Job Posting Insights: Text Analytics & Firm Categorization

Student: Shuyang Lin | Instructor: Prof. Peter Haslag

# Executive Summary

This capstone project aims to examine the applicability of text analytics to describe the pattern of companies in the job demand market. Distinguishing itself from prior research, this study uniquely combines elements of text analytics, company and industry perspectives on the job market, and a macro-level examination of employment trends. A large dataset containing 607,795 rows of job descriptions from various companies and years was utilized. The analytics process includes the following steps: data cleaning, vectorizing, dimensionality reduction, clustering, visualization, topic modeling, and similarity score calculation. The output consists of two datasets: one with cluster labels for each company in a given year, representing the type of job demand, and another with pairwise cosine similarity scores for comparing job demands between companies in the same year or the same company across different years.

After cleaning the raw text by removing stopwords, numbers, and punctuation, the top 500 companies with the most job postings were selected for further analysis. The cleaned descriptions were concatenated by firm-year pairs and transformed into a vectorized format using TF-IDF with lemmatization.

TruncatedSVD was employed for dimensionality reduction to 100 dimensions. DBSCAN clustering proved ineffective, producing only one dense cluster and noise. K-means clustering resulted in 9 clusters identified using the elbow method. However, the silhouette scores indicated that some clusters were sparse.

T-SNE was used for visualization, revealing two well-distinguished clusters. Latent Dirichlet Allocation (LDA) was applied to extract topic words representing the clusters' themes. By linking industry names to the data, it was found that five clusters had a dominant industry (>50%). In contrast, the remaining clusters were composed of 2-3 industries. This result showed that it’s possible to apply text analytics to classify the “job demand group” of companies.

Finally, cosine similarity scores were calculated for pairwise comparisons between companies in the same year and the same company across different years. The results were saved in a CSV file for future analysis.

# Background & Introduction

This capstone project was inspired by the innovative concept proposed by Professor Peter Haslag, Assistant Professor of Finance at Owen Graduate School of Management, Vanderbilt University. As the project's mentor and primary data source, Professor Haslag suggested utilizing the textual content of job postings from various organizations to monitor the job market in real time. The previous application of text analytics in job market research, such as job market trend exploration (Mbah et al., 2017), was primarily focused on web crawling and job type analysis. As a result, the methodology to determine a company's position in the job market remains in its infancy, warranting further investigation.

Conventional industry categorization methods, such as SIC or NAICS, are designed to place companies in predefined industry categories based on production processes, as stated by the United States Census Bureau. In this project, the North American Industry Classification System (NAICS), employed by Federal statistical agencies for business establishments classification, was utilized for examining the results of job posting analytics.

In recent years, research has begun to incorporate text data analytics into industry competitiveness evaluations. A notable example is the Text-Based Network Industries Classification (TNIC) by Gerard Hoberg and Gordon Phillips. They devised a pairwise product similarity scoring system and positioned firms within a spatial representation. Their initial approach involved web crawling engineering and analyzing the gained dataset containing business description text from firms' annual 10-K filings with the SEC.

Drawing upon Hoberg's approach, this project aims to leverage vector representations derived from job postings to create spatial representations of firms during specific years. By focusing on the demand side of the job market, this methodology will contribute to future research on company categorization, labor competitiveness assessment, and job market trend analysis.

# Data

This capstone project made use of a condensed sample from LinkUp's proprietary dataset, originally containing over 230 million job records from over 60,000 company career sites. The access was provided to Prof. Haslag by LinkUp, a software company that aggregates job listings and develops an employment media platform for labor demand assessment and research.

The sample dataset employed in this project was a 1.66GB CSV file supplied by Prof. Haslag, containing the first queried job description for each firm-occupation pair on an annual basis. It contains 607795 rows and six columns. Table 2-1 shows the data dictionary:

Table 2-1

| Variable Name | Data Type | Description |
| --- | --- | --- |
| description | String | The first full-text descriptions scraped from company career sites, given the gvkey, year and onet\_occupation\_code |
| tic | String | Stock ticker of the company in the given year |
| onet\_occupation\_code | String | An identifier from LinkUp’s NLP solution using job records and descriptions to get normalized titles |
| year | Integer | Year truncated from job post creation timestamp |
| number\_of\_posts\_all | Integer | Count of job posts given the onet\_occupation\_code, gvkey, and year |
| gvkey | Integer | A unique global company identifier assigned to each company in the Compustat-Capital IQ database |

The term "condensed" refers to the dataset containing only a single job description for a combination key (gvkey, year, onet\_occupation\_code) obtained through SQL-like queries from the LinkUp dataset.

The distribution of year and the first two digits of onet\_occupation\_code (stands for the NAICS industry) of the dataset is illustrated in Figure 2-1. We find that the dataset covers most job postings during 2012-2018, with several dominant occupations.

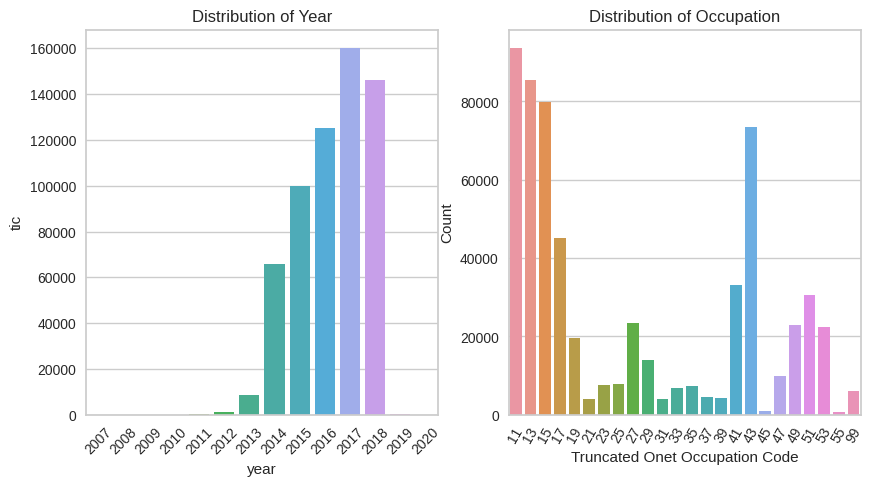


Figure 2-1 Distribution of Year and Onet Occupation Code

Since the description format consists of raw text extracted from HTML pages, text cleaning is necessary. For optimization of the analysis at the firm level, the project focused on the top 500 companies with the most significant sum of number\_of\_posts\_all among the 2,952 firms in the original dataset. The text-cleaning process began by removing URLs, star words, numbers, punctuations, stop words, and special UTF-8 characters. Figure 2-2 compares the raw length of the original dataset and the cleaned length of descriptions for the top 500 companies.

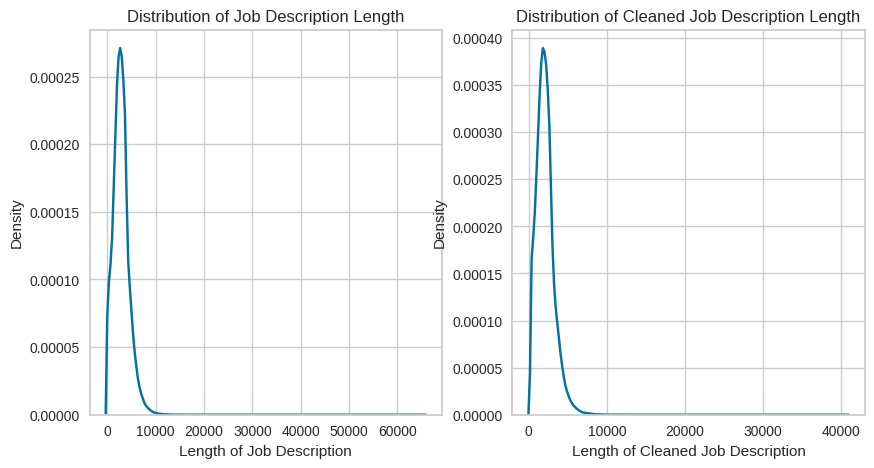


Figure 2-2: Distribution of Job Description Length

# Methodology

After truncating and cleaning the data, four steps were implemented to solve the unsupervised learning question and deliver the expected output.

## TF-IDF Vectorization

To employ clustering algorithms, the text data must first be transformed into numbers. The statistical measure Term Frequency-Inverse Document Frequency (TF-IDF) was utilized to vectorize concatenated texts of cleaned job descriptions, representing the job demand for each company in a given year. Prior to inputting the series into the vectorizer, a LemmaTokenizer was applied to perform lemmatization. This process groups together various inflected forms of the original word, reducing bias and extracting common topics. The loss of sentiment elements in negative sentences was deemed less significant in relation to the project's objective.

The text length distribution informed the decision to set the max\_features argument of the TF-IDF vectorizer to 4096, preventing the generation of numerous insignificant columns. The resulting output was a sparse matrix of size 2950x4096.

## Dimensionality Reduction

Given that the number of columns exceeded the number of rows, dimensionality reduction was employed. Truncated SVD, a widely used method, does not center the data before computing the singular value decomposition, making it efficient for working with sparse matrices, unlike PCA. Following scikit-learn's user guide recommendations, the number of components was set to 100, yielding an output of size 2950x100. The explained variance ratio was 0.478, with the singular values ranging from 2.06 to 28.99.

## Clustering

As the data distribution was unknown, DBSCAN, K-Means, and Hierarchical clustering algorithms were utilized for comparison, which was also the largest challenge in this project. A grid search was performed for DBSCAN as Figure 3-1 of Epsilon values ranging from 0.2 to 2 and Min Samples values between 20 and 80. The optimal Silhouette score was 0.30, achieved by an Epsilon of 0.95 and Min Samples of 20. However, after the fit and transform process, the labels consisted of only 1 cluster containing 2,948 samples, with 2 samples identified as noise. This outcome suggested that the data was too dense to effectively apply DBSCAN.

|  |  |
| --- | --- |
| Figure 3-1 Grid Search of DBSCAN w/ Silhouette Score & Noise Ratio | |

Following the unsuccessful attempt with DBSCAN, a search for the optimal K for K-Means was conducted. The Distortion Score Elbow is depicted in Figure 3-2, with the Elbow method suggesting a K of 9.

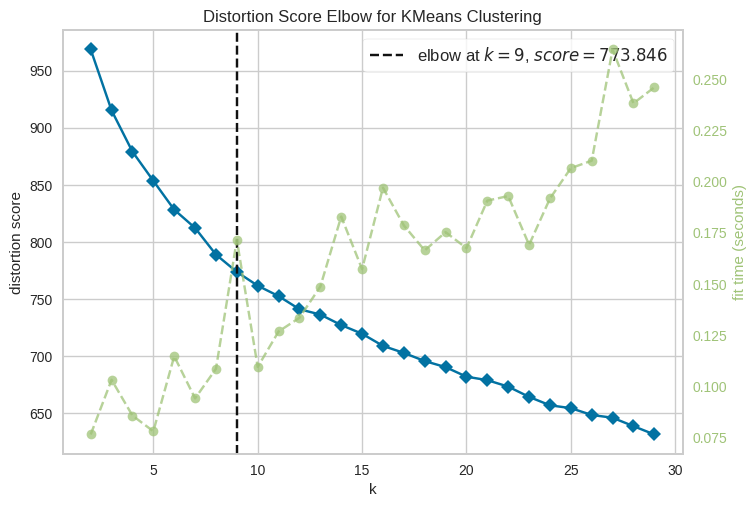


Figure 3-2 Distortion Score with respect to K

Silhouette Scores were also calculated for different Ks ranging from 2 to 15, as shown in Figure 3-4. It was observed that several groups had negative Silhouette scores, indicating the presence of non-spherical or non-convex clusters, which K-Means might struggle to delineate accurately. The highest score of 0.10 was obtained with a K of 4, while a K of 9 resulted in a score of 0.07. A similar outcome was found for the sparse matrix before dimensionality reduction.

A search for the optimal number of clusters between 2 and 16 was conducted using Agglomerative Hierarchical Clustering with ward linkage. The results yielded Silhouette scores and Calinski-Harabasz scores, as depicted in Figure 3-3. Concurrently, a dendrogram was generated to confirm the number of clusters, as displayed in Figure 3-5.

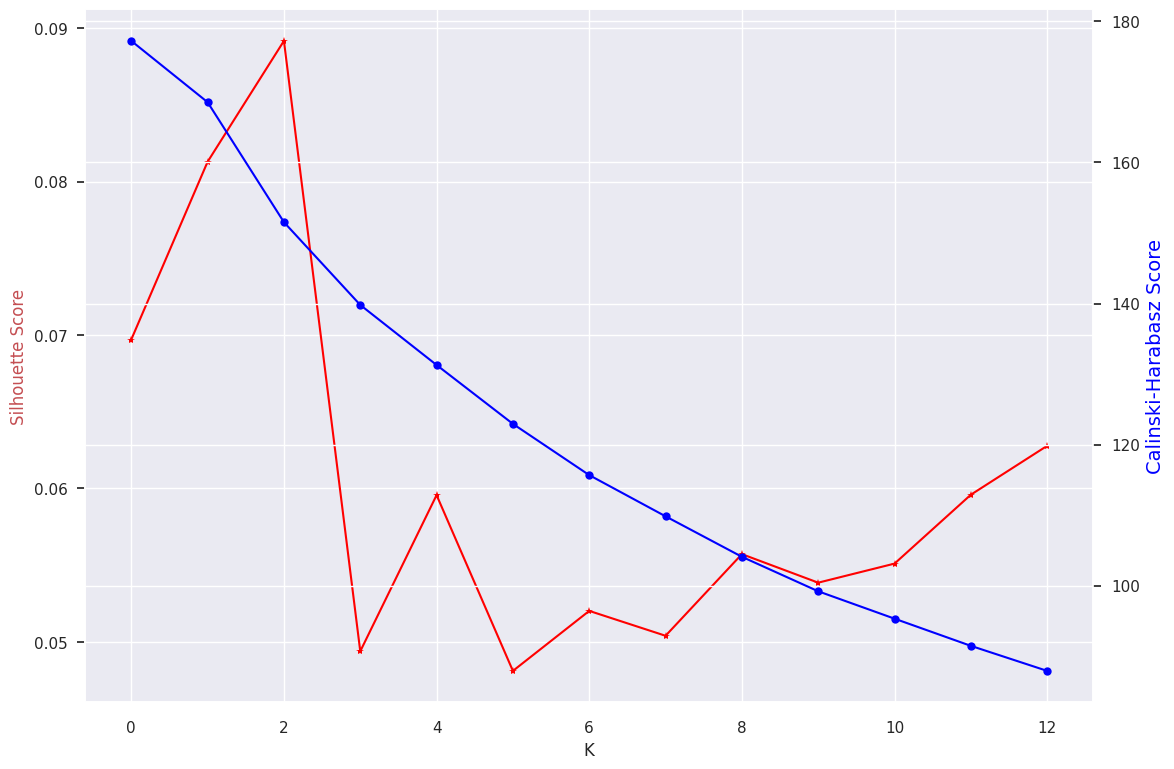


Figure 3-3 Silhouette Scores and Calinski-Harabasz Scores with respect to K

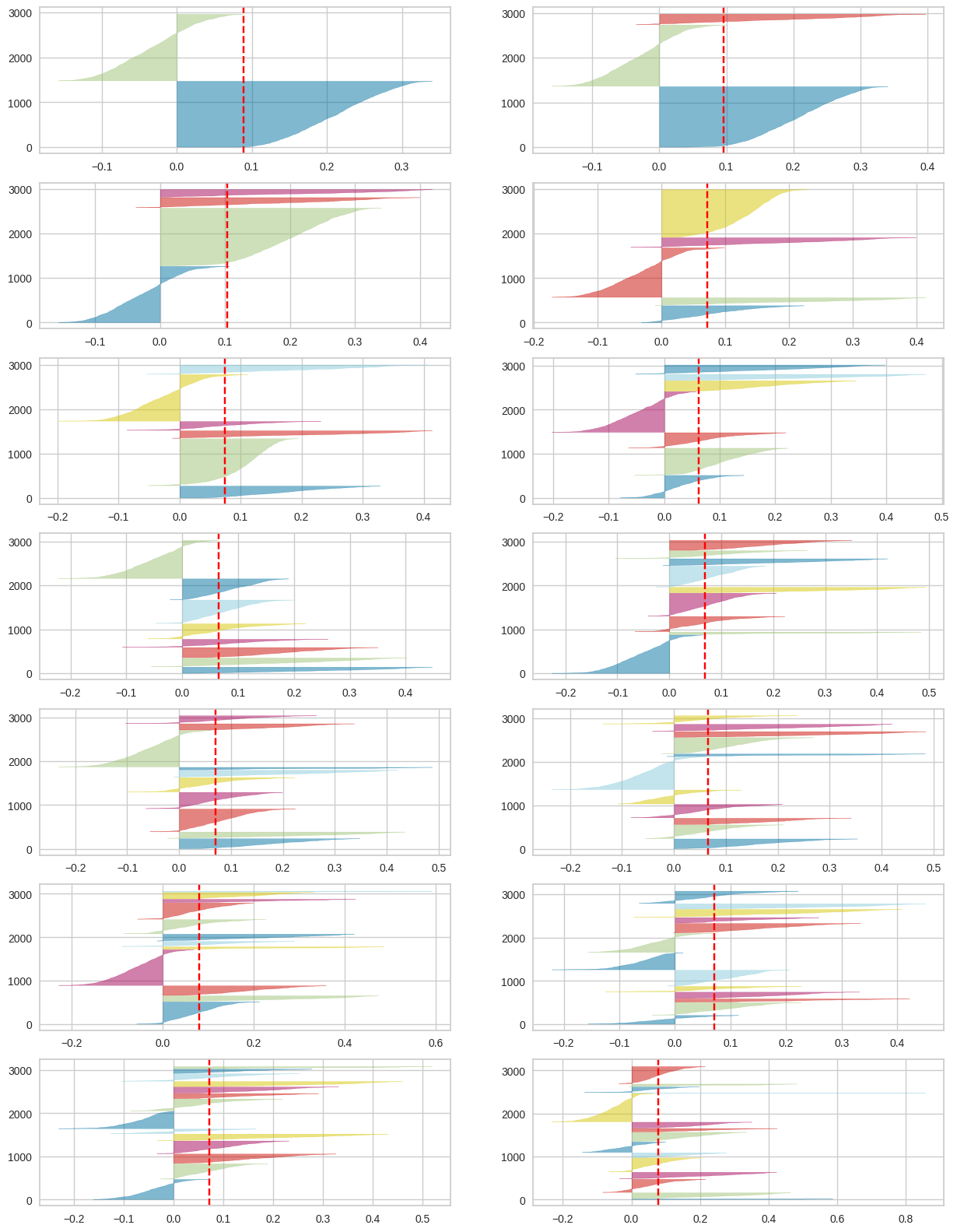


Figure 3-4 Silhouette Score with respect to K

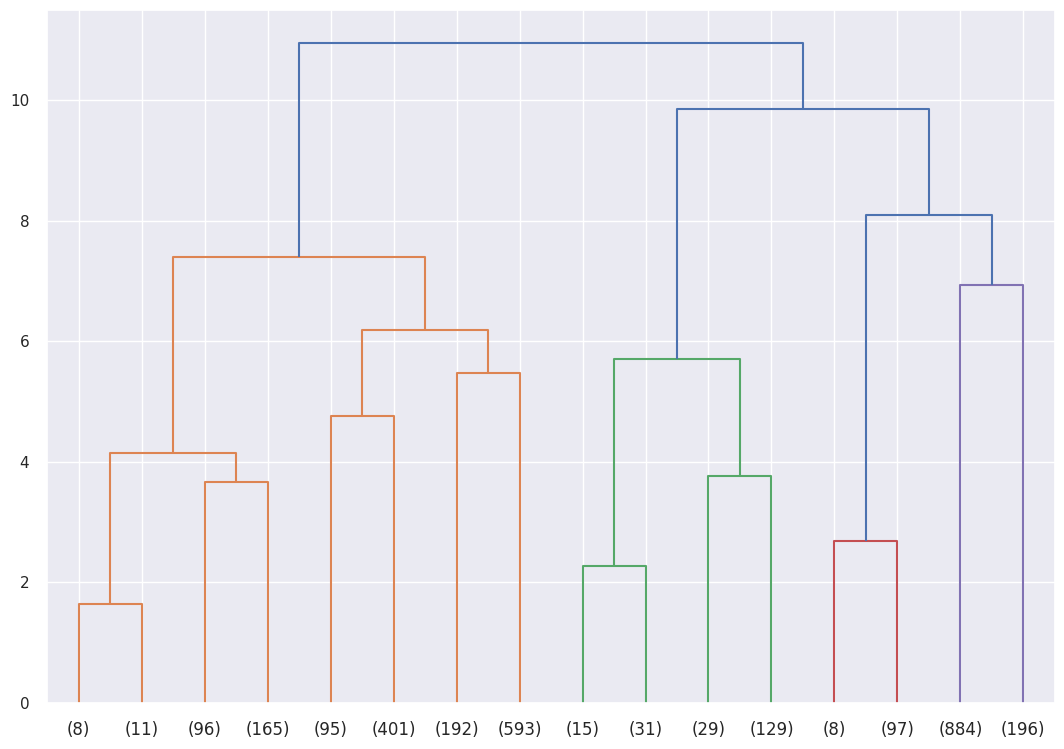


Figure 3-5 Debdrogram of Agglomerative Clustering

From Figures 3-4 and 3-5, it is evident that the scores remained low. Besides, the number of samples in hierarchical clusters was highly unbalanced compared to the one of K-Means. Consequently, the final model employs K-Means with a cluster count of 9. Table 3-1 presents the distribution of clusters.

Table 3-1 K-Means 9 Clusters Distribution

| Cluster | Count | Cluster | Count |
| --- | --- | --- | --- |
| 0 | 874 | 5 | 473 |
| 1 | 57 | 6 | 158 |
| 2 | 345 | 7 | 173 |
| 3 | 520 | 8 | 226 |
| 4 | 124 |  |  |

## Visualization

Applying t-SNE to the original 2950x4096 vectors, the clustering result is shown in Figure 3-6. This represents an initial 2-D spatial representation of companies in a specific year. Clusters 4, 7, and 8 appeared to be denser. The investigation of the underlying meaning of these clusters can be found in the topic modeling section.

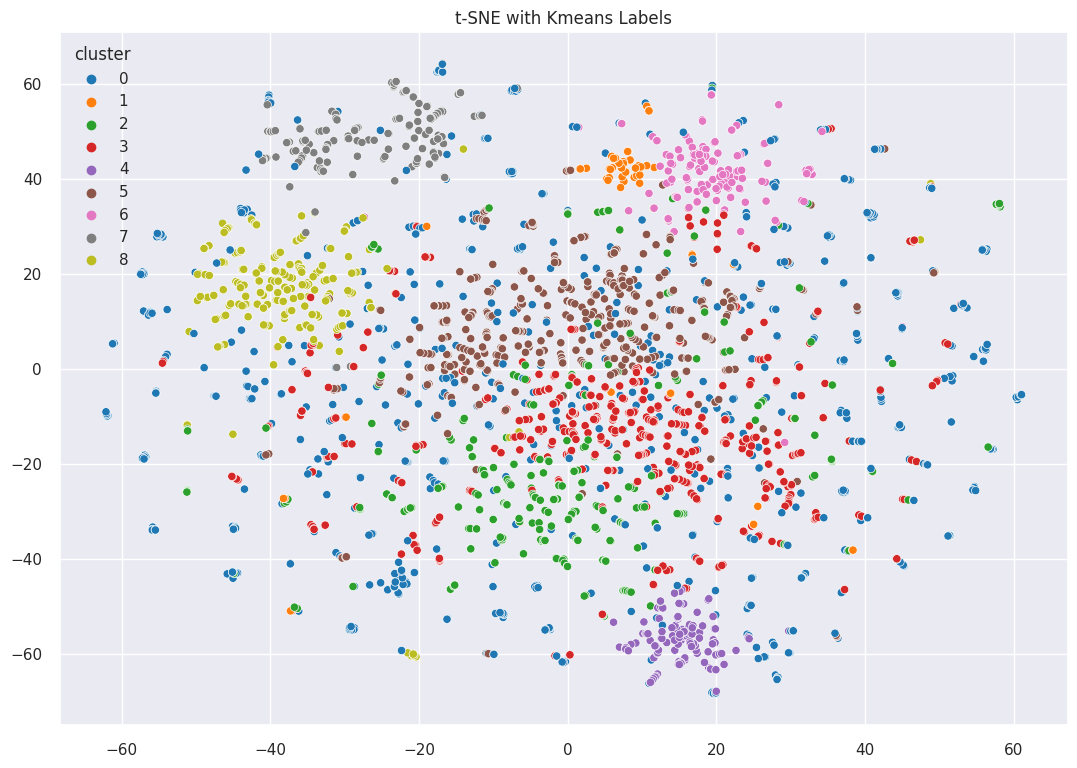


Figure 3-6 t-SNE with K-Means Labels

## Topic Modeling

The clusters' clean descriptions were vectorized using count vectorizers, and the keywords were extracted based on their frequency. Further results are displayed in the following chapter as Figure 4-1.

## Similarity Score

The raw 2950x4096 matrix was utilized to compute the pairwise similarity scores presented in Table 3-2. The distribution of these scores is illustrated in Figure 3-7, appearing to follow a skewed normal distribution. A similarity score of 1.0 suggests a shortcut for queries involving the same company in the same year.

Table 3-2 Sample Table of Pair-wise Similarity Score

| gvkey | year | gvkey\_r | year\_r | score |
| --- | --- | --- | --- | --- |
| 64768 | 2012 | 12141 | 2012 | 0.224809 |
| 64768 | 2013 | 12141 | 2013 | 0.708214 |

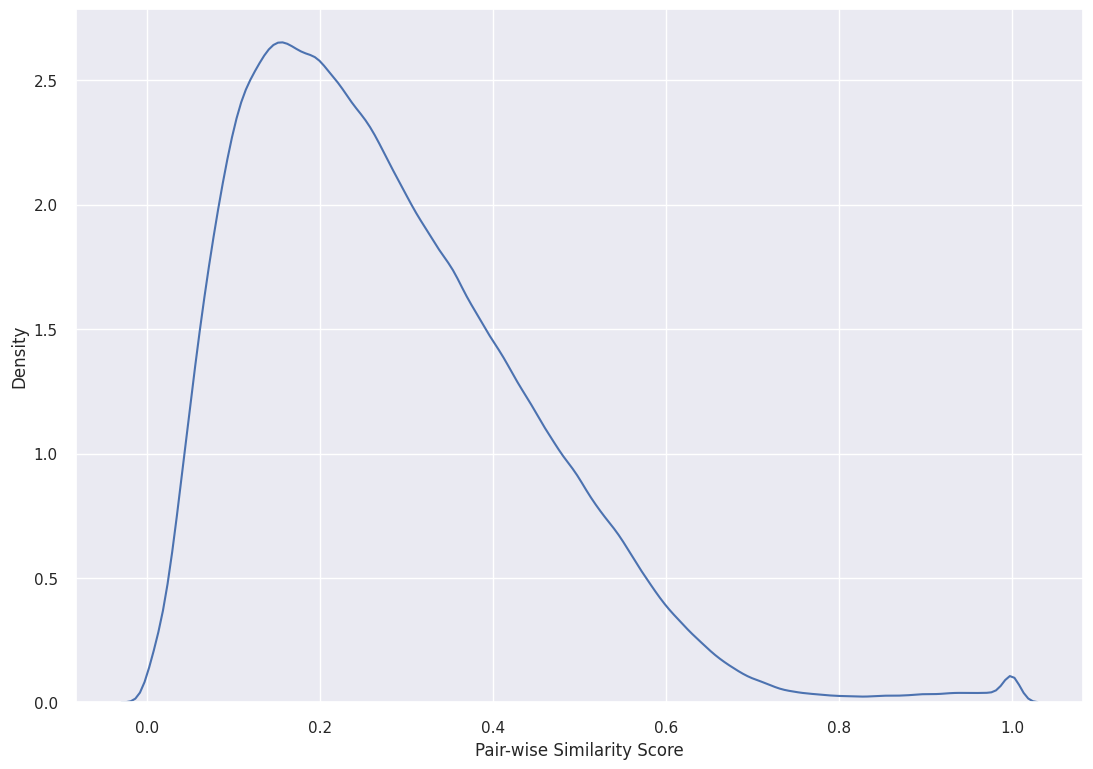


Figure 3-7 Distribution of Similarity Scores

# Result

Two datasets were created to support queries: Job Demand Yearly Inter-Firm Similarity Scores and Firm Yearly Job Demand Labels. The connection between the clusters' keywords and NAICS industry categorizations is illustrated in Figure 4-1, considering the number of firms. Denser clusters, such as No. 4, 7, and 8, appear to include a dominant production industry, which could explain the specific orientation of specialized labor demand. This implies that both the clusters and the applied analytics methodology for job post text are meaningful.

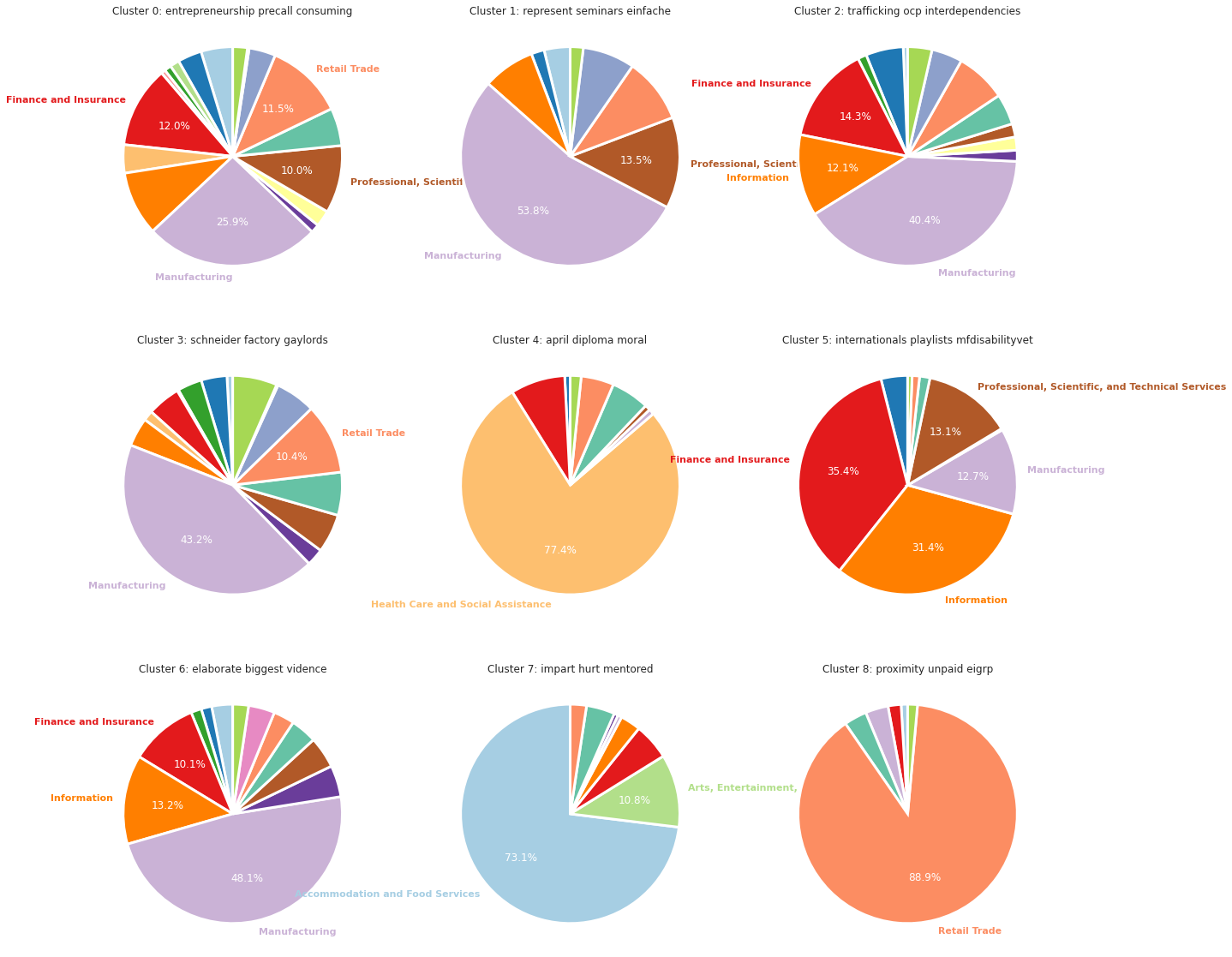


Figure 4-1 Job Demand Clusters with Production Category Breakdown

This observation also provides some evidence for the effectiveness of the job posting clustering methodology. Additionally, labor, production, and product markets may not always overlap. For instance, two firms could hire sales associates while engaging in distinct production processes and offering different products, potentially resulting in three separate categorization aspects.

# Conclusion & Next Steps

This project investigated the potential of text analytics in studying patterns in the job demand market for various companies. The analytical process encompassed data cleaning, vectorizing, dimensionality reduction, clustering, visualization, topic modeling, and similarity score calculation. The outcomes produced two datasets that could be valuable for future research: Job Demand Yearly Inter-Firm Similarity Scores and Firm Yearly Job Demand Labels.

The study has illuminated the possible connection between job demand patterns and industry categorizations. Interestingly, denser clusters, such as No. 4, 7, and 8, are primarily composed of specific production industries like Health, Food Services, and Retail, suggesting a potential link between specialized labor demand and these industries.

Future work can build on these findings by exploring the following directions:

* Analyzing a larger dataset with more companies, industries, years, and variables such as locations to validate the findings and uncover additional insights.
* Refining the text analytics methodology to improve the quality of the clustering and similarity scores by incorporating advanced natural language processing techniques, alternative clustering algorithms, or better scaling skills.
* Investigating the relationship between the identified job demand patterns and other firm-level characteristics, such as size, profitability, or growth, to better understand labor demand drivers that require domain knowledge.
* Focusing back on individual job postings to refine the clustering unit to a more granular level, where exploration of the content and fundamental components in the job market can be enhanced.

Overall, this project has taken an essential first step toward leveraging textual analysis to understand firm-level job demand characteristics. By building on the insights gained and addressing the current study's limitations, future research can continue to advance our knowledge in this field. As such, the analysis can contribute to developing more effective tools and strategies for labor market analysis, management, and comparison, as well as the associated implications.

# Reference

[1] Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, *124*(5), 1423-1465.

[2] United States Census Bureau. (n.d.). *North American Industry Classification System - NAICS*. United States Census Bureau. Retrieved March 29, 2023, from https://www.census.gov/naics/

[3] Mbah, R. B., Rege, M., & Misra, B. (2017, December). Discovering job market trends with text analytics. In *2017 International Conference on Information Technology (ICIT)* (pp. 137-142). IEEE.