**Report**

**Aim:**

The aim of the project was to map job titles scraped from indeed website for specific companies to the given functional areas

**Project:**

The project is divided into three parts

1. Scraping required data
2. Training, tuning parameters using grid search and saving model
3. Classifying the company wise job titles to the functional areas

**Overview**:

First data was scraped from indeed website for random companies for given functional areas and is used to train the classifier model using supervised learning. After tuning, testing and validation the trained model with the best parameters is saved. The company wise data is passed and a output csv file called final\_results.csv containing the unique company names and jobs per functional area is outputted.

**Note:**

    The final\_results.csv has more the 100 companies given is because of various names the company uses while posting and some companies are excluded they did not produce any job posting results.

**Algorithms used:**

Naïve bayes, logistic regression, svm and Random forest

Among the above Svm performed better and hence was chosen with alpha of 0.001 and showed a best cross validation score of 0.8626226583407671

**Best parameters for the constructed pipeline are**

{'classifier\_\_alpha': 0.001, 'vectorizer\_\_ngram\_range': (1, 1)} where the n\_gram is used for the tfidf vectorizer used in the pipeline.

For results of grid search for the above mentioned algorithms see the attached grid search text file.

**Observations:**

1. It is observed that using the job title for training showed accuracy of 70 percent and is hence excluded and hence job description was taken into consideration

2.Among different types of preprocessing steps for text data bag of words performed well in algorithm spot check.

3.while searching for jobs by functional area it is observed that leadership and other gave  ambiguous results and were causing poor best accuracy of 75 percent in algorithm spot check. hence search is done for specific keywords involving supervisor, visionary for leadership and delivery for other as the companies list includes home delivery companies like eBay, amazon etc.

4. metrics like accuracy is chosen because the data is not very skewed.

Though other metrics like precision, recall and f1 score are also considered

**To Run:**

run the **classifier\_data.py** file tooutput the required final\_results.csv file which has the sample description of output file and the result.csv file has the more extensive results with job description

**preprocessing steps:**

1. Removed stop words
2. Removed punctuations
3. Removed other noisy data
4. Used Tfidf vectorizer with user defined tokenizer called spacy\_tokenizer

Kindly refer the code for more elaborate details and flow of the code.

**End notes:**

Kindly refer to the code for description of functions written on top of functions for more information about the project. The accuracy of the model could be increased by doing more extensive grid search by trying out different parameters and by scraping more data which was limited in this case because of the limited computational power.