**Project 3) TEXT MiNING ONLINE JOB POSTING**

**Objective** –

Our main business objectives are to understand the dynamics of the labour market of Armenia using

the online job portal post as a proxy. A secondary objective is to implement advanced text analytics

as a proof of concept to create additional features such as enhanced search function that can add

additional value to the users of the job portal.

**Business Problem** –

Business questions answering to our business objectives are defined as follows:

**Job Nature and Company Profiles:** What are the types of jobs that are in demand in Armenia? How

are the job natures changing over time?

**Desired Characteristics and Skill-Sets:**

What are the desired characteristics and skill-set of the candidates based on the job description

dataset? How these are desired characteristics changing over time?

**IT Job Classification:** Build a classifier that can tell us from the job description and company

description whether a job is IT or not, so that this column can be automatically populated for new

job postings. After doing so, understand what important factors are which drives this classification.

**Similarity of Jobs:** Given a job title, find the 5 top jobs that are of a similar nature, based on the job

post.

**Text Mining Goals**:

The text mining goals is a set of sub-goals to answer our business questions:

For the IT Job classification business question, you should aim to create supervised learning

classification models that are able to classify based on the job text data accurately, is it an IT job.

On the business question of Job Nature and Company Profiles. Unsupervised learning techniques,

such as topic modelling and other techniques such as term frequency counting will be applied to the

data, including time period segmented dataset. Qualitative assessment will be done on the results to

help us understand the job postings.

To understand the desired characteristics and skill-sets demanded by employers in the job ads,

unsupervised learning methods such as K-means clustering will be used after appropriate dimension

reduction.

For Job Queries business question, we propose exploring the usage of Latent Semantic Model and

Matrix Similarity methods for information retrieval. The results will be assessed qualitatively. To

return the top 5 most similar job posting, the job text data are vectorised using different models

such as word2vec, and doc2vec and similarity scores are obtained using cosine similarity scores,

ranked and returned as the answer which is then evaluated individually for relevance.

**Data Availability** –

The dataset representation is tabular, but many of the columns are textual/unstructured in nature. Most

notably, the columns jobDescription, JobRequirement, RequiredQual, ApplicationP and AboutC are

textual. The column jobpost is an amalgamation of these various textual columns.

.

**Historical Data** –

* Last 6 months data
* The was obtained in last year kaggle competition

**Techniques –**

* Used **NLP Corpus, Stemming and Naive Bayes Classifier**, classify the agenda to their respective bucket (Package Issue, Feature Issue, Compliance Issue, PI Issue , Audit Issue .. etc) and assign them to teams accordingly.
* Made use of **python, nltk, re library** to pick up key identifiers such as process name, team ID, manager name/ID etc from plain text messages.

**Tools Used –** Python (numpy, pandas, re, nltk, sklearn, textblob, nlp etc.), Excel

**Metrices**-

* We used Confusion Matrix for multiclass classification. Confusion matrix can tell if there’s a class that’s constantly mistaken for some other class.

**Validation –**

• It will train my model using train data set I will verify result against test data set. If we will get

Test data classified to accurate class then I will say my model is good model.

**Challenges** –

* Data Collection
* Identify the Positive Drivers