Firm Categorization by Job Demand: A Text Analytics Approach

Student: Shuyang Lin | Instructor: Prof. Peter Haslag

# Introduction

In today's highly dynamic labor market, understanding the labor market competition is crucial for organizations to remain competitive, for employees to make informed career decisions, and for policymakers to craft effective labor regulations. This blog post presents the findings of a capstone project at the Data Science Institute of Vanderbilt University that uses text analytics to study job demand patterns among various companies. By analyzing a dataset of job postings, the project sheds light on the potential links between job demand and industry categorizations and identifies areas for further research.

# Importance of the Problem

The labor market is constantly evolving, with shifts in job demand driven by factors such as technological advancements, economic conditions, and societal trends. Previous studies in firm categorization are mainly from two aspects. The most popular one is from production. A good example is the North American Industry Classification System (NAICS). It’s predefined by the federal authority and updated manually, leading to a possible lack of flexibility. Recent studies like the Hoberg and Phillips text-based network industries classification utilized text clustering on product description text so that knowledge is learned from data rather than predefined. By analyzing job demand patterns to categorize firms, we can gain insights into the skills and occupations that are most in demand and identify trends that may have implications for workforce development, company strategies, and public policy. This project uses a text analytics approach to perform similarity score calculation and clustering, providing a basis for further research and analysis in this area. Professor Peter Haslag, Assistant Professor of Finance at Owen Graduate School of Management, Vanderbilt University, proposed the innovative concept.

# Data

This capstone project used a condensed sample from a proprietary dataset from LinkUp. This software company aggregates job listings and develops an employment media platform for labor demand assessment and research, as they cooperated with Owen Graduate School of Management. The dataset initially contained over 230 million job records from over 60,000 company career sites.

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 annually. It contains 607795 rows and six columns. Table 1 shows the data dictionary, and Figure 1 shows the distribution of year and onet\_occupation\_code:

Table 1: Data dictionary

| 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 |

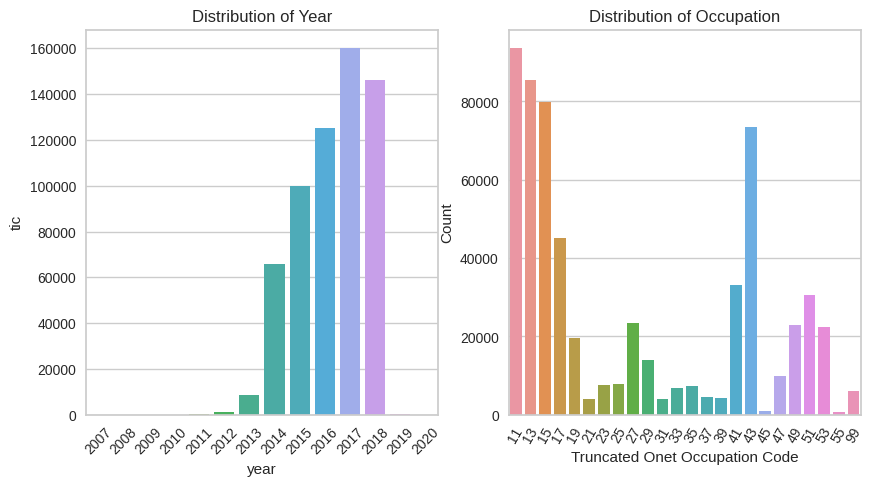


Figure 1: Distribution of Year and Onet Occupation Code

# Process & Challenges

Figure 2 shows the comprehensive pipeline of this project, and it starts with the raw data provided by LinkUp.

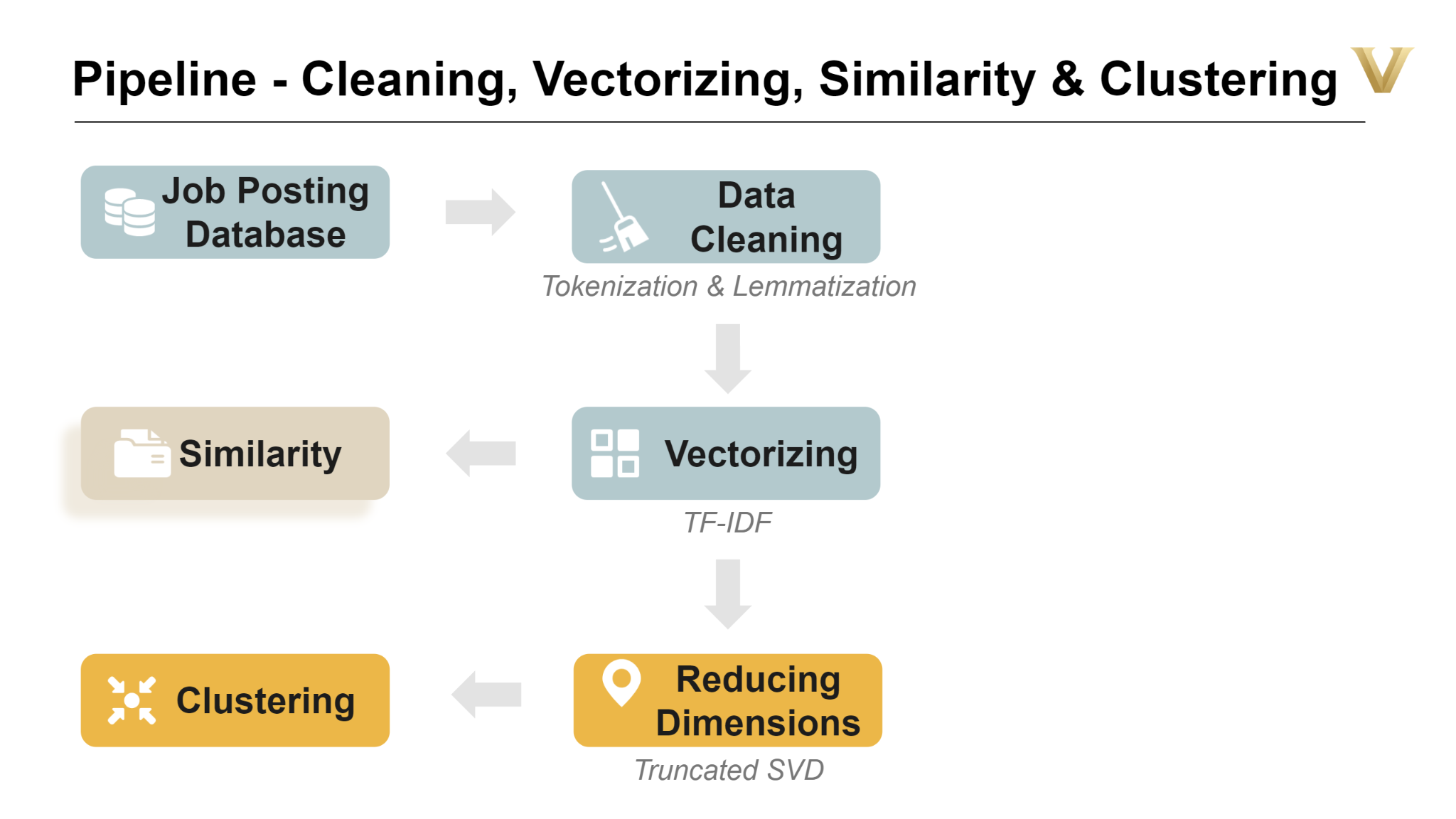


Figure 2: Project Pipeline

The first step, and also one of the main challenges faced in this project, was the data preprocessing and cleaning, given that job descriptions were raw text extracted from HTML pages. To optimize 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. Besides, the data cleaning process also involved tokenization and lemmatization in further investigating the meaning of words.

Once the data was prepped, TF-IDF was used to convert the textual information into vectors, enabling the analysis of firms' job demand patterns.

With the job demand vectors in hand, cosine similarities were calculated, resulting in scores that follow a positively skewed distribution. This calculation provided the first key outcome of the project: a measure of similarity between different firms' job demands.

Next, dimensionality reduction was tackled using truncated SVD to streamline the data for clustering. This process allowed for identifying meaningful patterns and groupings in the job demand data, which is crucial for understanding the underlying trends and relationships.

Another major challenge was selecting the most appropriate clustering algorithm for the data distribution. DBSCAN, K-Means, and Hierarchical clustering algorithms were utilized for comparison. While DBSCAN and Agglomerative clustering did not yield satisfactory results due to the density of the data, the K-Means method with a cluster count of 9 was ultimately chosen as the final model. The Silhouette scores of the different algorithms informed this decision. Figure 3 shows the result of K-Means clustering. Some clusters, like clusters 0 in blue, 3 in red, and 6 in pink, are dense and concentrated. At the same time, some, like cluster 2 in green and 4 in purple, are very sparse. This indicated that improvements are still needed on either the data or the clustering design.

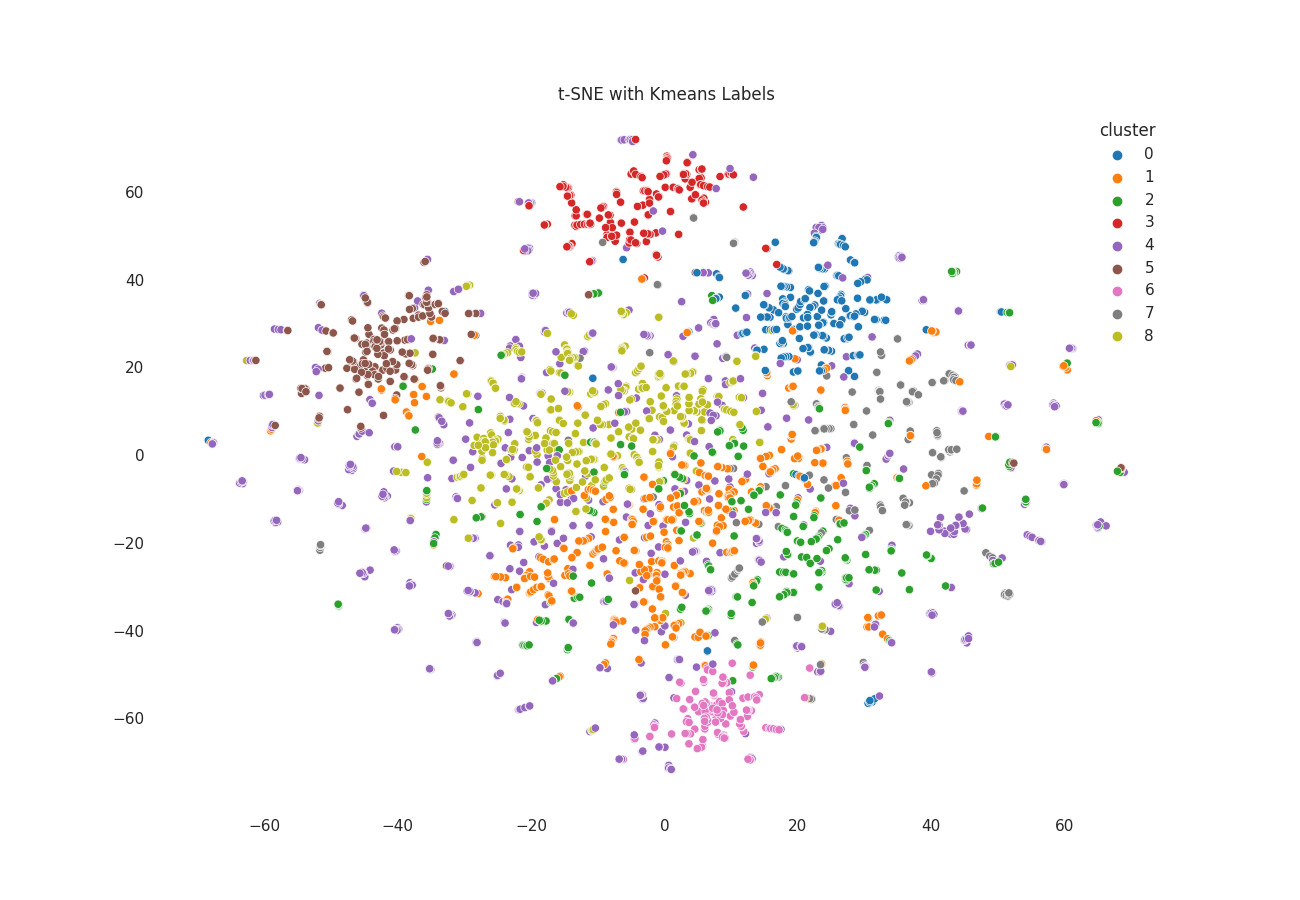


Figure 3: t-SNE Visualization with K-Means Labels

# Key Results

The project produced two datasets that could be valuable for future research: Job Demand Yearly Inter-Firm Similarity Scores and Firm Yearly Job Demand Labels. Figure 4 shows the positively skewed distribution of the similarity scores, which remains in a reasonable range.

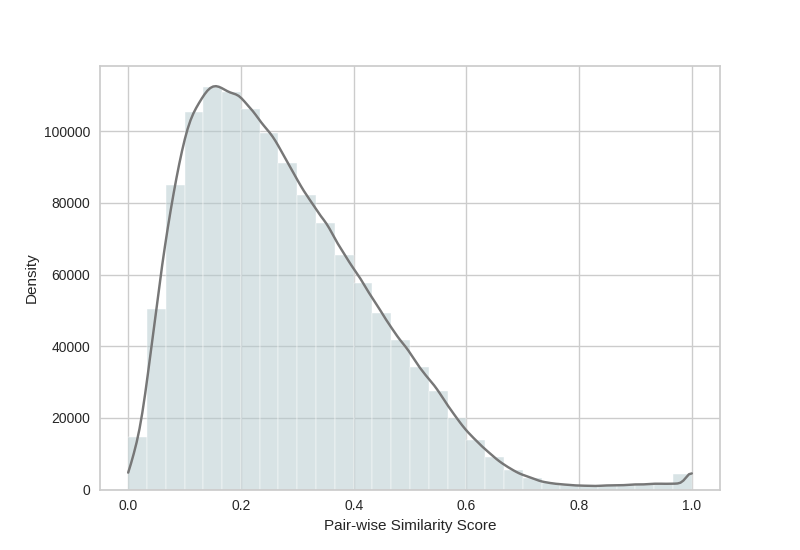


Figure 4: Job Demand Clusters with Production Category Breakdown

The analysis continued with a breakdown of the clusters by production categories, as shown in Figure 5. It revealed that denser clusters, such as No. 4, 7, and 8, primarily comprise specific production industries like Health, Food Services, and Retail. This finding suggests a potential link between specialized labor demand and these industries and provides evidence for the effectiveness of the job posting clustering methodology.

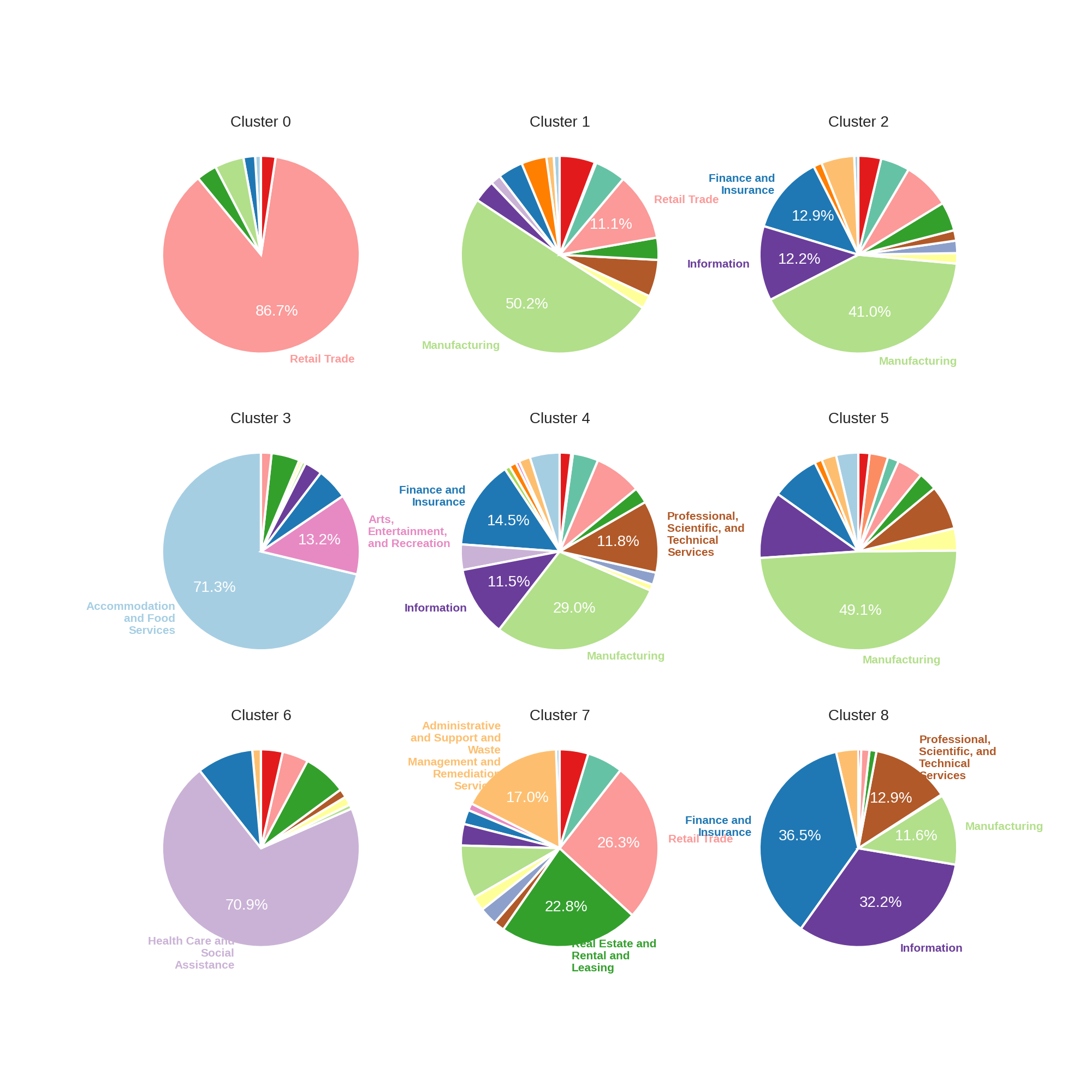


Figure 5: Job Demand Clusters with Production Category Breakdown

# Next Steps and Implications

The insights gained from this project can be further developed 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.

Building on the insights gained from this project, future research can continue to advance our understanding of firm-level job demand characteristics. This will contribute to developing more effective tools and strategies for labor market analysis, management, and comparison, as well as inform the associated implications for individuals, companies, and policymakers.