VANDERBILT UNIVERSITY

Job Posting Insights: Text Analytics & Firm Categorization

Student: Shuyang Lin

Instructor: Prof. Peter Haslag

April 24, 2023

Job Demand - New Perspective for Firm Categorization

Production

North American Industry Classification System[1]

Predefined categories based on production processes, employed by government

- Authoritative definition by federal
- Updating manually and outdated
- Requiring domain knowledge

Product

The Hoberg and Phillips Text-based Network Industries Classification[2]

Annual pairwise similarities based on product description text analytics

- Extracting knowledge flexibly
- Specializing only in product
- Not providing absolute label

Job Demand

Annual pairwise similarities & clustering labels by mining job posting data provided by

VANDERBILT Owen Graduate School of Management

LinkUp JOB MARKET DATA

- Feature separating from product
- Similar, yet apart from production
- Adapting to job market trends
- Measuring competitiveness

[1] United States Census Bureau. (n.d.). North American Industry Classification System - NAICS. United States Census Bureau.
[2] Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124(5), 1423-1465.

Pipeline - Cleaning, Vectorizing, Similarity & Clustering

Job Posting Database

Data Cleaning

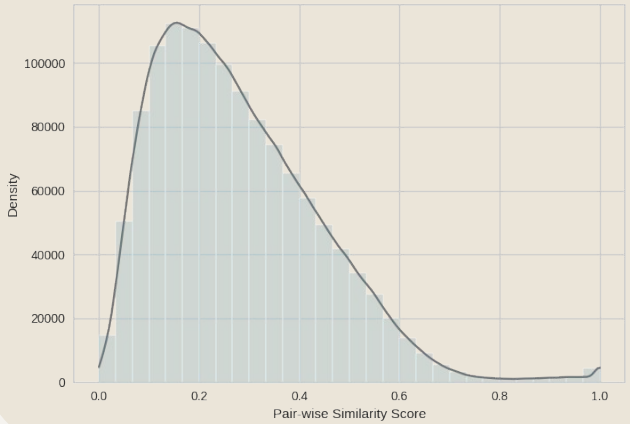
Tokenization & Lemmatization

Vectorizing

TF-IDF

Similarity

Distribution of Annual Pairwise Job Posting Cosine Similarity



- Data point unit *(firm, year)*
- Cosine similarity
- Same firm in different years
- Different firms in same year
- *Similarity of 1.0:*
  - *Same firm, same year*

Reducing Dimensions

Truncated SVD

Clustering

Challenge - Choosing Appropriate Clustering Algorithm

DBSCAN

Density-based

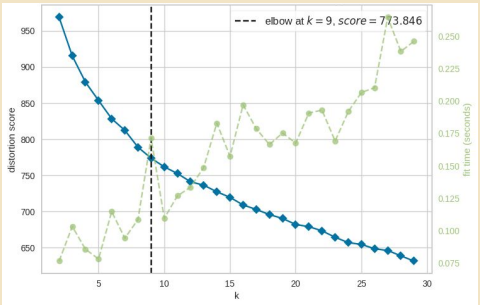
Parameters
- Epsilon - the radius
- Minimum Points

Results
- Grid search done
- Only 1 cluster
- 1‰ noise

K-Means

Simple
Scaled for large
Adaptive to new

Parameters
- *K* - elbow method



Distortion Score Elbow for K-Means

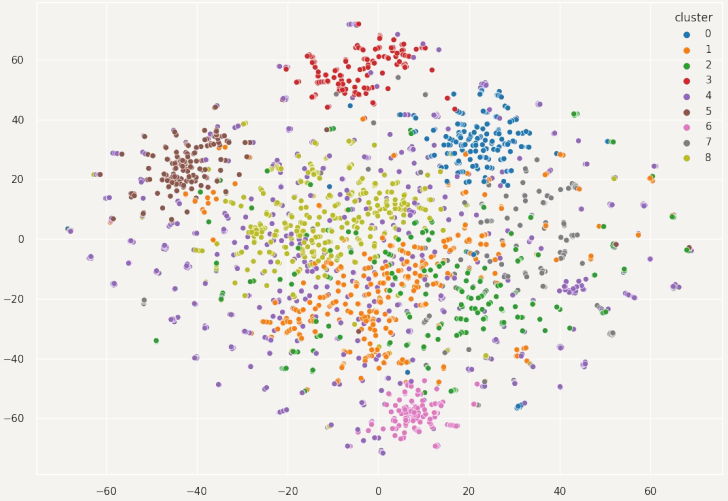
Agglomerative

Hierarchical

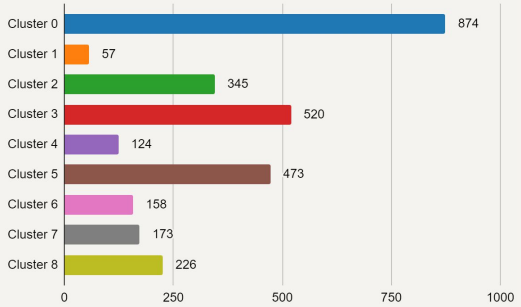
Parameters
- Ward method
- Number of clusters

Results
- Only 2 clusters
- Highly uneven

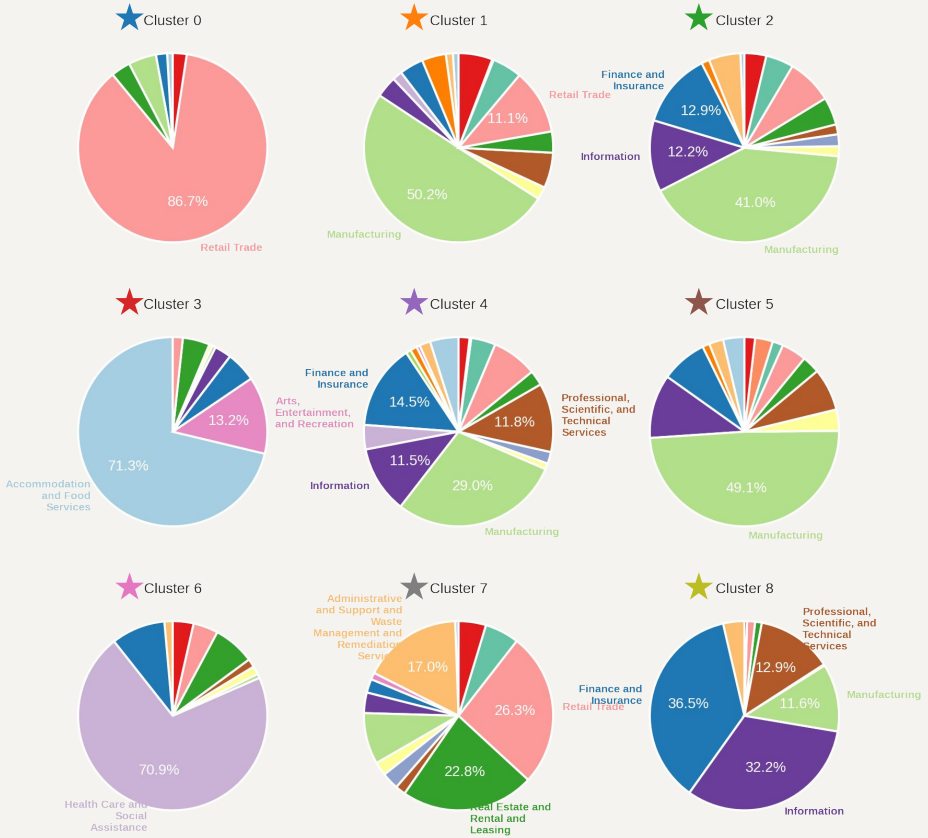
Job Demand Clusters Potentially Related to Production



2D Visualization of Job Demand Clustering by t-SNE



Number of Data Points in Clusters



Job Demand Clusters with Production Category Breakdown

Conclusions

- **2** output datasets as “products”

 *Similarity Scores*  *Cluster Labels*

- **1** pipeline using text analytics

 **Job Postings** →  **Clustering**

- Feasibility of text analytics on firm-level job market demand analysis
- Dense clusters of specialized labor demand in specific production groups

Limitations

- Constrained dataset size and scope
- Suboptimal clustering outcomes
- Analysis pending restructuring

Next Steps

- Introduce larger datasets
- Optimize analytics methodology
- Integrate additional variables
- Improve clustering approach

VANDERBILT UNIVERSITY

April 24, 2023

Thank you!