

Analysis of Job Descriptions Data

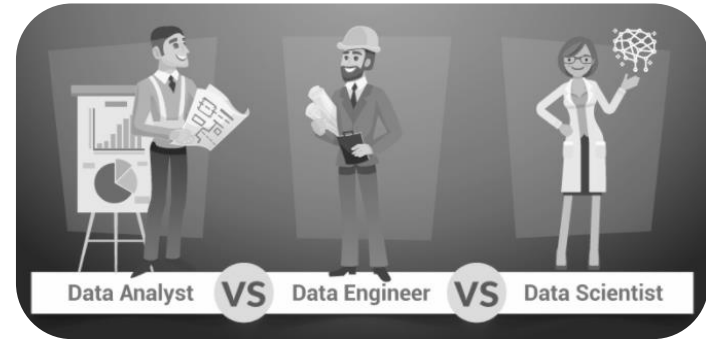
Final Project (ECON 691)

Aboli Khairnar

Why to analyze job descriptions data?

Job titles can be misleading!

- May not provide an accurate description
- Uninformative (too brief)
- Low quality
- Clickbait to attract a larger pool of applicants
- Inconsistent across organizations






Objectives


Classification

Train NLP model to identify high skilled jobs just using job descriptions.

- 
- Low skilled jobs are at risk for automation
 - High Skilled jobs have higher pay

Clustering

Recommending similar jobs based on job description and experience

- 
- Creates homogenous and broader occupation families
 - Develops a logical system to evaluate and differentiate positions within the organization
 - Enables the staff to explore career development opportunities



Data



Total 670 unique job titles and job descriptions covering all sectors

The job descriptions are a short summary of essential responsibilities, activities, and skills for a role.



Job level data is collected from O*Net

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.

Assumption: An effective job description provides enough detail for candidates to determine if they're qualified for the position.



Data Preprocessing

Aim: Removing meaningless features from job descriptions and retrieving relevant features from raw data records.

A. Tokenization:

- Breaks a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.

B. Stemming (lemmatization):

- Identifies each word from using its stem form.
- Reduces the number of unique vocabulary items that need to be tracked
- Speeds up a variety of computational operations

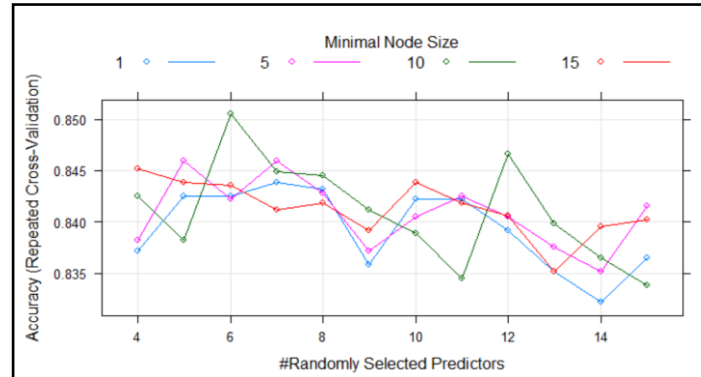
C. Stopwords removal

- Using stopwords dictionary to get rid of words that does not provide much information

Results

I. Classification

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



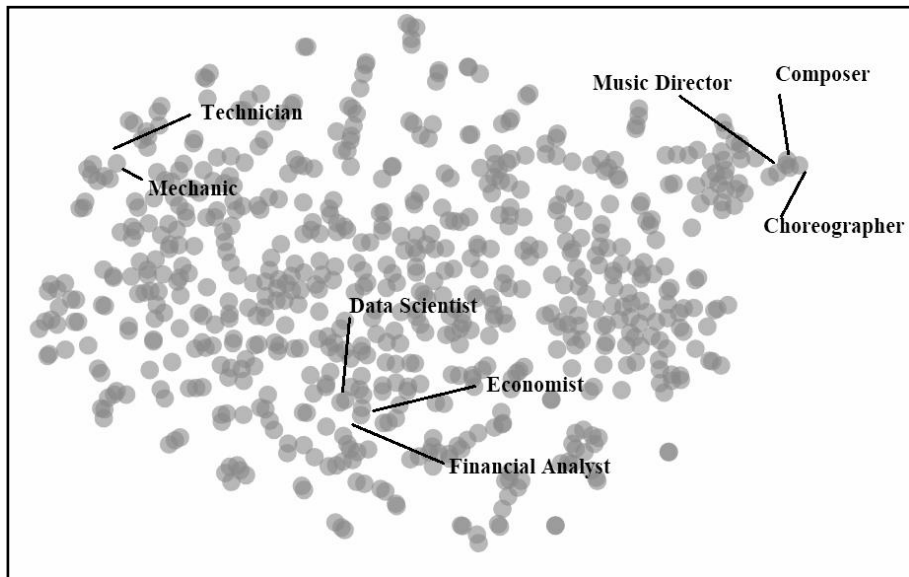
Model characteristics:

- ❖ 80-20 Split in Training and Testing
- ❖ Caret Package: Resampling and K fold cross validation

Performance Metric:

AUC score (classes are nearly balanced)

II. Clustering: Visualizing word vectors in 2D space



Cosine Similarity:

Uses the cosine of the angle between two vectors.

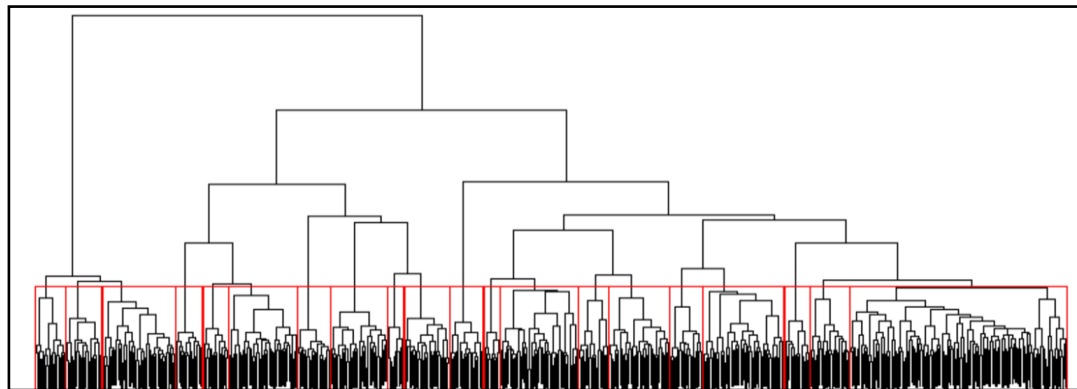
Dimensionality Reduction:

- t-Distributed Stochastic Neighbor Embedding (t-SNE)
- Non-linear dimensionality reduction algorithm used for exploring high-dimensional data.
- Maps multi-dimensional data to two or more dimensions suitable for human observation.
- More effective than linear Principal Component Analysis (PCA) which is incapable of understanding complex polynomial relationships between features.

II. Clustering: Topic Modeling

Hierarchical clustering

- Groups a set of observations based on “distance”
- Starts with each observation as its own cluster and then successively combine these clusters based on some aggregate measure of the distance between them.
- Distance measure = ‘Ward’s Euclidean distance’.
- Creates dendrogram which can be “cut” at a height to create distinct clusters.
- Creates a good visualization.

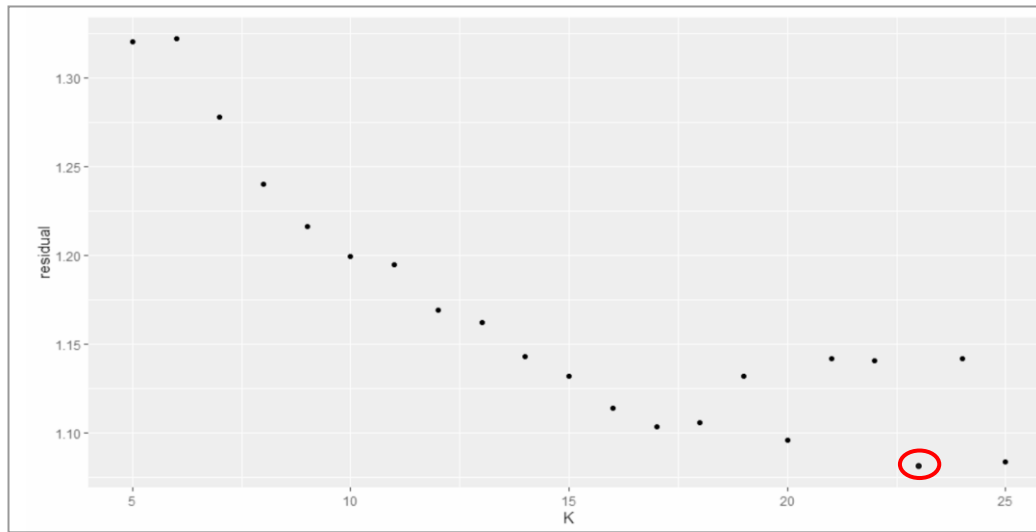


However, there is no criterion to finalize the optimum number of clusters which is one of the drawbacks of this method.

II. Clustering: Topic Modeling

Latent Dirichlet Allocation (LDA)

- Discovers topics within a collection of job descriptions.
- Each topic is defined by a probability distribution with support over many words
- 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



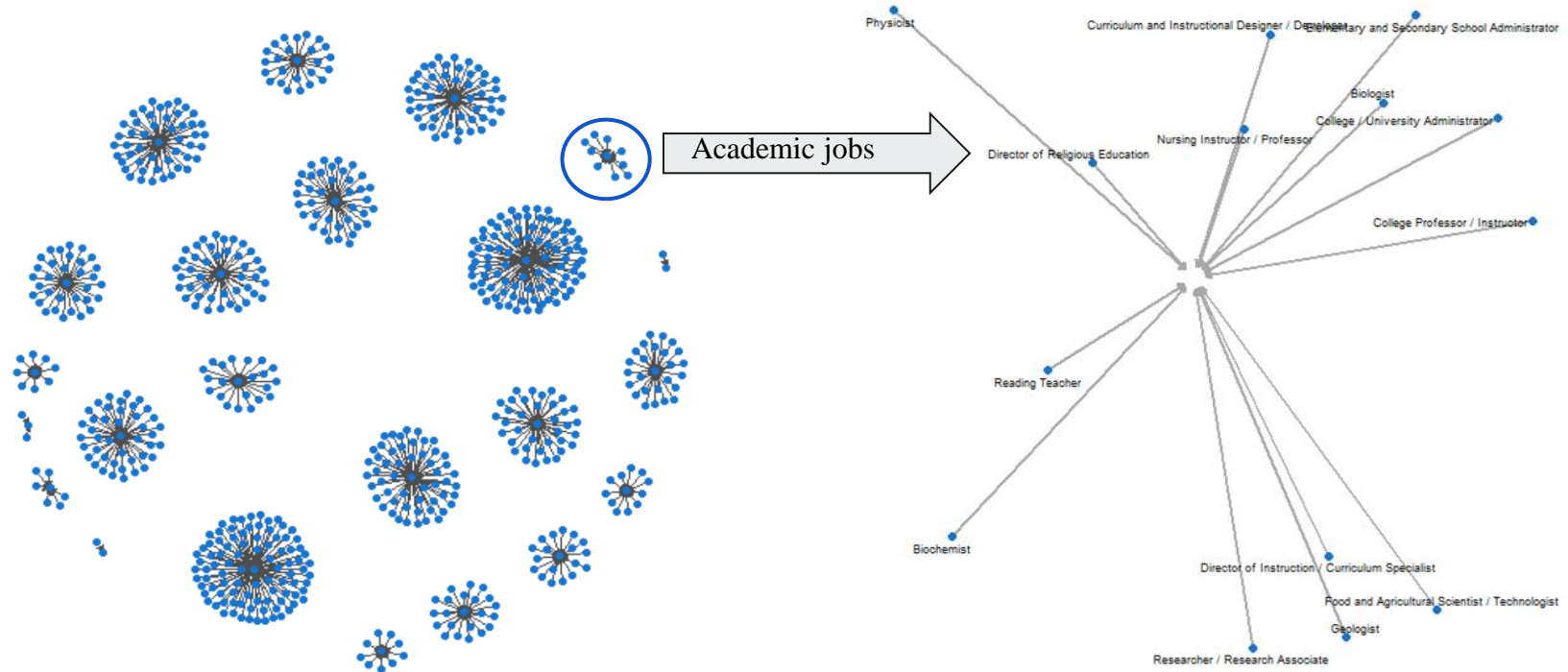
Topic distribution for job of Tile / Granite Worker



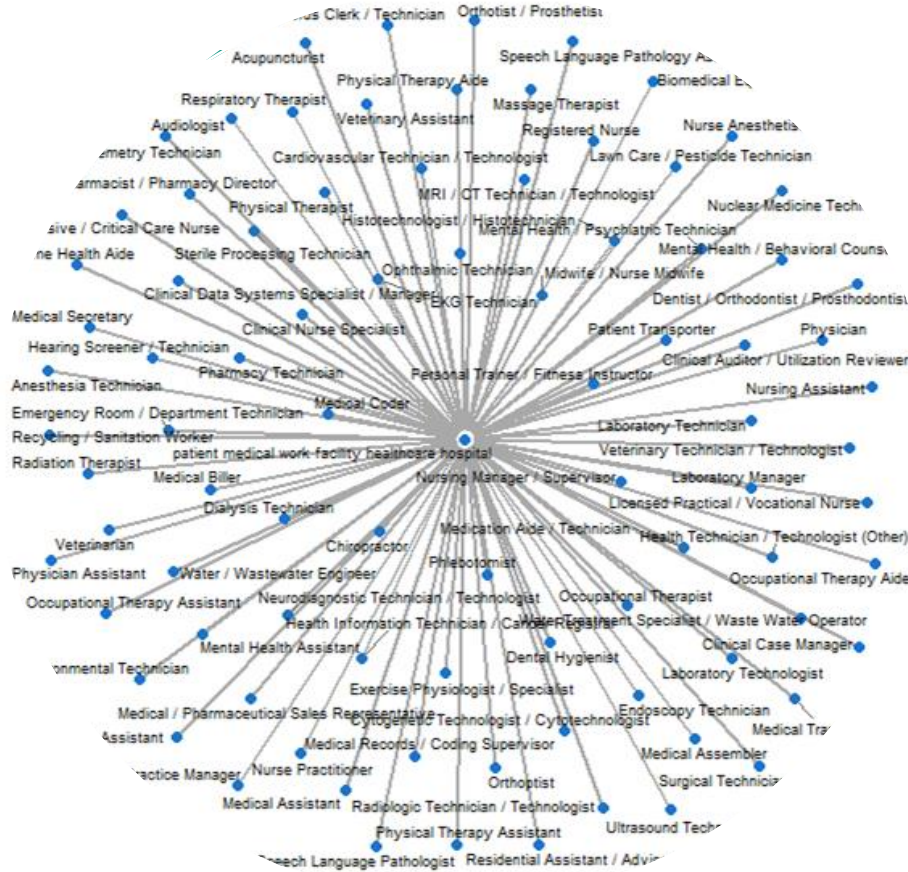
Sample Topics

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

Network of occupation families



Healthcare jobs



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.
- Less accurate for more specialized jobs (high skilled jobs)



Future Scope

- Expand the research beyond clustering
- Generate career paths/ scope for each job
- Create key skills from job description
- Create Risk to Automation Score for each job



Thank you!