Final Project Prospectus (ECON 691)

Date: October 16, 2020

Author: Aboli Khairnar

Objectives:

The research will try to answer the following questions:

1. Classification: Can we identify high skilled jobs using only job description?
2. Clustering: Can we recommend similar jobs given the job description?

Approach:

Objective 1:

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.

Steps:

1. Preprocessing job description data
2. Labeling jobs as High skilled based on experience and education requirement
3. Splitting data into the train (80) and test set (20)
4. Tokenizing training dataset, creating vocabulary and DTM_train (Document Term Matrix for Training dataset)
5. Fitting a logistic regression model using the train set
6. Tokenizing testing dataset, creating vocabulary and DTM_test (Document Term Matrix for Testing dataset)
7. Checking the performance of the model on the Test set.
8. Hyperparameter tuning

Objective 2:

Clustering: Recommending similar jobs based on job description and experience

Steps:

1. Creating Document Term Matrix (DTM) and Document Feature Matrix (DFM) for the whole corpus of a job description
2. Calculating Cosine distance and implementing dimensionality reduction technique
3. Implementing Hierarchical Clustering to find optimum job categories for given corpus

Final Project Prospectus (ECON 691)

4. Compare obtained job categories with optimum topics (Job categories) obtained from the Latent Dirichlet Allocation (LDA) model by implementing *the Ksearch* function
5. Clustering job titles based on associated topics and experience identify similar jobs.
6. Visualizing sample clusters

Data sources:

- Job description data will be collected from proprietary API
- O*Net also offers job level data which can help to identify high skilled jobs

Challenges:

- Job descriptions are a short summary of job duties. It might not be extensive enough to to classify high skilled jobs and recommending similar jobs, can be challenging.
- Identifying the best method to identify the optimum number of job categories. Choosing between the distance-based method or Topic modeling?