but ended up finding more useful information. The UMAP separated two types of groups of clusters: Cities and Fortune 500 companies. We realized we could gain further analysis by creating different PCAs and UMAPs of different cluster sizes.

Doing UMAP on our data-set with seven clusters yields very interesting results. We see two main areas of note in the plot, the top right and the bottom left. The bottom left, which is cluster 3, consists mainly of wiki articles on cities and states which are where the companies in the Fortune 500 are headquartered. The top right corner consists mainly of the actual companies in the Fortune 500. Cluster 1 includes companies in the Finance and Insurance industries, which are very similar. Cluster 5 includes big retailers like “Target” and “Best Buy”. Cluster 7 includes Energy companies like “ExxonMobil” and “Dominion Energy”. Cluster 2 seems to include a mix of Manufacturing and Engineering companies like “Clorox” and “Howmet Aerospace” which is interesting. Cluster 4 includes a mix of technology companies like “Alphabet Inc.” and bank/financial companies like “Capital One”. Perhaps with the prevalence of Fintech, banks and Technology companies are growing more and more alike. Finally, cluster 6 includes a mixture of Technology, Automobile, and Airline companies. Perhaps the relationships in cluster 6 come from the use of Technology in Automobiles, especially self-driving cars, and in Airplanes.

We then looked at page clustering using only two clusters. The data was run through PCA, and was plotted based on two of the principal components. Doing this revealed two very distinct clusters. We then dove into what was contained in those clusters, and it turned out that it managed to create a cluster of companies, and a cluster of locations in the US, which we found very interesting. The last thing we looked at was topic modeling using Latent Dirichlet Allocation. We allowed it to generate topics, and found that a lot of the topics that it generated happened to actually mimic some of the clusters that were generated earlier in our discovery. Another cool result that came from topic modeling using LDA was that it managed to combine keywords in interesting ways. For example it lumped all "visual media" type companies into one topic (network television, film, gaming, etc).