Final Project (ECON 691)
Aboli Khairnar

Analysis of Job Descriptions Data

Abstract

The job titles can sometimes be misleading. The job descriptions convey the actual duties which are consistent with the position level and job title. For this project, I analyzed around 677 job descriptions data. The contributions of this paper are: (1) Introducing methods to identify high skilled jobs from job description (2) I further explored this data to find homogenous occupation families.

1 Introduction

Job titles can be misleading!
The job titles can be of low quality and misleading. Low-quality job titles may lead candidates to believe the job post is spam. The job title can be used as a clickbait to attract a larger pool of applicants. Sometimes a great opportunity can be hidden by a misleading job title. The good job descriptions will reflect what the position may involve and require which helps the candidate assess if they are suitable for the position.

Low skilled jobs are at risk for automation!
In middle and low-skilled jobs, AI systems will complete the easily automated tasks while humans continue to perform those that cannot be automated. A high probability of automation may also be associated with the creation of new tasks and jobs though the productivity gains from adopting AI technologies, but these jobs and tasks will most likely be high-skilled.

Clustering jobs as per job families offer some benefits!
Identifying semantic connections between job skills, machine learning techniques create interesting and powerful possibilities. The occupation family not only helps human resources to develop a logical system to evaluate and differentiate positions within the organization but also helps the staff to explore career development opportunities.

2 Data Description

Job titles and job description data are collected from a proprietary API. The job descriptions are a short summary of essential responsibilities, activities, and skills for a role. An effective job description is supposed to provide enough detail for candidates to determine if they're qualified for the position.

O*Net offers a public occupational dataset. It assigns each job into a different job level (1-6) based on education and skills needed to perform specific job tasks. Higher the job level more skill and experience is needed to perform that job. This data can help to identify high skilled jobs. For this analysis, I collected 677 unique job descriptions data. Each job title is further labeled as high skilled and low skilled based on *ONET specification for job levels. I have labelled job levels higher than 4 as high skilled.

3 Objectives and Methodology

3.1 Preprocessing and Terms Weighing

This phase is aimed at removing meaningless features from job descriptions and retrieving relevant features from raw data records. In the proposed methodology, data preprocessing consists of three steps[3]:

A. Tokenization: It is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining.

B. Stemming (lemmatization): It is necessary to identify each word from using its stem form. As stemming reduces the number of unique vocabulary items that need to be tracked and speeds up a variety of computational operations.

C. Stopwords removal: Stopwords are nuisance and often do not carry much meaning. For many NLP purposes, stopwords removal is a common preprocessing step, some of these words are: the, a, of, in, etc.

3.2 Objectives

a. Classification:
   High skilled jobs are the jobs which need extensive training and education. We will train our NLP model to predict high skilled jobs just using job descriptions.

b. Clustering: Recommending similar jobs based on job description and experience

4 Results and Discussion

4.1 Classification

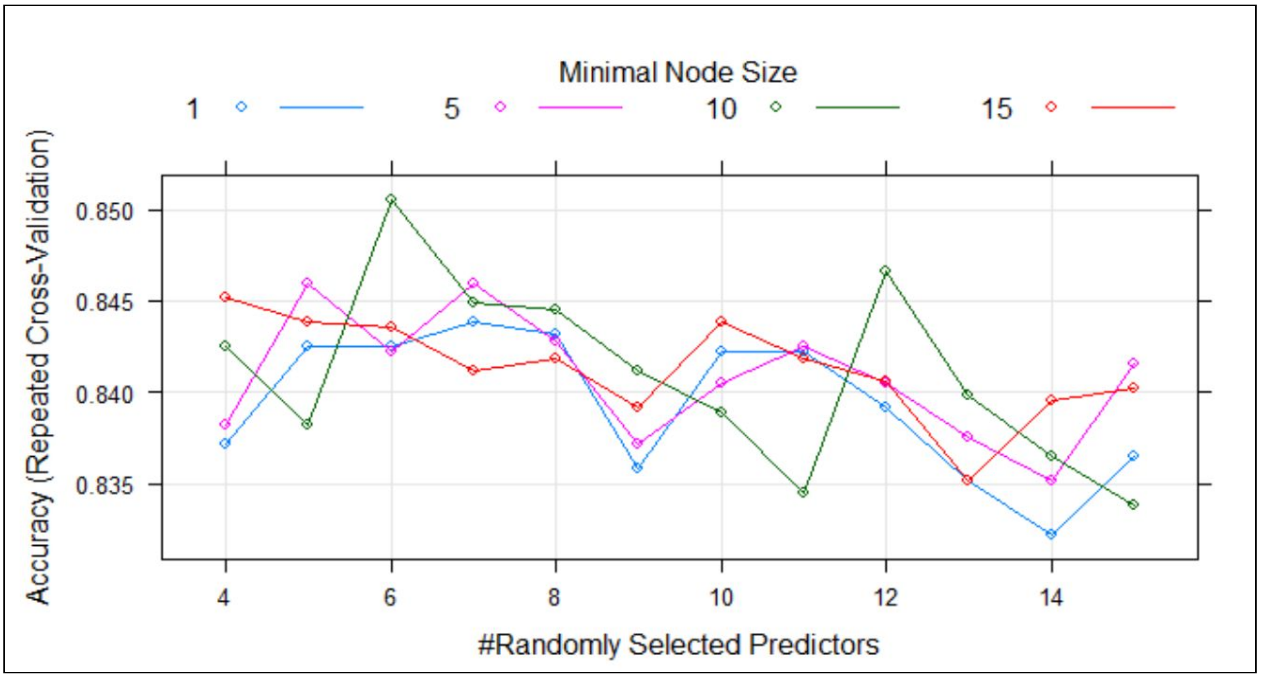
Data is splitted into training (80%) and testing set (20%). Table 1 shows the results for different models. This analysis uses bootstrapping and k-fold cross validation functionalities in the Caret package in R to get the robust estimate of the accuracy of various machine learning algorithms like Logistic Regression, Support Vector Machines, Naive Bayes, Random Forest, and Neural Network. Since classes are roughly balanced, I have selected AUC as a performance metric to compare different models.

Out of all five models, the Random Forest model performs the best in identifying high skilled jobs based on job description. The hyperparameters for the Random Forest are tuned using the TuneGrid approach in the Caret function. Figure 1 shows the accuracy across selected hyper parameters (Number of variables randomly sampled as candidates at each split & minimum node size) using the Random Forest algorithm.

Table 1 Model Performances

Model	AUC
Logistic Regression	0.7
Support Vector Machine	0.71
Naive Bayes	0.63
Random Forest	**0.79**
Neural Network	0.71

Figure 1 Hyper parameters tuning for Random Forest Classifier

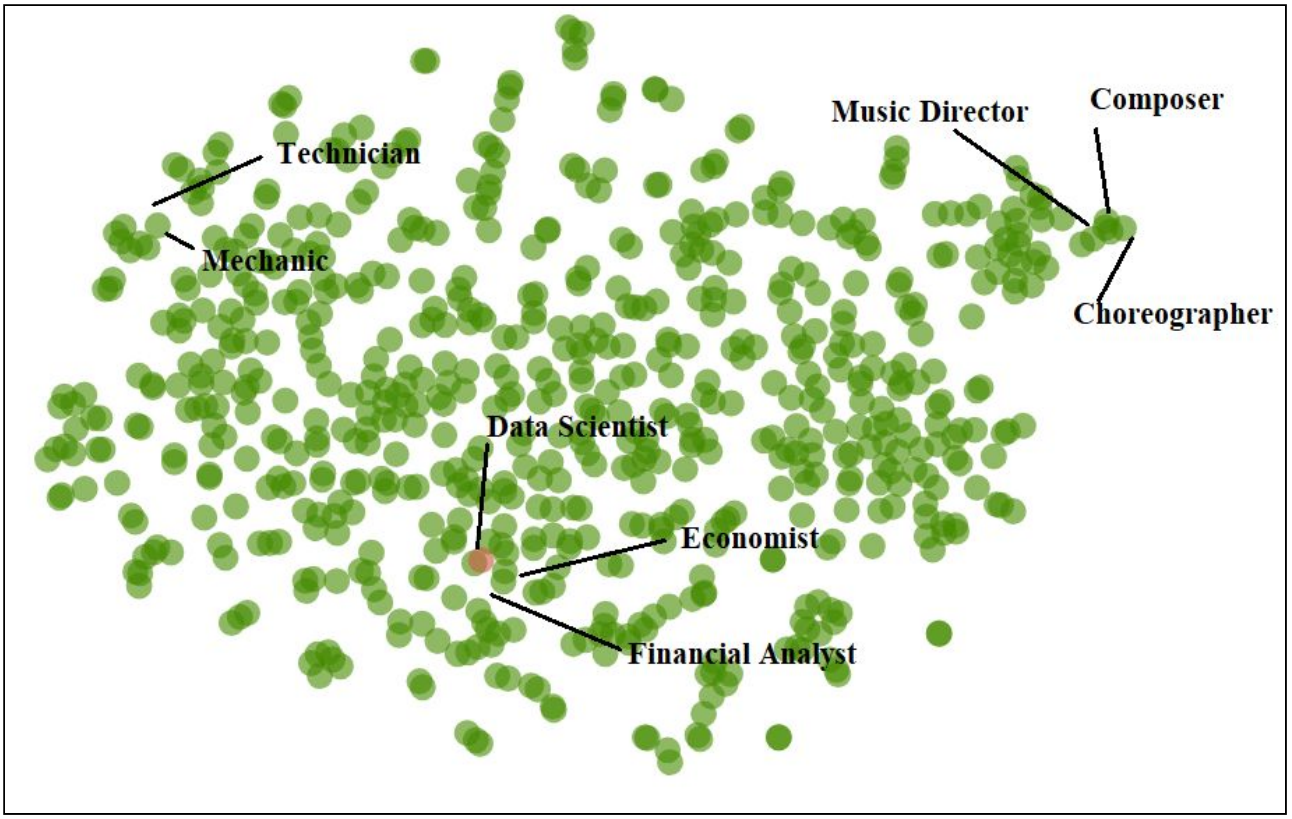


4.2 Clustering

4.2.1. Visualizing word vectors in 2D space

Figure 2 shows visualization of word vectors in 2D space. Similarity function used in this analysis is cosine similarity which uses the cosine of the angle between two vectors. In order to visualize the word vectors in 2D space, we need to implement dimensionality reduction technique. t-Distributed Stochastic Neighbor (t-SNE) Embedding is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It maps multi-dimensional data to two or more dimensions suitable for human observation. It is a new and popular dimensionality reduction algorithm used for plotting the vectors in 2D space. This algorithm is much more effective than linear Principal Component Analysis (PCA) which is incapable of understanding complex polynomial relationships between features.

Figure 2 Visualization of word vectors in 2D space

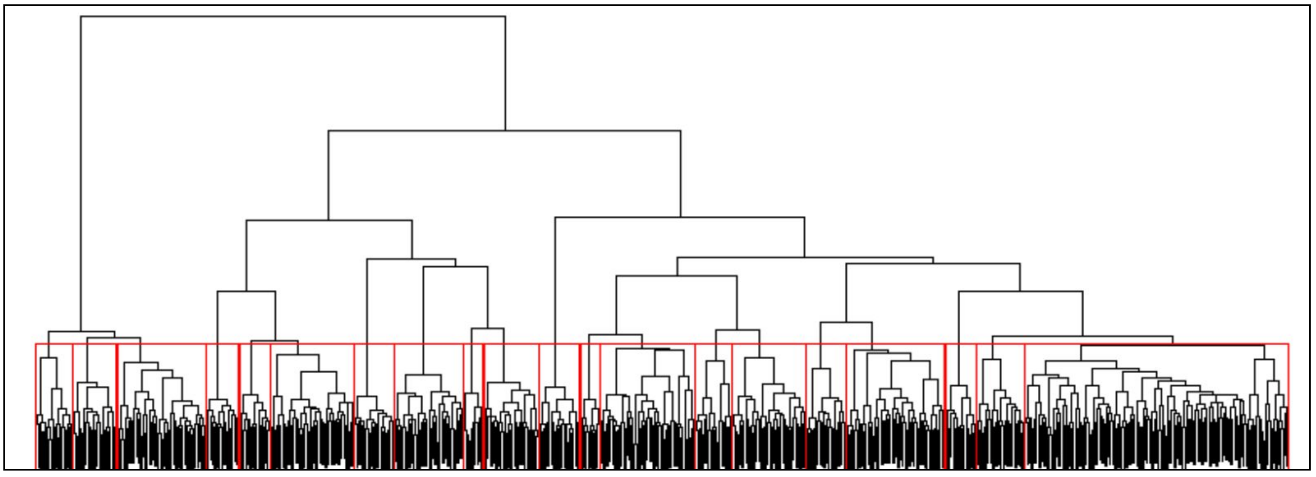


4.2.2. Implementing topic modeling techniques

Hierarchical clustering

Hierarchical clustering groups a set of observations based on some "distance", but instead of fixing the number of groups from the outset, the procedure is to start with each observation as its own cluster and then successively combine these clusters based on some aggregate measure of the distance between them. The distance measure between clusters is measured by 'Ward's Euclidean distance'. Figure 3 shows the dendrogram which is "cut" at a height to create 20 distinct clusters. This method creates a good visualization. However, there is no criterion to finalize the optimum number of clusters which is one of the drawbacks of this method.

Figure 3 Dendrogram for HC



Latent Dirichlet Allocation (LDA)

LDA topic model helps us to discover topics within a collection of job descriptions. In general, we use a collection of job descriptions as a corpus, and occupation families as the terms we want to discover. Since each topic is defined by a probability distribution with support over many words, it's hard to interpret topics. This method overcomes the drawback of Hierarchical Clustering method by providing the optimum number of topics are approximately 23 (least residual) which can be seen in the residual plot shown in Figure 4. Figure 5 shows the topic distribution for Tile/ Granite worker's job.

Few topics are listed in Table 2. Topic names are formulated by combining top six important words in specific topics.

Figure 4 Residual plot for LDA Topic Modeling

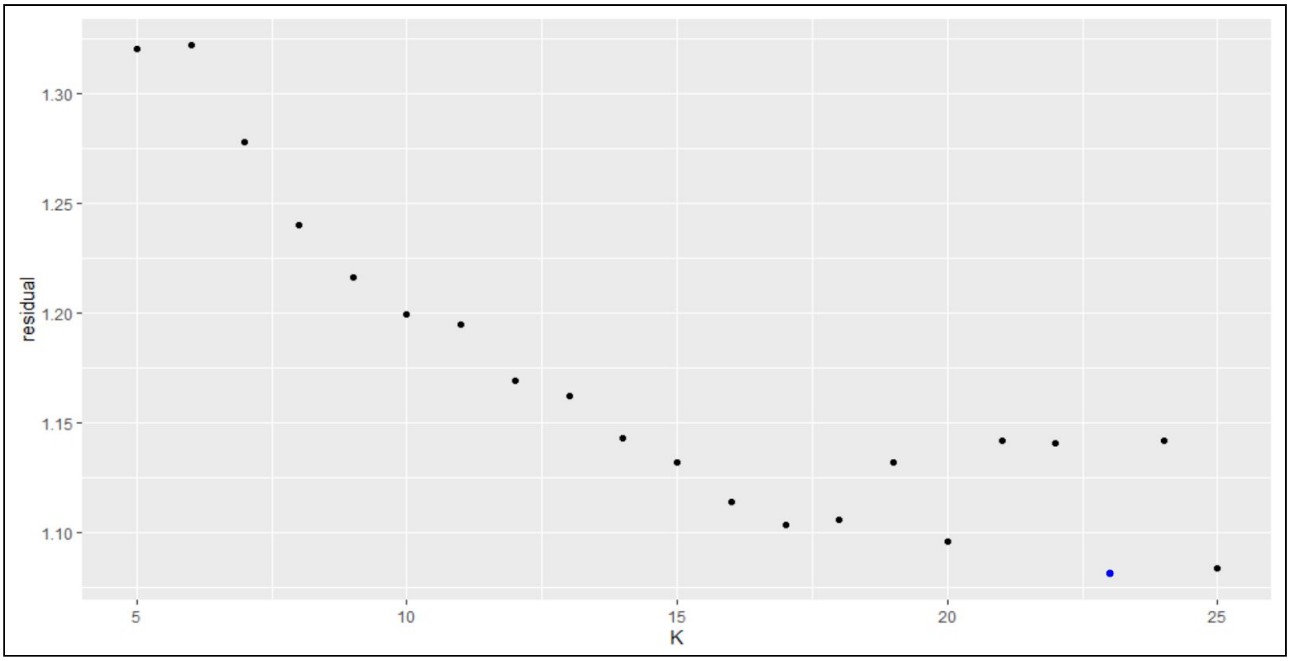


Figure 5 Topic distribution for job of Tile / Granite Worker

Table 2 Sample Topics

Topic	Name	Topic	Name
1	**Construction jobs** repair site build construction machine machinery	5	**Information Technology jobs** product computer engineer technology test project
2	**Academic jobs** school student teach study animal skill	6	**Corporate jobs** company system organization company_organization need maintain
3	**Healthcare jobs** patient medical work facility healthcare hospital	7	**Factory jobs** use make process chemical metal production
4	**Office jobs** work assist office help prepare file	8	**Advertising jobs** business sale market product advertise sell

Figure 6 Job path network

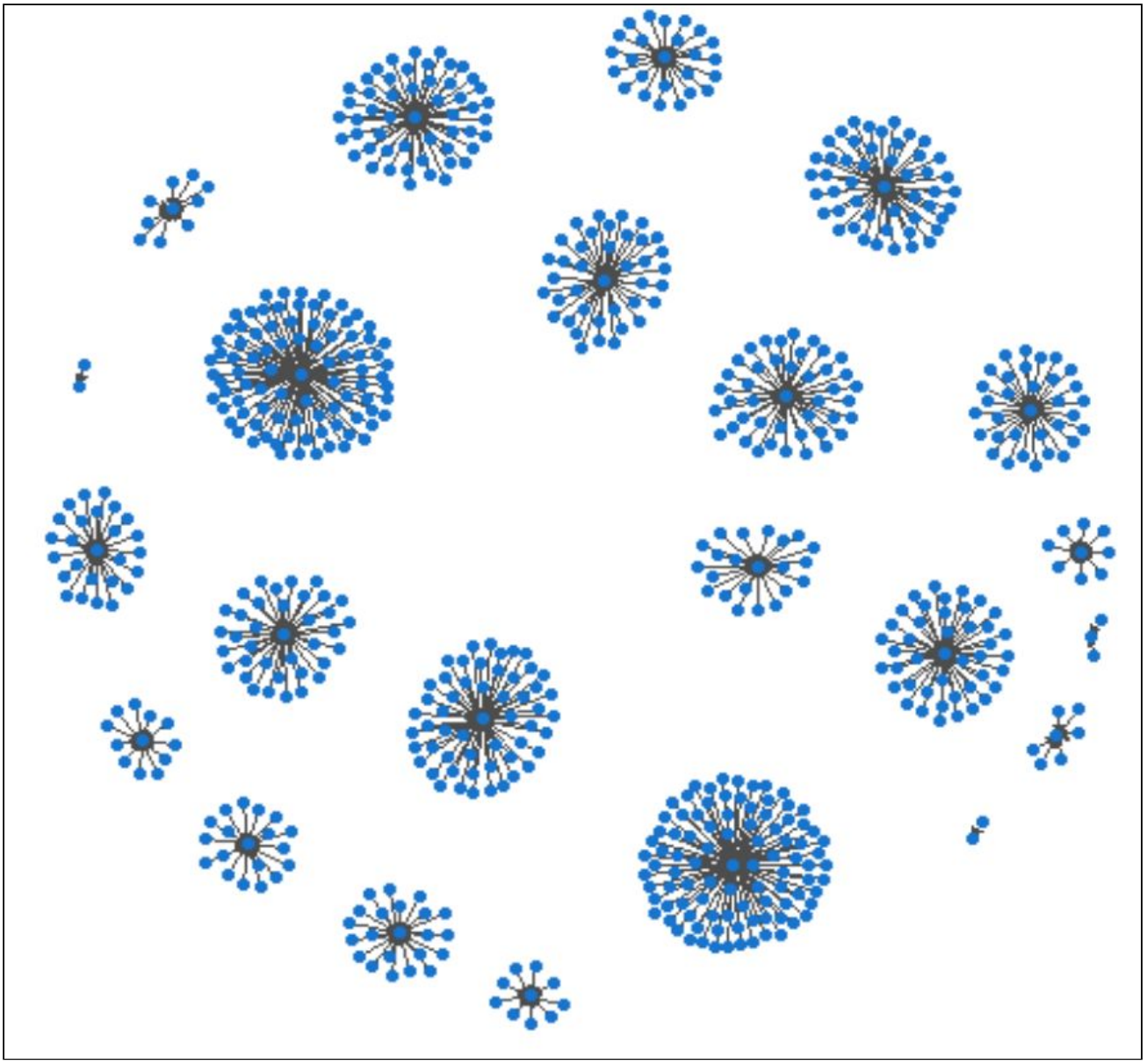
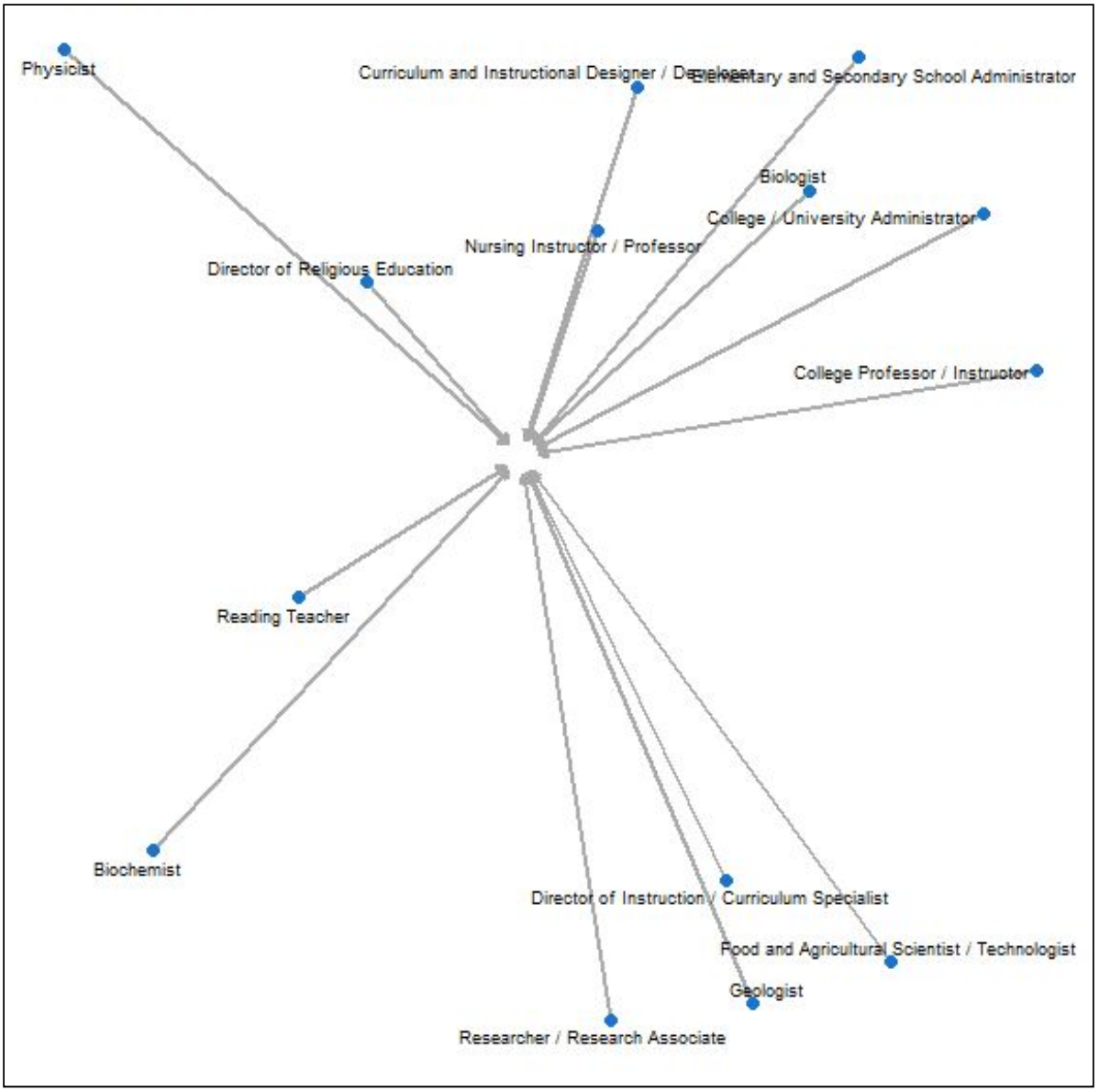


Figure 6 shows the job path networks for all 23 clusters. Each cluster represents different occupation families. Figure 7 & 8, shows job networks corresponding to Health care (which is one of the biggest clusters) and Academic jobs (one of the smallest clusters) respectively. We can conclude that this analysis has successfully identified homogeneous occupation families based on job description.

Figure 7 Job path sub network 1 (Health Care jobs)

Figure 8 Job path sub network 2 (Academia jobs)



5. Job Recommendation Framework

Clusters can be further divided into sub occupation families by applying job level and experience filters

For e.g. Occupations which are of job level 1 (comparatively low skilled) and needs less than 2 years of experience (Home Health Aide, Dietary Aide, Pharmacy Aide, Patient Transporter, Occupational Therapy Aide) creates a homogenous sub occupation family.

Limitations

The proposed job recommendation framework works well with low skilled jobs. But it is less accurate for recommending more specialized jobs (high skilled jobs).

5 Conclusions and Future Work

This paper presents the results for job posting analysis. It also presents a comprehensive set of methods for classification and clustering job descriptions data. This study investigated different types of models for classification: Logistic Regression, Support Vector Machines, Naive Bayes, Random Forest, and ANN. We identified that the Random Forest classification algorithm performs the best to identify high skilled jobs based on job description with AUC score of around 0.78. For topic clustering, this paper investigates two approaches: Hierarchical Clustering (HC) as well as LDA approach, and demonstrates how LDA outperforms HC for the job descriptions data. The generated subnetworks can help to create a broader level job recommendation system by adding an extra filter for experience requirements. In future work, I want to expand its research beyond clustering and give more emphasis on generating career paths/ scope based on given job description and experience within each cluster.

References

1. https://www.bruegel.org/2020/06/artificial-intelligences-great-impact-on-low-and-middle-skilled-jobs/
2. https://nycdatascience.com/blog/student-works/using-machine-learning-measure-job-skill-similarities/
3. Chokor, A, Naganathan, H, Chong, O & El Asmar, M 2016, 'Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning', Unknown Journal, vol. 145, pp. 1588-1593.
4. https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/
5. https://www.onetonline.org/