*Cody, Drake*

*May 9th, 2021*

*|*

BUSINESS DESCRPTION CLUSTERING

**Abstract:**

Text data from the business descriptions of the holdings of the S&P 100 in 2019 was used to discover if companies in the same Global Industry Classification Standard (GICS) sector have similar business descriptions in 10-K filings. Applying hierarchical clustering to the business descriptions transformed into TF-IDF vector representations, the analysis found only certain sectors had similar business descriptions. The results appear to group companies with similar operations together potentially suggesting that the cosine similarity of business descriptions can be used as a proxy for how similar two company’s operations are.

**Abstract:**

Text data from the business descriptions of the holdings of the S&P 100 in 2019 was used to discover if companies in the same Global Industry Classification Standard (GICS) sector have similar business descriptions in 10-K filings. Applying hierarchical clustering to the business descriptions transformed into TF-IDF vector representations, the analysis found only certain sectors had similar business descriptions. The results appear to group companies with similar operations together potentially suggesting that the cosine similarity of business descriptions can be used as a proxy for how similar two company’s operations are.

**Data Summary:**

The S&P 100 is a subset of the S&P 500 index. The S&P 100 is designed to track the performance of public companies that have a market capitalization of at least $14.6 billion, are listed on United States stock exchanges, and have a holdings GICS sector representation like the S&P 500. This made the S&P 100 an attractive index for this analysis since the holdings of the index provided a good sample of companies in all of the GICS sectors.

All public companies are required by the SEC to file an annual report known as a 10-K. The report is made up of multiple items and sub-items with clear definitions for what information is to be available to the public through 10-K annual reports. The item that the analysis in this paper focuses on is Item 1. Also known as the business description item, the definition from the SEC for item 1 is as follows:

**“Item 1**requires a description of the company’s business, including its main products and services, what subsidiaries it owns, and what markets it operates in. This section may also include information about recent events, competition the company faces, regulations that apply to it, labor issues, special operating costs, or seasonal factors. This is a good place to start to understand how the company operates.”

The year chosen for the analysis was 2019. The tickers that made up the S&P 100 on 1/15/2019 were sourced from Bloomberg. Due to issues with company identifiers being different in Bloomberg and SEC EDGAR, 10-Ks from 90 out of the 100 tickers that were members of the S&P 100 on 1/15/2019 were used in the analysis. The business descriptions were extracted from these documents and transformed into TF-IDF vector representations post the removal of numbers, punctuation, and common words with no meaning (see NLTK stop words in appendix). The cleaned dataset used in this analysis consisted of 90 vectors with 19,187 dimensions each.

**Methods of Analysis**

A vector space model was used to transform the business descriptions into TF-IDF vectors. This process consisted of finding the term frequency of each term in each business description then weighting the term frequency by the inverse document frequency of the term based on how many business descriptions the term appears in. The vectors were then transformed so that the root sum of squares equals 1. Normalization was done so that the length of the business description (defined by word count) did not influence how similar two descriptions were.

The vectors were clustered using a hierarchical clustering algorithm using cosine distance. Cosine distance is the most popular distance measure for document clustering (Steinbach, Michael). Complete linkage, average linkage, single linkage, and centroid linkage were tested to see which method yields the most balanced dendrogram. Complete linkage was chosen for the analysis since this linkage method had the least amount of chaining and no inversion in the dendrogram.

To understand which terms were separating the business descriptions, principal component analysis was used. However, since the dataset was so high dimensional, little information can be extracted from the principal component analysis visuals. To further understand which terms were separating the documents, the TF-IDF values of each term were averaged in each cluster to create word clouds.

**Results and Interpretation:**

Using the dendrogram created by the results of hierarchical clustering, the companies were split into 10 clusters. Companies that have similar operations like Ford (F) and General Motors (GM), US Bancorp (USB) and Wells Fargo (WFC), and most of the communication Diagram, schematic

Description automatically generatedservices companies merged lowest on the dendrogram indicating that firms with similar operations use similar terms in their business Chart

Description automatically generateddescriptions.

The sector breakdowns of each cluster shows that companies in different GICS sectors can have dissimilar business descriptions except for Utilities, Financials, and Health Care companies. Companies in the same GICS sector were assigned to different clusters indicating that firms in the same GICS sector do not always use similar terms in the business descriptions filed in the company’s 10-K.

Because many of the companies in the S&P 100 have a diverse set of operations, companies with different GICS sector classifications can be more similar in terms of their operations than another company classified under the same GICS sector. For example, the energy companies who are involved in the buying and selling of oil and gas, Exon Mobil (XOM) and Chevron (CVX), were clustered with utility companies who are highly integrated in energy markets. The other two energy companies, Halliburton (HAL) and Schlumberger (SLB), produce technology for companies involved in energy markets. The difference in operations of the energy firms not grouped with utilities explains why these two companies could have higher similarity to industrials and technology firms than other energy sector firms. Therefore, cluster assignments could potentially be a better indicator of how similar company’s operations are than the GICS sector classifications.

To understand what is separating the business descriptions into their respective clusters, principal component analysis is used. Each principal component representants a dimension in the dataset that explains a fraction of the dataset’s total variation. In any dataset, there are *M* principal components where *M* ≤ min(n-1,p). Since the data contains 90 company’s description Shape

Description automatically generatedvectors (n = 90) with a length of 19,187 terms each (p = 19,187), there are 89 principal components.

Chart, scatter chart

Description automatically generatedAnalyzing the proportion explained by each principal component, each dimension explains almost equal amount of variance. This shows that a model cannot be made to project the dataset into a lower dimensional space. Therefore, the first three principal components of the dataset do not give much information about what terms are separating the business descriptions. This is seen in the scatter plots of the PCA projections in a two-dimensional space using the first/second principal components and the first/third principal components. However, the data points are scattered out on the plot indicating there are natural groupings in the data that hierarchical clustering can find.

Timeline

Description automatically generatedUsing Word clouds, the terms separating cluster become somewhat clearer. Focusing on the four largest clusters (1,2, and 6), word clouds were generated by the average TF-IDF values of the vectors in each cluster. Cluster 6 has many financial sector related terms like banking, capital, FDIC, reserve, and regulatory explain why all these companies are classified as Financials. Cluster 1 did not have a balance of companies across GICS sectors explaining the mix of terms in the cluster 1 word cloud. Cluster 2 was mostly healthcare companies explaining why terms like pharmaceutical, health, therapy, device, disease, and FDA, and appear in the word cloud. Cluster 3 had Information Technology and Industrial companies explaining why terms like network, production, system, material, application, cloud, and software appear in the word cloud. Overall, these word clouds are showing the difference in the terms used in the descriptions making up each cluster.

**Limitations and Concerns**

Using 19,187 terms made the results of the clusters difficult to fully interpret using the principal components. Many words that show up in the word clouds like company, business, industry, cost, etc. could be eliminated since they are so commonly used in business regardless of the difference in operations. Seeing in the scree plots (see appendix) was relatively flat indicates that this data set cannot be represented in a lower dimensional space. Eliminating terms could help make principal component analysis more useful.

Further analysis is needed to confirm if cosine similarity of business descriptions is a good proxy for how related operations of two companies are. Only in certain cases did this claim make sense using clusters in this analysis. Having a measure of similarity between business would help with comparable pricing models for investment analysts trying to find the fair value of an investment security using the price of an investment security from another company. A robust measure of similarity could also further research in seeing if M&A transactions are more successful when the firms are more similar or less similar.

**References:**

Ben. *Build a Search Engine in 20 Minutes or Less*, 1 Jan. 1970, [https://anythingbutrbitrary.blogspot.com/2013/03/build-search-engine-in-20-minutes-or.html.](applewebdata://DDED28CE-7148-4D18-BC30-6FDFD9F81607/Ben.%20Build%20a%20Search%20Engine%20in%2020%20Minutes%20or%20Less,%201%20Jan.%201970,%20https:/anythingbutrbitrary.blogspot.com/2013/03/build-search-engine-in-20-minutes-or.html.)

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O’Reilly Media Inc

“Fast Answers.” *SEC Emblem*, 1 July 2011, <https://www.sec.gov/fast-answers/answersreada10khtm.html>.

*Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.*

[Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

Steinbach, Michael, et al. *Www.stat.cmu.edu*. https://www.stat.cmu.edu/~rnugent/PCMI2016/papers/DocClusterComparison.pdf.