KE-5205 (TEXT MINING) PROJECT REPORT

TEXT MINING ON ARMENIAN ONLINE JOB POSTINGS

GROUP TM01

ANURAG CHATTERJEE **(A0178373U)**
BHUJBAL VAIBHAV SHIVAJI**(A0178321H)**
GOH CHUNG TAT KENRICK**(A0080891Y)**
GOPALAKRISHNAN SAISUBRAMANIAM**(A0178249N)**
LIM PIER **(A0178254X)**
LIU THEODORUS DAVID LEONARDI **(A0178263X)**
TSAN YEE SOON**(A0178316Y)**

MASTER OF TECHNOLOGY IN
KNOWLEDGE ENGINEERING
BATCH KE-30(2018)

Contents

Executive Summary

Online job advertisements have become the dominant job searching and employer-employee job matching model in most developed economies around the world and gaining popularity in all parts of the world. It is estimated that in 2014 that more than 70 percent of job openings are posted online in the United States of America and by researching the detailed data of the online jobs ads, researchers can better understand the labour market.

This project analysed the online job ads posting from 2004 to 2015 posted on the CareerCenter.am, an Armenian online human resource centre. The main business objective is to understand the dynamics of the labour market of Armenia and relevant business questions were defined. A secondary objective is to implement advanced text analytics as a proof of concept to create classification and information retrieval function that can add additional value to the job portal.

We followed the CRISP-DM methodology for the project. After the business and data understanding with visualisation, we prepared the data. The Armenian job dataset was cleaned of missing data and duplicated sets. The pre-processing involves heavily the careful removal of non-essential characters such as newlines and punctuations and conversion of text to lower case. Also, depending on the task, tokenisation followed by lemmatisation is also done. Patterns that did not add much value such as emails and telephone numbers were also removed. We also expanded the stop-words dictionary for this task and other common words such as 'Armenian' which carries no additional information value.

By applying K-Means clustering, we created an understanding of the required qualifications and skillset in the Armenia labour market over the 10-year period from 2004 to 2015. IT related skills demands have shown to be on an increasing trend increased over the period. This increase and shift towards IT-related jobs are also validated by topic modelling that clearly shown that software development 'topic' has the strongest growth over the same period.

Applying supervised text mining techniques, we also demonstrated it is possible to create accurate classification models that create a filter out IT related job posting. We are also able to identify the type of companies that create job ads with our custom regex filters. In addition, we have shown that a job similarity search function is possible given the embeddings on the job ad text using cosine similarity between the vectors.

It is recommended that the Armenian government place more emphasis and effort to provide better education pathway for their citizens to gain IT skills to fulfil the IT skills labour demand. Likewise, for current and future job seekers, it is recommended that if they have interest, they should not hesitate to equip themselves with IT skillset, in particularly the IT operations and applications developments knowledge. In addition, the CareerCenter.am job portal website can utilise our project Information Extraction methods to improve the search capabilities and features of the website, enhancing the user experiences and capabilities of the job matching function.

1 Business Understanding

1.1 Introduction

The project seeks to understand the overall demand for labour in the Armenian online job market from the 19,000 job postings from 2004 to 2015 posted on CareerCenter, an Armenian human resource portal. Through text mining on this data, we will be able to understand the nature of the ever-changing job market, as well as the overall demand for labour in the Armenia economy. The data was originally scraped from a Yahoo! Mailing group.

1.2 Business Objectives

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 questions answering to our business objectives are defined as follows:

1.2.1 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?

1.2.2 Desired Characteristics and Skill-Sets

What are the desired characteristics and skill-set of the candidates based on the job description dataset? How are these desired characteristics changing over time?

1.2.3 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 are important factors which drives this classification.

1.2.4 Similarity of Jobs

Given a job title, find the 5 top jobs that are of a similar nature, based on the job post.

1.3 Text Mining Goals and Project Plan

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

For the IT Job classification business question, we aim to create supervised learning classification models that are able to classify based on the job text data accurately, is it an IT job. The evaluation will be the F1 score for the classification model.

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. Silhouette scoring and qualitative assessment will be used to determine the quality of the clustering results.

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.

2 Data Understanding

2.1 Initial Data

The data was obtained from Kaggle here. Each row represents a job post. 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.

2.2 Data Description

2.2.1 Key Data Columns Description

Col Name	Description	Col Name	Description
Jobpost	The full text for the job post	**Date**	Date that the job was posted
Title	Title of the job	**Company**	Company for the job
Announcement Code	Announcement code, which is some internal code and is usually missing.	**Term**	Announcement code, which is some internal code and is usually missing.
Eligibility	Eligibility of the candidates.	**Audience**	Who can apply?
StartDate	Start date of work.	**Duration**	Duration of the employment.
Location	Employment location.	**JobDescription**	Job Description.
JobRequirement	Job requirements.	**RequiredQual**	Required qualifications.
Salary	Job salary.	**ApplicationP**	Application procedure.
OpeningDate	Opening date of the job announcement.	**Deadline**	Deadline for the job announcement.
Notes	Additional notes.	**AboutC**	About the company.
Attach	Attachments.	**Year**	Year of the announcement (derived from the field date).
Month	The month of the announcement (derived from the field date).	**IT**	TRUE if the job is an IT job.

Table 1: Key Data Columns Description

2.3 Data Exploration

2.3.1 Job Postings by Year

We begin by doing performing non-text mining with data exploration, doing a simple plot of the job postings by the year. We see that the number of job postings has increased throughout the years, with a dip in the year 2009, likely due to the 2008 global financial crisis. The number of job postings increasing is either indicative of the economy of Armenia expanding, or there is a shift towards online job postings as opposed to traditional media job postings in general or a function of both.

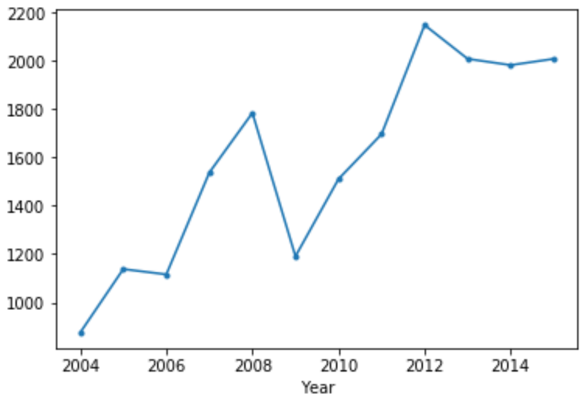


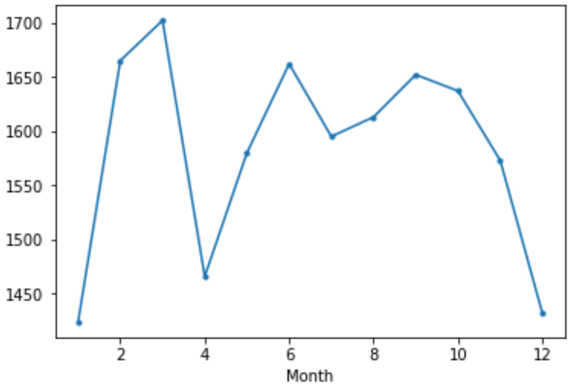
Figure 1 No. of Job Ads Posts over different periods

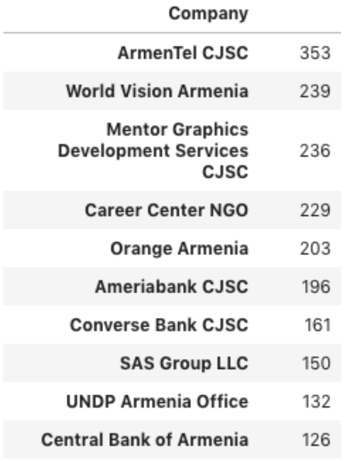
2.3.2 Job Postings by Month

Next, we did a plot of all the job postings grouped at the month level. As can be seen in the plot, job postings peak in the months of March and June and are at a low in December and January.

2.3.3 Top Companies Posting Jobs

Next, a summation of the number of jobs posted, grouped by the company was done. Here is a list of the top 20 companies posting in this dataset. It can be roughly seen that most jobs are from banking or technology in this dataset. More in-depth analysis of the different job sectors will follow in the text mining portion of this report.





Company	
ArmenTel CJSC	353
World Vision Armenia	239
Mentor Graphics Development Services CJSC	236
Career Center NGO	229
Orange Armenia	203
Ameriabank CJSC	196
Converse Bank CJSC	161
SAS Group LLC	150
UNDP Armenia Office	132
Central Bank of Armenia	126

Figure 2 No. of Job Postings by Month (left) and Top Companies Posting Jobs (Right)

2.3.4 Length of Job Post for Each Entry

The plot below shows the length of the JobPost column for each entry. Most posts are under 2000 wor ds, with one particular exception, JobPost with index 105, which has the job title ‘International Prize fo r R&D in Biomedicine and New Technologies’.

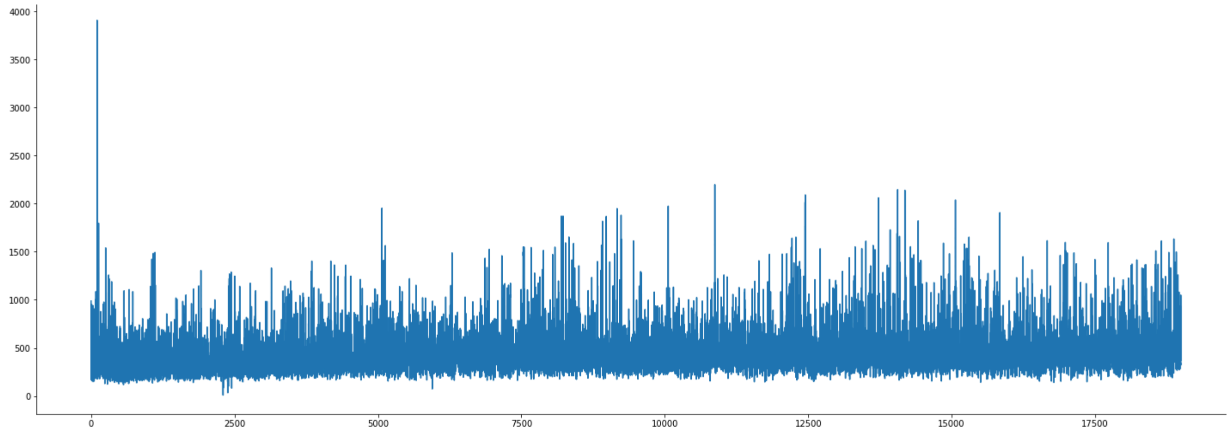


Figure 3 Length of Job Posting on the Job Portal

2.4 Data Quality

From our examination of the data, we see the following issues in the dataset:

- **Missing values / NAs values**
- **Duplicate posts for the same job**

We have dealt with this in the pre-processing stage, which will be described in the following section. In addition, we have made the assumption that these job postings have been proof-read by the Human Resources Department of the respective companies. Hence there are minimal typos and a quick check with pyenchant package – English language dictionary also confirms this is indeed the case.

3 Data Preparation

3.1 Data Selection and Construction

We selected the following columns in the table for analysis:

- Jobpost
- Title
- JobRequirement
- JobDescription
- RequiredQual
- AboutC
- IT
- Year

The full description of these fields are detailed in Table 1.

3.2 Data Cleaning

Removal of duplicate job posts was done as these are likely due to the same poster posting multiple times when the job is not filled. After doing so, it was found that from 19001 rows, there were 18892 rows left, meaning that 109 duplicate rows were removed. The condition for detecting duplicates was based on considering both the jobpost and the title column.
For the columns that are labelled NAs, we simply did not consider them when doing the text mining.

3.3 Data Pre-processing

Note that these data pre-processing tasks were performed for a majority of the tasks. In the cases where the placement of words in a phrase is important, we did not do some of the tasks. An example is Word2Vec, which considers the words before and after the target word. In such a case, we did not do lemmatization.

3.3.1 Removal of Newlines and Single Quotes

Unnecessary characters like newline '\n' and single quotes were removed as these did not contribute to the text mining goals.

3.3.2 Removal of Emails, Web addresses and Telephone numbers

Emails, web addresses and telephone numbers did not align with most of the text mining tasks that we performed and were removed for the most part of the tasks.

3.3.3 Removal of Punctuations and Conversion to Lower-case

To not recognise two words that are the same but do not have the same capitalisation as two different tokens, all words were set to lower-case for most tasks. For tasks where part of speech tagging was required to be performed this step was not performed.

3.3.4 Tokenisation

Tokenisation was done prior to lemmatisation and stop-word removal using NLTK's word_tokenize function.

3.3.5 Lemmatisation

Using lemmatisation, we modified tokens with similar meaning and changed them all to the lemma, the base or dictionary form of the word. NLTK's WordNet lemmatiser was used in most cases.

3.3.6 Extension of Stop-words

When processing the job post column, we extended the stop words to include words like 'armenian', 'armenia', 'job', 'title', 'position', 'location', 'responsibilities', 'application', 'procedures', 'deadline', 'required', 'qualifications', 'renumeration', 'salary', 'date', 'company' and 'llc'.

These were mostly headings of sections in the jobpost column and did not serve much purpose in our text mining tasks. For 'armenia', it was removed as we already know that the jobs are based in Armenia based on the context of the dataset. For some of the text mining tasks, we also removed calendar months to further refine the clustering/topic modelling outcomes.

3.3.7 Bigrams

We were able to get meaningful bigrams like "financial management" from using Gensim's Phrases and Phraser classes. This helped us to derive more meaning from the job posts.

3.3.8 Test Pre-processing on a Software Developer Job Post

After the above tasks were done, we tested the processed data on a single row, which had the title 'Software Developer'.

The following word cloud also shows the nature of the tokens in this row in a friendlier way. We can see that it is commensurate with the title ‘Software Developer’, and we know that this particular job needs SQL and server expertise and is database-centric. It is also likely to be based in Yerevan.

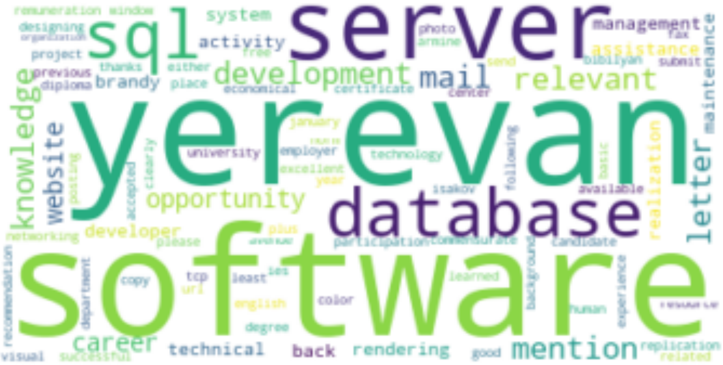


Figure 4 Word Cloud generated from the Job Ads Postings

4 Text Mining

4.1 Supervised Learning

This section seeks to classify if a job posting is related to IT sector or not which is the business objective 1.2.3.

4.1.1 TF-IDF Based Classification

TD-IDF, also known as Term Frequency-Inverse Document Frequency is a statistical method to reflect the importance of a word to a document in a collection.

In order to correctly classify whether a particular job posting is an IT job or not, we used two columns from the dataset, “jobpost” as the text corpus and “IT” as the label (IT job or not).

We have reused the Pre-Processing methods in the Pre-processing section. As performed previously, we have generated the TF-IDF and reduced the dimension through Singular Value Decomposition (SVD) to 1500 columns.

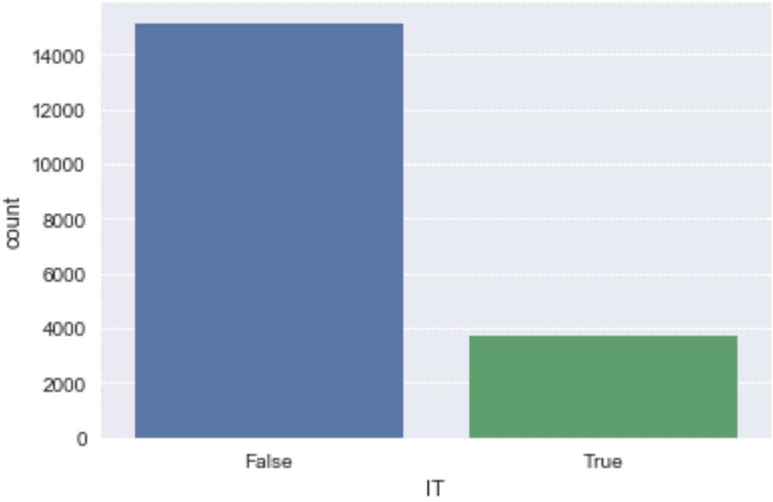


Figure 5 Dataset IT Job Class Distribution

After which, Support Vector Machine (SVM) is used to perform classification on whether the job is an IT job or not. Initial Inspection showed that there is an imbalanced class as shown in the figure below. As such, we will change the increase the weightage for those job which are IT jobs.

After splitting the available into 70% training and 30% testing, we train our SVM model and we predict on the test partition to obtain the results as shown in the confusion matrix.

This model performs fairly well as F1 score is high on both classes. To improve on the accuracy, we used more advanced techniques as shown in the next section.

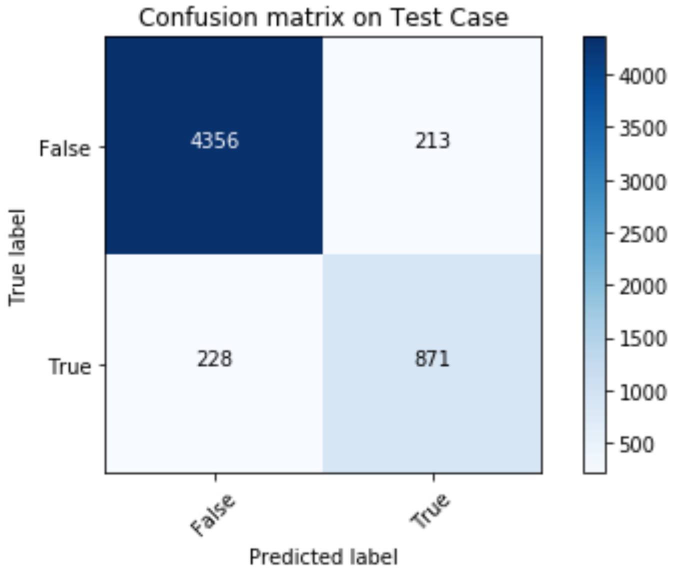


Figure 6 Confusion Matrix Result

	Precision	Recall	F1-score	Support
False	0.9503	0.9534	0.9518	4569
True	0.8035	0.7925	0.7980	1099
Binary Average	0.8035	0.7925	0.798	5668
Micro Average	0.9222	0.9222	0.9222	5668
Macro Average	0.8769	0.873	0.8749	5668
Weighted Average	0.9218	0.9222	0.922	5668

Figure 7 Evaluation Scores using Support Vectors Classifier

4.1.2 RNN-Based Classification

Recurrent Neural Networks (RNN) are powerful neural networks that works really well with sequential data like time series, speech, text, financial data, audio, video, weather, and much more.

In order to correctly classify whether a particular job posting is an IT job or not, we used two columns from the dataset, “jobpost” as the text corpus and “IT” as the label (IT job or not). The text corpus were then tokenised using Spacy (https://www.spacy.io), and no stop words nor punctuation were removed. We use GloVe, Global Vectors for Word Representation from Stanford NLP, which were pre-trained on Wikipedia 2014+ Gigaword 5 (6 billion tokens, 400k vocabularies, 100 dimension vectors) as the pre-trained word embeddings model to initialise default embedding layer’s weight vectors.

The following is the Recurrent Neural Network architecture:

Layer	Dimension
Embedding	vocab_size=35896, embedding_dim=100
Dropout	dropout_rate=0.5

LSTM	embedding_dim=100, hidden_dim=256, num_layers=2, bidirectional=True, dropout_rate=0.5
Dropout	dropout_rate=0.5
Linear	Input_features=512, output_features=1

	Function
Optimizer	Adam (learning rate=0.001)
Loss Function	Binary Cross Entropy with Logits

Table 2 RNN Architecture and Parameters

The job posting dataset were split into three parts (train: 9,311 posts, validation: 3,990 posts, and test: 5,701 posts) for our classification task. The model was trained in 5 epochs, in which the progress of its training can be seen in the images below.

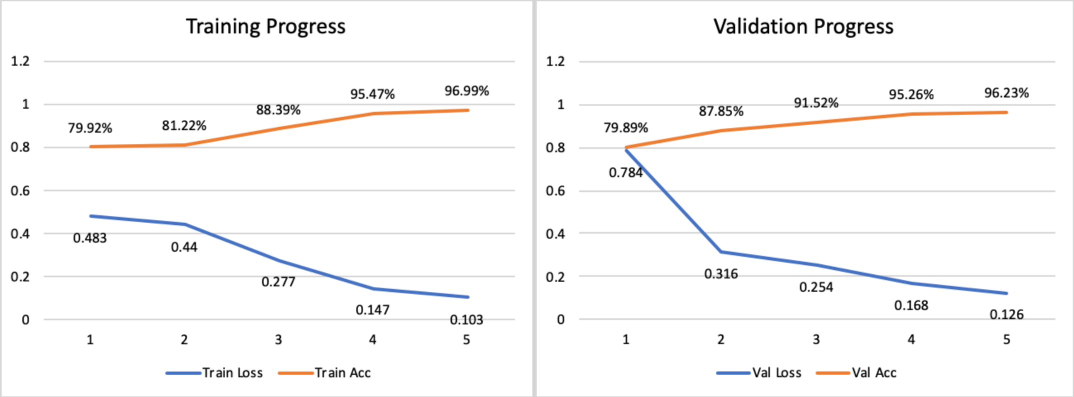
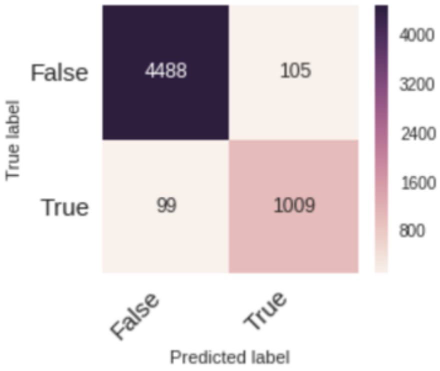


Figure 8 Training and Validation Progression Charts

The model could achieve 96.42% accuracy on test dataset. These are the confusion matrix and others scores that were calculated based on the comparison between predicted versus actual target values (IT job or not) of test dataset.



IT Job or Not	F1-score	Precision	Recall	Support
False	0.9778	0.9784	0.9771	4,593
True	0.9082	0.9057	0.9106	1,108
Binary avg.	0.9082	0.9057	0.9106	5,701
Micro avg.	0.9642	0.9642	0.9642	5,701
Macro avg.	0.9430	0.9421	0.9439	5,701
Weighted avg.	0.9643	0.9643	0.9642	5,701

Table 3 Confusion Matrix and the Evaluation Matrix Table

4.2 Unsupervised Learning

This section seeks to explore the job nature, company profiles and the skills that are demanded by the companies which align the business questions 1.2.3 and 1.2.4. We leveraged unsupervised learning methods including clustering, topic modelling and named entity extraction for the analysis.

4.2.1 K-Means Clustering

To have a global understanding of the required qualifications for the job postings and to understand how these qualifications have changed over the time-period evaluated the records were clustered based on the content of the Required qualification column.

The data from this column was tokenised and pre-processed using most of the steps mentioned in the earlier section. The TF-IDF matrix was generated for all the job postings and the terms that remained after the pre-processing with the condition that terms which occur in more than 70% of the documents or in less than three documents be ignored. The dimension of the resulting TF-IDF matrix was (16689, 5591) where 16,689 is the number of job postings and 5591 is the number of terms in the TF-IDF matrix. To efficiently perform the clustering, the dimensions of the job postings (document) vector was reduced from 5591 to 1500 using the Singular Value Decomposition technique.

After performing SVD, 89% of the variance of the data could still be explained by the 1500 components. The job posts which were now represented by the SVD values were clustered using KMeans clustering. Cluster sizes between 4 to 10 were tried. Both the interpretation of the clusters and the Silhouette score were considered to evaluate the optimal number of clusters. Based on the business objective of understanding the global demand of qualifications over all the job postings, interpretability of the clusters was considered the more important criterion to evaluate the number of clusters.

We chose KMeans with K=7. Although the Silhouette score of 0.013 is lower but we chose K=7 since it gave the clusters with distinct interpretations. The centroid of the clusters revealed the below terms and the clusters were interpreted based on the terms as described in the below table.

Cluster number	Terms	Interpretation
1	net sql web development good server javascript database html php	IT application development
2	ability management work degree year good excellent project field	Project management
3	higher education work russian excellent good computer field ability	Higher education and language
4	accounting finance tax work financial legislation good excellent standard	Accounting and finance
5	marketing sale business excellent ability work russian communication strong degree	Marketing and sales
6	development testing software design good programming ability plus tool linux	IT operations
7	work excellent ability russian good university communication degree strong	Communication

Table 4 Interpretation of required qualification clusters over every job postings

To understand the how the required qualifications have changed over the years, clustering was performed in 3 distinct sets of filtered data based on the year of the job posting and the cluster for that skill is tabulated against the proportion of the number of job postings for those years demanding the skills in that cluster as shown in the below figure.

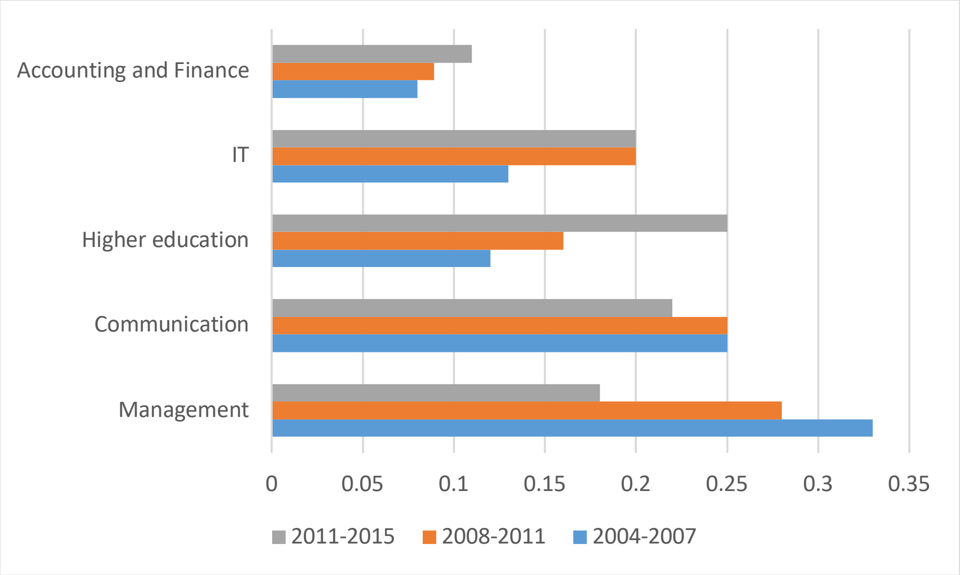


Figure 9 Skills in demand over the years

As can be seen, executive skills like management, communication in different languages and higher education are popular over the time period evaluated making up approx. 70 % of the job postings. Job postings that demand core IT skills can also be seen to increase from 13% in the 2004-07 period to 20% during 2008-11 and 2011-15 periods.

4.2.2 Named Entity Recognition

The Spacy library in Python was leveraged to perform Named Entity Recognition on the required qualifications column to understand highly sought-after skills over the entire period of job postings. Among the different entity types extracted by Spacy, we focussed on the "ORG" entity types as they contain the names of the skills, software tools, etc. identified from within the corpus. The results were visualised using a word cloud, and as can be seen below, Microsoft Office and Computer Science skills are highly sought after.

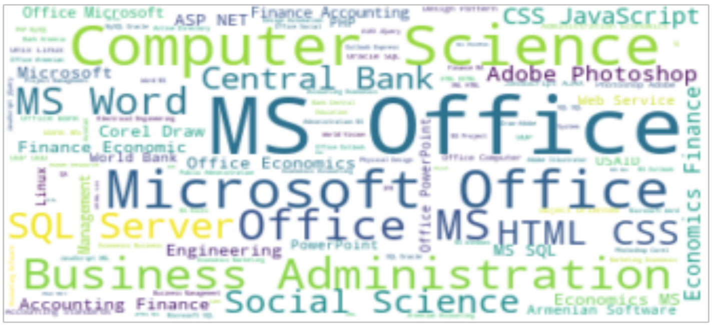


Figure 10 Highly sought-after qualifications based on all job postings

4.2.3 Topic Modelling to Determine Nature of Jobs over Time

We employed the use of genism's LdaModel class to do Latent Dirichlet Allocation (LDA) for topic modelling on the jobpost column. After running the topic modellings with various parameters, we decided that the following gave the best results. After trying number of topics from 6 to 10, we felt that the best interpretability was given by the following parameters.

- Number of Topics : 7
- Filtering Dictionary by Extremes – no_below = 3
- Filtering Dictionary by Extremes – no_above = 0.7

We also found that we got better results by doing POS tagging and restricting the words to nouns and verbs. This gave us the following results for topic modelling:

```
[(0,
  '0.024*"ability" + 0.021*"communication" + 0.019*"customer" + 0.018*"term" + 0.015*"sale" + 0.015*"line" + 0.015*"llc" + 0.014*"ma
rketing" + 0.013*"service" + 0.012*"send"'),
 (1,
  '0.066*"project" + 0.036*"development" + 0.022*"program" + 0.019*"activity" + 0.018*"support" + 0.018*"implementation" + 0.013*
"sector" + 0.013*"include" + 0.012*"community" + 0.011*"ensure"'),
 (2,
  '0.024*"report" + 0.016*"bank" + 0.015*"datum" + 0.014*"prepare" + 0.012*"indicate" + 0.011*"year" + 0.011*"accounting" + 0.010*"
finance" + 0.010*"branch" + 0.010*"account"'),
 (3,
  '0.039*"development" + 0.037*"software" + 0.033*"design" + 0.025*"team" + 0.022*"developer" + 0.018*"web" + 0.017*"develop" +
0.016*"test" + 0.015*"system" + 0.015*"technology"'),
 (4,
  '0.034*"engineering" + 0.022*"construction" + 0.018*"safety" + 0.017*"test" + 0.014*"engineer" + 0.014*"store" + 0.013*"site" + 0.01
2*"register" + 0.011*"amd" + 0.011*"answer"'),
 (5,
  '0.018*"office" + 0.017*"course" + 0.015*"training" + 0.012*"expert" + 0.011*"study" + 0.010*"student" + 0.010*"material" + 0.010*"e
ducation" + 0.009*"medium" + 0.009*"month"'),
 (6,
  '0.035*"management" + 0.025*"ensure" + 0.023*"business" + 0.020*"manage" + 0.019*"plan" + 0.017*"develop" + 0.016*"manager"
+ 0.013*"system" + 0.013*"service" + 0.012*"process"')]
```

We then derived the following topics from the results. Note that the topics derived are the nature of the jobs:

- Topic 0: Sales and Marketing
- Topic 1: Project Management and Development
- Topic 2: Banking and Finance
- Topic 3: Software Development
- Topic 4: Construction and Safety Engineering
- Topic 5: Education and Training
- Topic 6: Business Development and Management

The following plot shows the topic frequencies, based on the top topic that each job post belongs to. We can see that most jobs have the topic Sales and Marketing. The Construction and Safety Engineering topic has the least number of jobs belonging to it from the dataset.

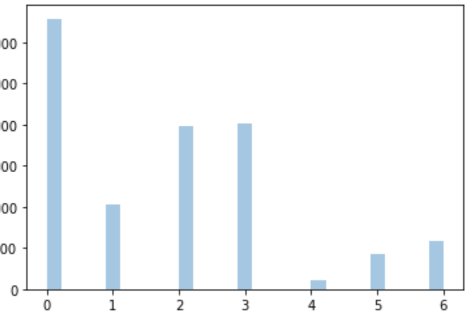
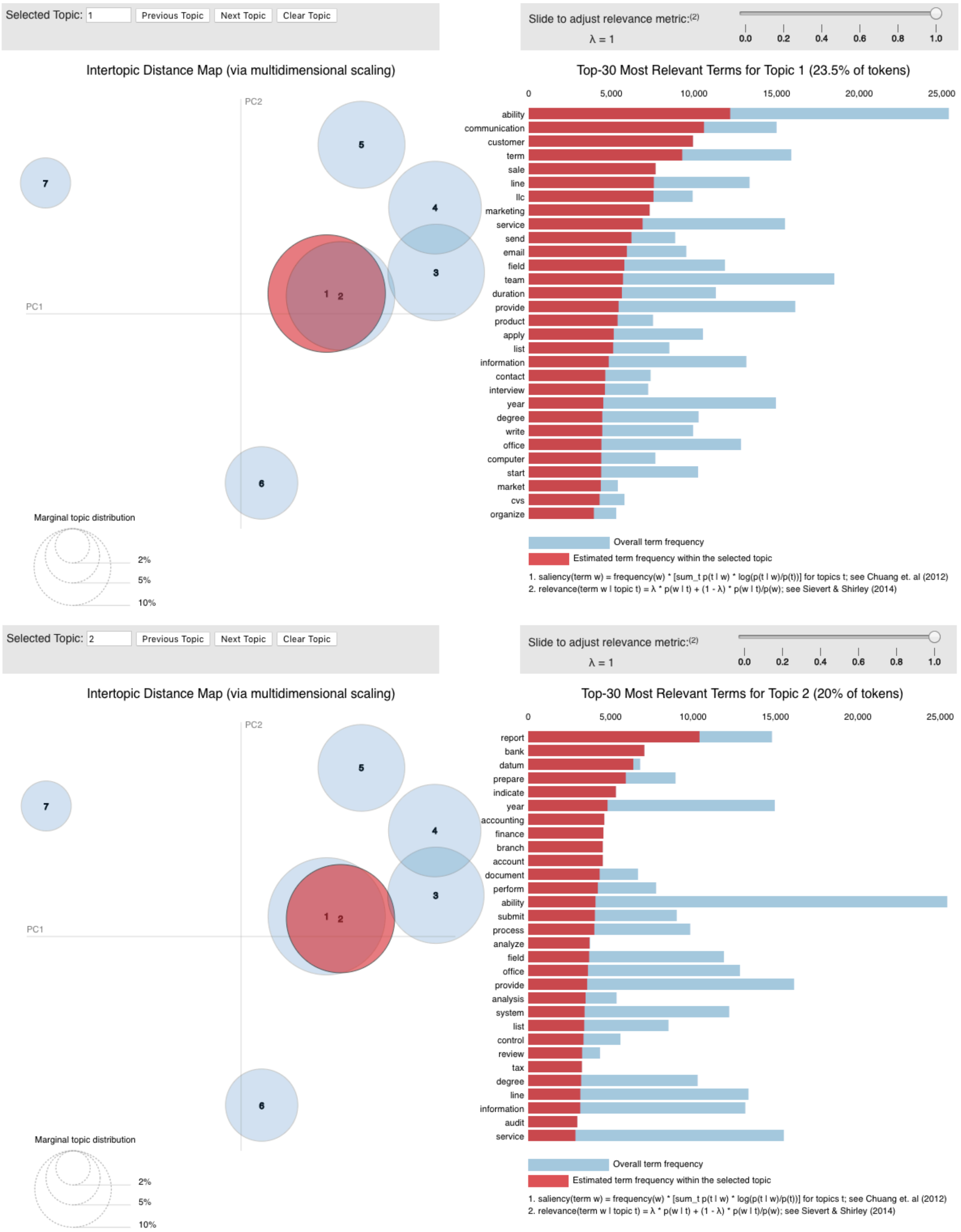


Figure 11 Topic Frequencies



Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Selected Topic: 2

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 12 Most Relevant Terms in Topic 1 (top) and Topic 2 (bottom)

We then use pyLDAvis to further visualise the results. Due to the nature of pyLDAvis, we did not have the same topic numbering. pyLDAvis created a inter-topic distance map based on multidimensional scaling.

Note that Topic 1 (Sales and Marketing) overlaps with Topic 2 (Banking and Finance). This suggests that these two topics are closely related, and that many of these banking and finance jobs most likely have a sales/marketing nature. The other topics are mostly standalone, except for a slight overlap between

Topic 3 (Project Management and Development) and Topic 4 (Business Development and Management), which makes sense due to the business orientated nature of both topics.

4.2.4 Observing the Changes in Topics over the Years

We also plotted the changes in the distribution of topic frequencies over the years in three demarcations period, 2004-2007, 2008-2011 and 2012-2015.

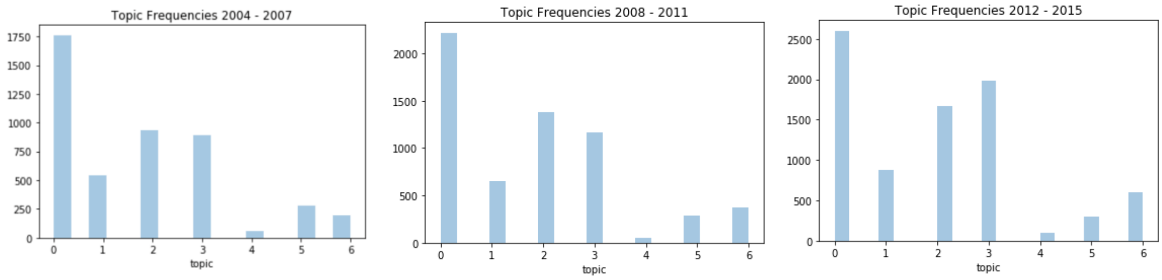


Figure 13 Charts showing Topic Freq. Count over the three time periods

- Topic 0: Sales and Marketing
- Topic 1: Project Management and Development
- Topic 2: Banking and Finance
- Topic 3: Software Development
- Topic 4: Construction and Safety Engineering
- Topic 5: Education and Training
- Topic 6: Business Development and Management

As can be observed, proportion-wise, the topic frequencies remain roughly similar. What is obvious is that the two topics Banking and Finance as well as Software Development have grown in proportion across the overall job market over the years. Also, the highest number of jobs is for a Sales and Marketing position. Construction and Safety Engineering kinds of job consistently had low numbers of job postings over the years.

There were slightly less Software Development jobs compared to Banking and Finance jobs initially from 2004-2007. This trend continued from 2008 to 2011. However, from 2012-2015, Software Development jobs were higher in demand than Banking and Finance jobs. A large proportion of the job posts are for jobs located in Yerevan, and we already know that Yerevan is Armenia’s finance hub. A possible explanation could be that banks and finance services have become more digitalised, and many jobs for the Banking and Finance industry require Software Development skills as well.

4.3 Information Extraction

This section starts by further evaluating the company profiles using pattern matching techniques and then evaluate the similarity between jobs by finding the top 5 most similar jobs given a job title (to align with objective 1.2.4).

4.3.1 Information Extraction to Obtain Company Profiles

We also did manual rule matching on the AboutC column, which has a description of the companies, to extract more precise information about the job natures.
The patterns used are (in Spacy Matcher format) :

- [{'POS':'ADJ'},{'ORTH': '-'}, {'POS':'NOUN', 'OP':'+'}, {'LOWER': 'organization'}]
- [{'POS':'NOUN', 'OP':'+'}, {'LOWER': 'organization'}]
- [{'POS':'NOUN', 'OP':'+'}, {'LOWER': 'company'}]
- [{'POS':'NOUN', 'OP':'+'}, {'LOWER': 'services'}]
- [{'LOWER': 'company'}, {'POS':'VERB'}, {'POS':'NOUN', 'OP':'+'}]
- [{'LOWER':'provider'}, {'POS':'ADP'}, {'POS':'NOUN', 'OP': '+'} , {'POS':'CCONJ','OP':'+'}, {'POS':'NOUN','OP':'+'}]

- [{'LOWER':'provider'}, {'POS':'ADP'}, {'POS':'NOUN', 'OP': '+'}]
- [{'POS':'VERB'}, {'POS':'NOUN', 'OP': '+'}]
- [{'POS':'ADJ'}, {'POS':'NOUN', 'OP': '+'}]
- [{'POS':'NOUN'}, {'POS':'ADP'}, {'POS':'NOUN', 'OP': '+'}]
- [{'POS':'NOUN'}, {'POS':'NOUN','OP':'+'}]

This allowed us to capture the following company profiles, visualised here in a simple frequency distribution. This information can be used to supplement what we found from the topic modelling to give greater insight into the companies available in Armenia. As expected, many of the posts are for finance and software companies.

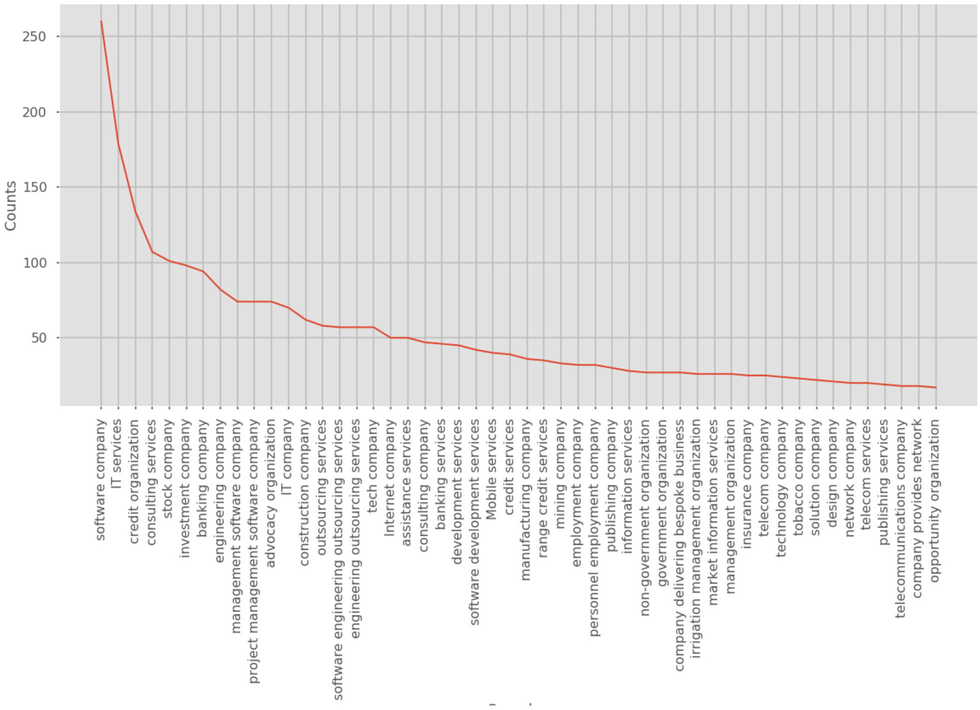


Figure 14 Company Profiles/Types that created the Job Ads

4.3.2 Job Similarity

To identify the similarity between jobs, the four columns – Title, JobDescription, RequiredQual and JobRequirement were selected as they provide different views of describing jobs. The pre-processing steps mentioned in section 3.3 were applied and stored as separate columns, to study the influence of raw text vs the processed.

To compute job similarity for a field, the chosen measure was Cosine Similarity. The normalised pairwise cosine distance gave the dissimilarity score between jobs and subtracting 1 from a distance gave the similarity score. To compute the overall job similarity, the individual four similarities based on the fields were averaged, which was used to rank jobs by their closeness to each other.

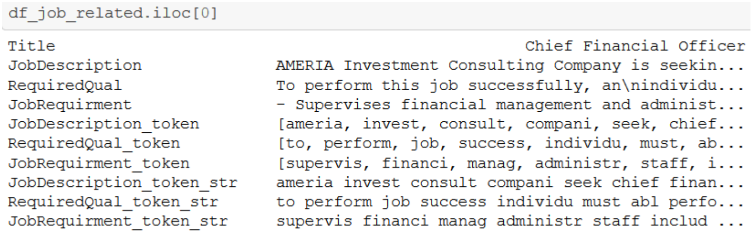
The subsection below discusses three approaches which serve as feature engineering steps that aid in computing similarity.

4.3.2.1 Using TF-IDF

The TF-IDF matrix was generated individually based on the pre-processed tokens of Title, Job Description, Required Qualification and Job Requirement columns. For a pair of job i and j, the cosine distance between the two vectors corresponding to i and j indices of the TF-IDF matrix gave the dissimilarity score between them. Simple average across all the scores was taken to be the overall score.

Below are sample results from the TF-IDF approach:

1. Chief Financial Officer Jobs



```
df_job_related.iloc[0]

Title                                          Chief Financial Officer
JobDescription             AMERIA Investment Consulting Company is seekin...
RequiredQual               To perform this job successfully, an\nindividu...
JobRequirment              - Supervises financial management and administ...
JobDescription_token       [ameria, invest, consult, compani, seek, chief...
RequiredQual_token         [to, perform, job, success, individu, must, ab...
JobRequirment_token        [supervis, financi, manag, administr, staff, i...
JobDescription_token_str   ameria invest consult compani seek chief finan...
RequiredQual_token_str     to perform job success individu must abl perfo...
JobRequirment_token_str    supervis financi manag administr staff includ ...
```

Figure 15 CFO Job Posting Results using TF-IDF model

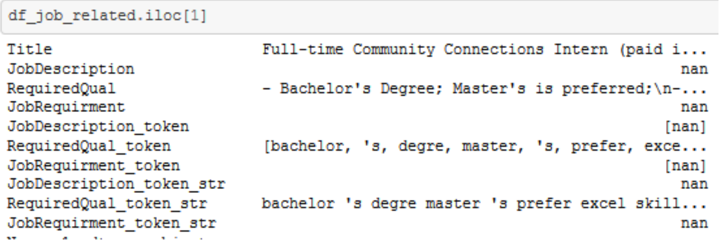
S

Similar jobs with scores:

Title	JobDescription	RequiredQual	JobRequirment	Similarity Score
Chief Financial Officer	River Island is seeking a Chief Financial Officer..	Master's degree in Management, Finance , Economics	Develop tools and systems to provide critical..	0.4898
Chief Financial Officer	Armenian Datacom Company (ADC) is seeking ..	Master's degree in Accounting, Finance	Responsibilities include but are not limited	0.4888
Chief Financial Officer	The Chief Financial Officer (CFO) will support..	- Degree in Accounting, Finance, Business, Law...	Financial management of the NSRCIP T1 & T2 pro...	0.4808
Chief Financial Officer	Gritti LLC is inviting highly qualified profe...	- Master's or equivalent university degree in ...	General Responsibilities: - Assist in perf...	0.4737
Chief Financial Officer (CFO)	The CFO will have full authority and responsib...	- Master's degree in Business Administration, .	- Ensure compliance and deal with local, state..	0.4647

Table 5 Top 5 CFO similar job postings

2. Intern Jobs



```
df_job_related.iloc[1]

Title                      Full-time Community Connections Intern (paid i...
JobDescription                                                           nan
RequiredQual               - Bachelor's Degree; Master's is preferred;\n-...
JobRequirment                                                            nan
JobDescription_token                                                   [nan]
RequiredQual_token         [bachelor, 's, degre, master, 's, prefer, exce...
JobRequirment_token                                                    [nan]
JobDescription_token_str                                                 nan
RequiredQual_token_str     bachelor 's degre master 's prefer excel skill...
JobRequirment_token_str                                                  nan
```

Figure 16 Intern Job Postings results using TF-IDF model

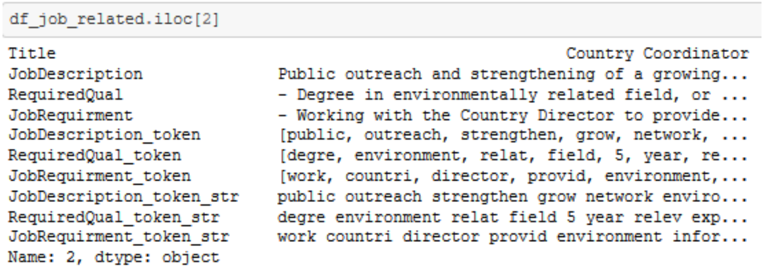
Similar jobs with scores:

Title	JobDescription	RequiredQual	JobRequirment	Similarity Score
IT Teacher (full time)	- Degree in Computer Science, Information Tech...	Nan	Nan	0.6631
Non-paid part or full time Administrative Intern	Nan	Nan	Nan	0.6341
Administrative and Programmatic Intern	Nan	- Fluency in English and Armenian;\n- Good com...	Nan	0.6269
Administrative and Programmatic Intern	Nan	- Fluency in English and Armenian;- Good com...	Nan	0.62697
Non-paid part or full time	Nan	Nan	Nan	0.62597

Programmatic Intern				

Table 6 Top 5 Intern similar job postings

3. Country Coordinator



```
df_job_related.iloc[2]

Title                                                  Country Coordinator
JobDescription            Public outreach and strengthening of a growing...
RequiredQual              - Degree in environmentally related field, or ...
JobRequirment             - Working with the Country Director to provide...
JobDescription_token      [public, outreach, strengthen, grow, network, ...
RequiredQual_token        [degre, environment, relat, field, 5, year, re...
JobRequirment_token       [work, countri, director, provid, environment,...
JobDescription_token_str  public outreach strengthen grow network enviro...
RequiredQual_token_str    degre environment relat field 5 year relev exp...
JobRequirment_token_str   work countri director provid environment infor...
Name: 2, dtype: object
```

Figure 17 Results of Country Coordinator using TF-IDF model

Similar jobs with scores:

Title	JobDescription	RequiredQual	JobRequirment	Similarity Score
Country Coordinator - Armenia	CENN - Caucasus Environmental NGO Network - is..	1. Education: University education in environm.	- Write first hand articles, conduct intervie.	0.4585
Country Coordinator	Veya Limited needs an experienced Business Ma..	- University degree in Middle East Studies, In..	- Coordinate activities of the office and staf...	0.2960
Environmental Coordinator	The Environmental Coordinator performs a cros...	- Higher education in Economics, Technical fie...	- Organize Environmental Committee meetings, c...	0.2795
Country Coordinator	Veya Ltd. is looking for a qualified Country...	- At least five years of experience in project ma..	- Negotiate with potential customers;- Lobby...	0.2727

Table 7 Top 5 similar jobs to Country Coordinator

It is observed that though the scores are not high, owing to the presence of common words and roles for a job, TF-IDF gives good results in identifying similar jobs through quality assessment.

4.3.2.2 Using Pretrained Word2Vec Embedding

To see if the scores can be improved through the capture of semantics, we used the functions and model weights present in the SpaCy package. 'en_core_web_lg' is a word embedding CNN model pre-trained on OntoNotes, with GloVe vectors trained on Common Crawl. It provides Vocabulary, syntax, entities, vectors and can be used to infer a 300-dimensional vector given the word. For documents or sentences, the corresponding vector is computed by the average of their individual word vectors.

Like before, individual similarity scores (cosine) were computed for each field and averaged over to get the overall similarity score between two jobs i and j.

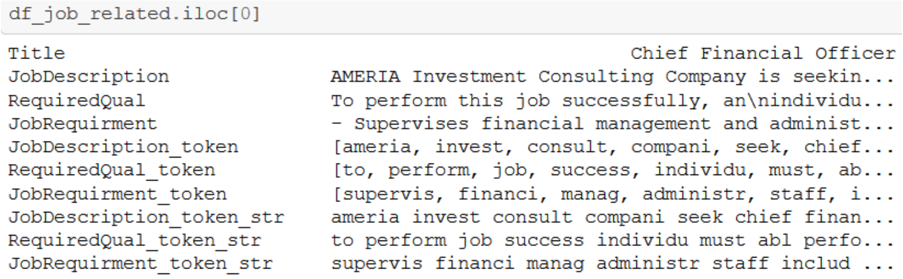
The out-of-the-box pre trained word embedding model gave decent results for the Title field. However, for the other fields, the model was not able to distinguish well, regardless of the full raw string or using the pre-processed tokens after lower case and punctuation and stopwords removal. Another approach that was tried was to make use of the keywords from POS + RegexMatcher as described in 4.3.3 and minimise the words. Results were slightly better than the former although could be improved.

For most of the job comparisons, it gave high similarity score (>0.8) The score number shows improvement over the TF-IDF, however, from a human's perspective the results do not exactly match. There is two possible explanation:

1. The pre-trained model might not have seen job-related documents
The roles and qualifications of two jobs may be semantically similar excluding the domain or field.

2. Below is a sample result from the Word2Vec approach.

Chief Financial Officer



```
df_job_related.iloc[0]

Title                                              Chief Financial Officer
JobDescription             AMERIA Investment Consulting Company is seekin...
RequiredQual               To perform this job successfully, an\nindividu...
JobRequirment              - Supervises financial management and administ...
JobDescription_token       [ameria, invest, consult, compani, seek, chief...
RequiredQual_token         [to, perform, job, success, individu, must, ab...
JobRequirment_token        [supervis, financi, manag, administr, staff, i...
JobDescription_token_str   ameria invest consult compani seek chief finan...
RequiredQual_token_str     to perform job success individu must abl perfo...
JobRequirment_token_str    supervis financi manag administr staff includ ...
```

Figure 18 CFO results using Word2Vec approach

Similar jobs with scores:

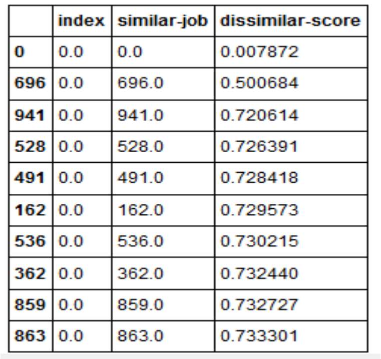
Title	JobDescription	RequiredQual	JobRequirment	Similarity Score
Chief Accountant	Chief Accountant will be supervised by Financial Director of the Company…	University degree in Accounting/ Finance; - Russian and Armenian language proficiency, knowledge of English language…	- Prepare monthly and annual financial and tax reports; - Tax management of projects; - Wire transfer maintenance…	0.8904
Chief Accountant	Rasco-Armenia cjsc is looking to recruit a highly professional Chief Accountant for a newly established insurance company. This position will carry out routine accounting and financial reporting of the company…	University degree in Economics / Finance / Accounting; - professional qualification certificate issued by the Ministry of Finance of the Republic of Armenia…	nan	0.8820
Chief Engineer	The Chief Engineer will be responsible for all engineering and technical outputs in a five-year public works program (PWP) financed by USAID…	- Five years' experience in engineering in development programs in Armenia; - Strong understanding of designs (assessment of design will be part of interview process…	- Management of technical staff; - Responsible for ensuring that all eight projects are completed with quality and on time…	0.8731
Director of Finance and Accounting	Excellent knowledge of Accounting/Tax filing both Central Bank and Tax Dept., budget formation, presentation and control.	at least 3-4 years' experience in audit/bank/lending	nan	0.8645
Chief of Party	Development Alternatives Inc. is seeking long-term Chief of Party candidates in the field of public policy/ policy advocacy to provide overall strategic direction and policy guidance	- Three to five years of COP experience preferably in Latin America; - Prior experience working in Peru…	- Lead DAI consultants in carrying out their work. - Promote policy changes in Peru by exercising public advocacy expertise.	0.858

Table 8 Top 5 similar jobs to CFO using Word2Vec

4.3.2.3 Using Custom Doc2Vec Embedding

The third approach was to train a Doc2Vec model to capture the relationships latent in paragraphs. We used the Gensim package for the purpose. The entire job post of all jobs was given as the train corpus to train the Doc2Vec model, with output embedding dimension of 300, number of epochs=200 and minimum word occurrence size of 2.

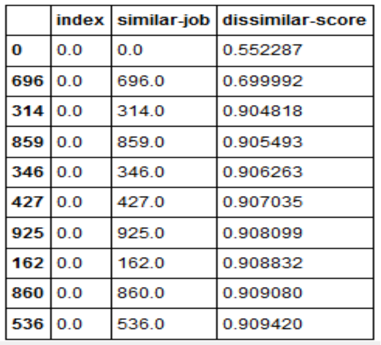
Using the raw text, the sorted dis-similarity scores showing the top 10 jobs after simple averaging is as follows:



	index	similar-job	dissimilar-score
0	0.0	0.0	0.007872
696	0.0	696.0	0.500684
941	0.0	941.0	0.720614
528	0.0	528.0	0.726391
491	0.0	491.0	0.728418
162	0.0	162.0	0.729573
536	0.0	536.0	0.730215
362	0.0	362.0	0.732440
859	0.0	859.0	0.732727
863	0.0	863.0	0.733301

Table 9 Similarity and Dissimilar scores using the simple averaging approach

Similarity was initially calculated by simple average across the fields. Weighted averaging across the fields in the ratio of: title_similarity*0.9 + jd_similarity*0.3 + jr_similarity*0.3 + rq_similarity*0.3 changed the order to an extent (weights assigned based on intuition). Since the title predominantly gives a hint to the job, it is assigned a high weight, whereas the other fields are given lower but equal weightages.



	index	similar-job	dissimilar-score
0	0.0	0.0	0.552287
696	0.0	696.0	0.699992
314	0.0	314.0	0.904818
859	0.0	859.0	0.905493
346	0.0	346.0	0.906263
427	0.0	427.0	0.907035
925	0.0	925.0	0.908099
162	0.0	162.0	0.908832
860	0.0	860.0	0.909080
536	0.0	536.0	0.909420

Table 10 Similarity and Dissimilar scores using custom weighting approach

Below are results of samples based on the Doc2Vec model. Though scores are lower as like in TF-IDF, it differs from the former in a way we can see differences in the job titles and roles but can relate to their similarities semantically.

Below are results of samples based on the Doc2Vec model. Though scores are lower as like in TF-IDF, it differs from the former in a way we can see differences in the job titles and roles but can relate to their similarities semantically.

Job	Similar Job	Matching Characteristics	Similarity Score
Chief Financial Officer	**Chief Financial Officer**	financial management, investment management, management, supervisory, and administrative skills	**0.3000**
Chief Financial Officer	**Chief Accountant**	Supervise financial aspects	**0.0945**
Programmer	**Moderator of Electronic Bulletin and Web site**	Excellent computer and editing skills.	**0.0800**

Programmer	**ASP.NET (C#) Web Developer**	Minimum two years experience in WEB programming; Knowledge of Visual Studio .NET.	**0.4530**
Full-time Community Connections Intern	**Administrative and Programmatic Intern**	Fluency in English and Armenian Good communication skills	**0.2310**

Table 11 Top 5 similar jobs to Chief Financial Officer job using Doc2Vec Model

Among the three approaches, the TF-IDF approach gives the best results from the perspective of the viewer who is interested in seeing similar jobs having the same position and from the same domain, whereas the other two embedding approaches have the tendency to show jobs from related but different positions and domains.

5 Findings & Recommendation

This text mining project shows that by carefully pre-processing the various job ads posts and the unstructured data, we can gain valuable insights about the Armenian labour market.

By using K-Means clustering explained in section 4.2.1, we created an understanding of the required qualifications and skillset in the Armenia labour market over the 10-year period from 2004 to 2015. We can clearly see that Management and communication skills are in constant high demand over the period and IT related skills have increased over the period.

With the application of LDA statistical model on the full text of the job post with tuned parameters in section 4.2.3, we are able to show the main job “topics” of the online job ads. And by counting the frequency of the job post topics, we revealed that the greatest number of job posting are related to Sales and Marketing with Software development in second place. Plotting the results over the 10-year period show clearly that software development jobs postings have the strongest growth.

Using pattern matching rules on the company’s description, we are able to get a feel of the type of companies that were posting on the job marketplace. This information can be used to supplement the findings in section 4.2.3 to get a better picture of the job market.

In addition to understanding the dynamics of the Armenia labour market, we can perform advanced text analytics to provide added value to the Armenian job portal. We have demonstrated that we are also able to train good classification model to enhance the job portal, automatically classifying if the job posting is IT related. Also, by implementing job similarity search via cosine similarity between the text columns embeddings in section 4.3.2, we are able to create a useful feature that job seekers can use to find similar jobs in the job portal that he/she can apply to.

Recommendation

As the Armenian economy grows and changes, we observed that IT is a skillset that is in demand. It is recommended that the government place more emphasis and effort to provide better education pathway for their citizens to gain IT skills to fulfil the demand. Likewise, for current and future job seekers, it is recommended that if they have interest, they should not hesitate to equip themselves with IT skillset, in particularly the IT operations and applications developments knowledge.

In addition, the CareerCenter.am website can utilize our project Information Extraction methods listed in section 4.3 to improve the search capabilities and features of the website. Enhancing the user experiences and capabilities of the job matching function.