Online Job posting Analysis

Objective Our main business objectives are to understand the dynamics of the labour market of Armenia using the online job portal post.

Business Problem

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.

Data Availability

The job posting informations were given in tabular format, which contains job title, location, description,
Required qualification, resposibilities, date, about company.

Sample job post is given below,

TITLE: Database Developer

LOCATION: Yerevan, Armenia

JOB DESCRIPTION: IUNetworks LLC is looking for a qualified Database Developer to design stable databases, according to the Company's needs. The incumbent will be responsible for developing, testing, improving and maintaining new and existing databases, ensuring the database systems run effectively and securely on a daily basis. He/ she will work closely with developers to ensure system consistency. He/ she will also collaborate with administrators and clients to provide technical support and identify new requirements.

JOB RESPONSIBILITIES:
- Design stable, reliable and effective databases;
- Optimize and maintain legacy systems;
- Modify databases according to requests and perform tests;
- Solve database usage issues and malfunctions;
- Liaise with developers to improve applications and establish best practices;
- Gather user requirements and identify new features;
- Develop technical and training manuals;
- Provide data management support to users;
- Ensure all database programs meet Company and performance requirements;
- Research and suggest new database products, services and protocols.

REQUIRED QUALIFICATIONS:
- BS in Computer Science or a relevant field;
- Proven work experience as a Database Developer;
- In-depth understanding of data management (e.g. permissions, recovery, security and monitoring);
- Knowledge of software development;
- Hands-on experience with SQL;
- Experience in programming with Oracle 11g Server, MS SQL Server, MySQL and PostgreSQL;
- Knowledge of Oracle Exadata is a plus;
- Experience in NoSQL (Couchbase/ MongoDB) is a plus;
- Excellent analytical and organization skills;
- Ability to understand users' requirements; a problem-solving attitude;
- Excellent verbal and written communication skills.

REMUNERATION/ SALARY: Competitive salary, based on skills and experience, plus a medical insurance and biannual company events.

APPLICATION PROCEDURES: Please apply to this job by sending your resumes to: job@iunetworks.am . Please mention the name of the position you are applying for in the subject line of the letter.
Please clearly mention in your application letter that you learned of this job opportunity through Career Center and mention the URL of its website - www.careercenter.am, Thanks.

OPENING DATE: 12 April 2017

APPLICATION DEADLINE: 08 May 2017

ABOUT COMPANY: IU Networks LLC is an information technology company that provides integrated solutions of hardware supply and software development. The Company was founded in March 2008.

Approach

Part 1

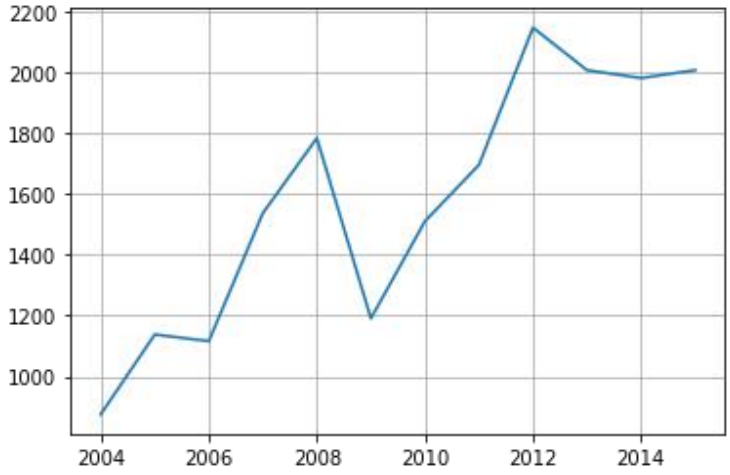
1 Importing the libraries

2 Data Audits

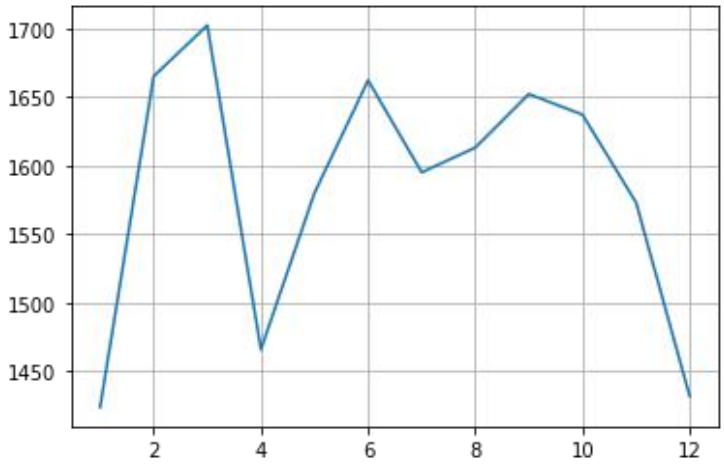
- Check for duplicates
- Check for missing values
- Check language of the data

3 Exploratory Data Analysis

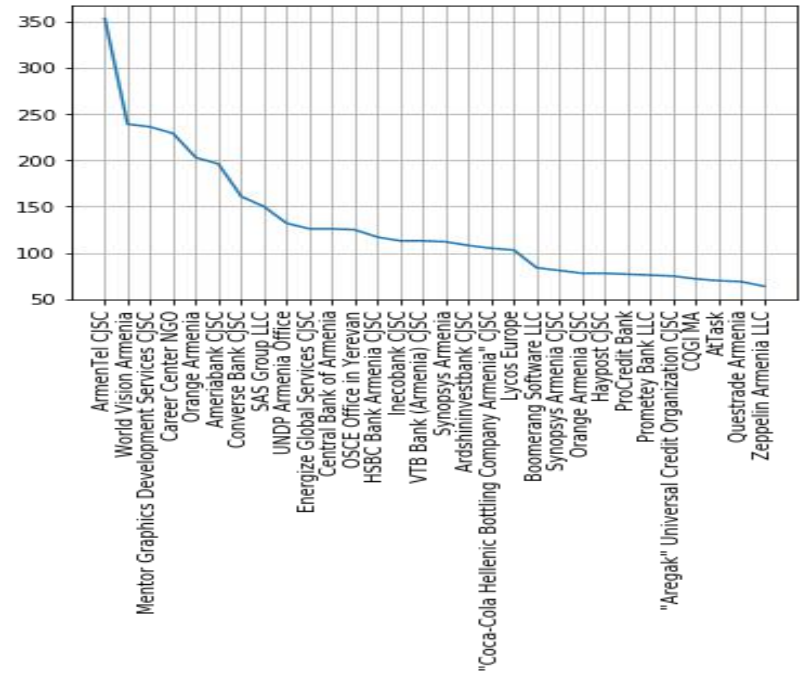
- Job postings by year
  Job post count Vs Year



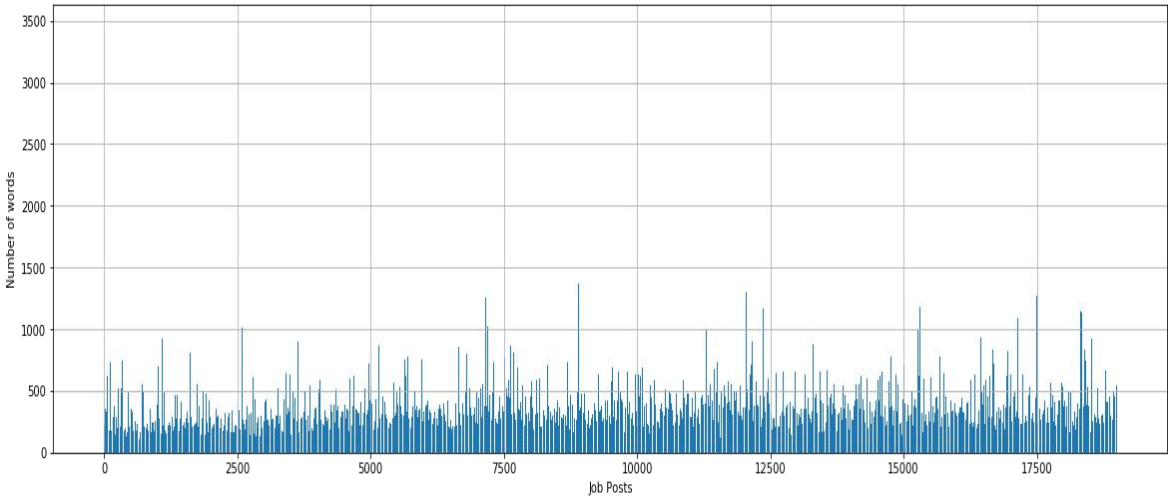
- Job postings by month
  Job post count Vs Month



- Top companies posting the jobs



- Length of job post for each entry



4 Preliminary data preparation

- Imputing the missing values with mode for categorical variables
- Imputing the missing values with median for numerical variables
- Removing the duplicate rows

5 Data Preparation

- Selecting features based on business
  Jobpost
  Title
  JobRequirment
  JobDescription
  RequiredQual
  AboutC
  IT Year

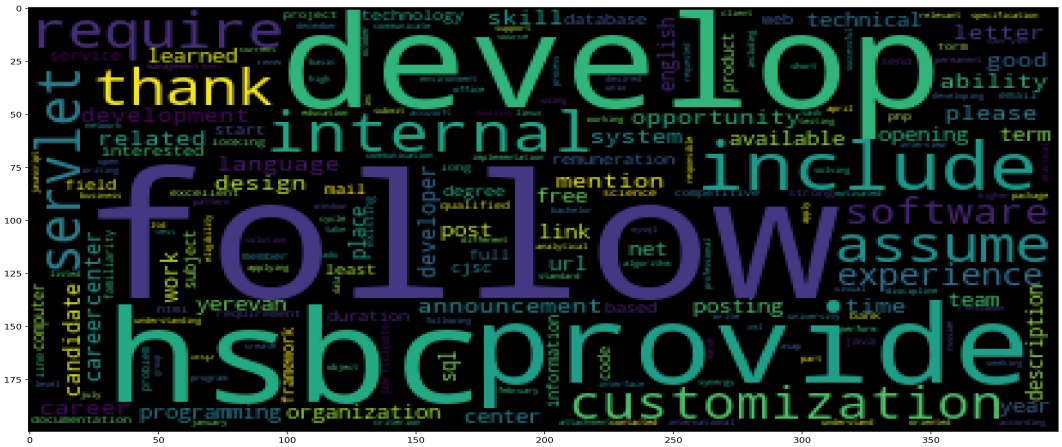
- Data Cleaning

    - Removing symbols, punctuations (Regular expressions)
    - Changing to lower case (Regular expressions)
    - Lemmatization / Stemming (nltk)
    - Removal of stop words (nltk corpus)
      Along with the stop words few other common words were also removed.
    - Removing words except nouns, adjectives, verbs using POS tagging

The final corpus was created after all the data preparation process.

6 Testing the Frequency distribution of the 'Software developer' job title

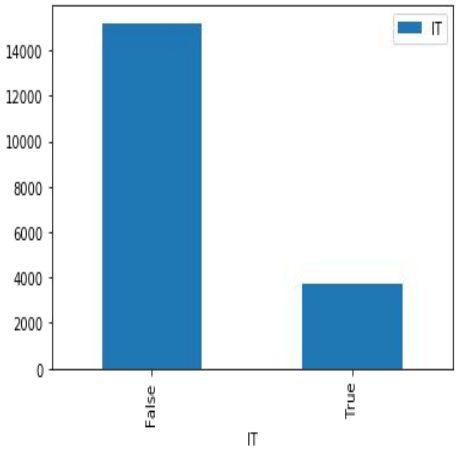
The above processed data was tested for a sample ‘software developer’ post.



7 IT job classification

In this section, we will classify the job postings if it is related to IT sector or not

IT job count Vs other job count



- Supervised learning

    - TF-IDF model
      The data was split into train and test and then the TF-IDF vectorizer was applied.

- Dimensionality Reduction (Singular Value Decomposition (SVD))
  Using the SVD the 21091 dimensions were reduced to 1500 dimensions which was able to explain the variance in the data of about 79%.

- Over sampling (SMOTE)
  Over sampling was done to balance the data, as the True and False percentage was not equal.

- Model Implementation + Evaluation

  - LinearSVC
  - Logistic Regression

The above two algorithms were fitted to the TF-IDF model and the evaluatoin metrics such as confusion matrix, accuracy, roc_auc score, precision and recall were examined.

It was observed that both the models were performing well and the test accuracy of the both the models came out to be 92.75%. Thus, anyone of the model can be used to classify the job as IT or not.

8 Job Qualification Clustering

This section seeks to cluster the job based on the Qualifications.
To understand the Qualification required for each job.

Unsupervised learning

- K Means clustering
  - Train and Test split
    Train and Test was done on the RequiredQual column.

  - TF-IDF vectorization

  - Dimensionality Reduction (Singular Value Decomposition (SVD))
    Using SVD, the dimensions 4876 was reduced to 1500.
    Which was able to explain the variance 0.91%.

  - Finding the Optimum #Clusters (Silhouette Analysis)
    Based on the Silhouette score it was identified that the optimum number clusters can be 7, 9, 11.

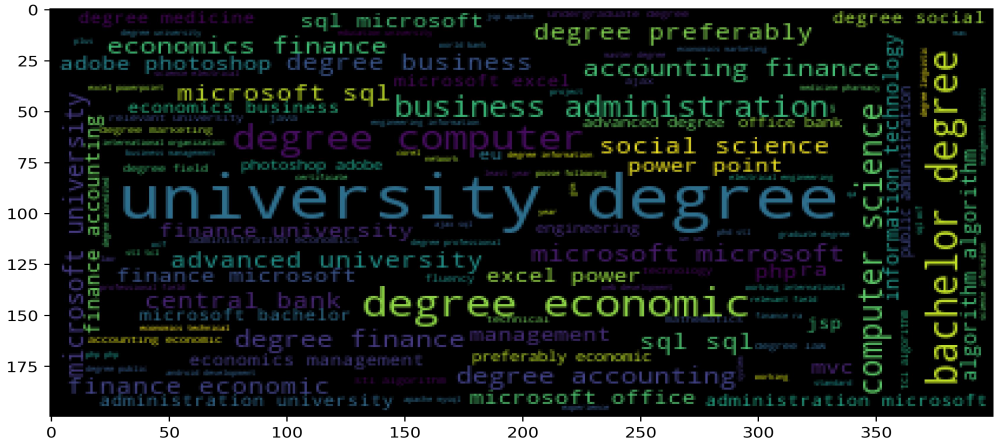
  - Model Implementation
    For each number of clusters each K Means model was created.

  - Cluster Inspection
    The Centroids of each cluster was printed, and examined manually to identify the cluster characterisics.

    Interpretation of the 7 #Clusters solution
    Communication - Cluster 0: work excellent ability russian university degree communication good strong field
    Management - Cluster 1: ability work degree management excellent year good field project least
    Higher education - Cluster 2: higher education work russian excellent good computer ability field year
    Web developer - Cluster 3: net sql web html development server good javascript cs php
    Communication - Cluster 4: eager communicative confident minded punctual timely open complete manner mail
    Accounting - Cluster 5: accounting finance work tax financial software excellent standard good ability
    IT test engineer - Cluster 6: development software testing good design ability programming system linux plus

9 Job skills requirement Entities

- Named Entity Recogition
  In order to understand what are the tools, skills entities requested in the job post,
  Using the Spacy library,
  Each entity whose label is 'ORG' was filteres to find the skills, software tools.



From the above word cloud the following things can be noted,
The most common required qualifications are Microsoft office, Microsoft sql, Power point, computer science degree, excel, finance, business administration

Part 2

10 Job topics

In the different job topics were identified from the jobposts.

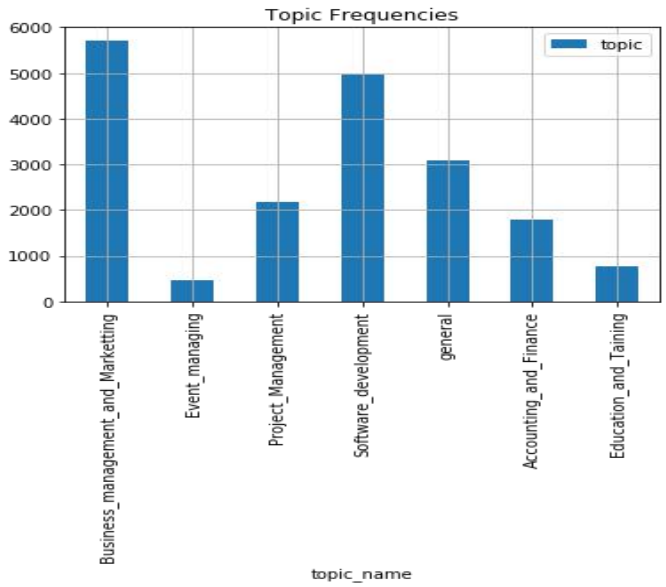
- Topic modelling (LDA)
  - Model Implementation
    The LDA from the gensim library was applied to the job jobpost document term matrix.
  - Analysis of the Topics derived – Number of topics was set as 7. The topic equations were given below.

[(0,
'0.016*"november" + 0.015*"course" + 0.014*"online" + 0.014*"http" + 0.014*"medium" + 0.013*"training" + 0.012*"form" + 0.012*"learning" + 0.012*"october" + 0.010*"expert"'),
(1,
'0.018*"management" + 0.017*"ability" + 0.012*"manager" + 0.011*"business" + 0.011*"line" + 0.010*"ensure" + 0.010*"service" + 0.010*"marketing" + 0.008*"activity" + 0.008*"communication"'),
(2,
'0.025*"interested" + 0.023*"term" + 0.019*"russian" + 0.018*"excellent" + 0.017*"responsible" + 0.017*"subject" + 0.016*"field" + 0.016*"duration" + 0.015*"good" + 0.014*"send"'),
(3,
'0.050*"project" + 0.024*"development" + 0.017*"program" + 0.013*"support" + 0.013*"implementation" + 0.011*"activity" + 0.010*"community" + 0.009*"sector" + 0.009*"including" + 0.009*"international"'),
(4,
'0.022*"development" + 0.020*"software" + 0.017*"team" + 0.016*"design" + 0.014*"customer" + 0.012*"sale" + 0.012*"system" + 0.011*"ability" + 0.010*"developer" + 0.009*"technical"'),
(5,
'0.029*"financial" + 0.026*"legal" + 0.022*"accounting" + 0.019*"report" + 0.014*"prepare" + 0.013*"internal" + 0.013*"finance" + 0.013*"ra" + 0.012*"document" + 0.012*"branch"'),
(6,
'0.030*"office" + 0.024*"administrative" + 0.018*"assistant" + 0.012*"call" + 0.012*"hotel" + 0.012*"task" + 0.011*"relevant" + 0.011*"meeting" + 0.011*"written" + 0.010*"event"')]

From the above the lda out the following topics are found for job posts,

Topic 0 : Education and Taining
Topic 1 : Business management and Marketting
Topic 2 : general
Topic 3 : Project Management
Topic 4 : Software development
Topic 5 : Accounting and Finance
Topic 6 : Event managing

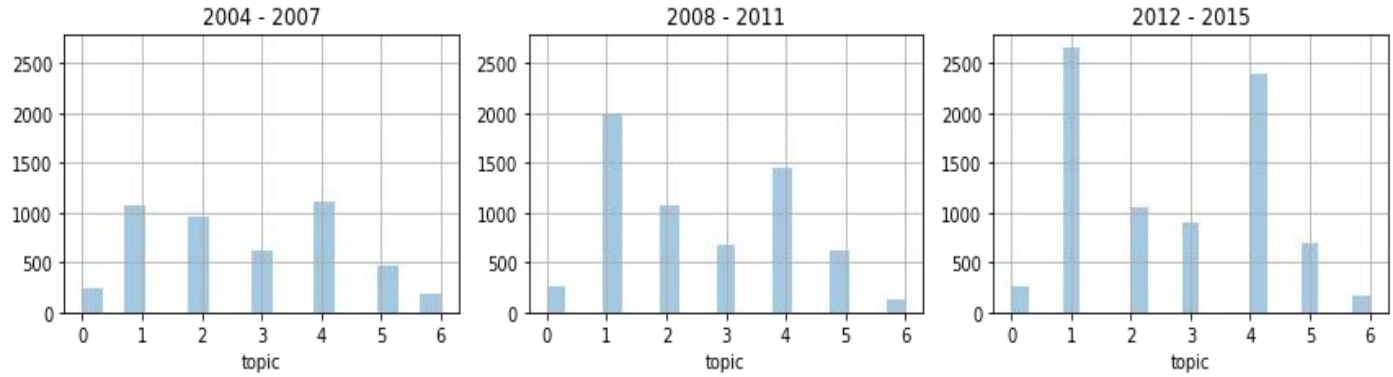
- Topic frequencies



- Visualization using pyLDAvis

topic_visualization.html

- Observing the changes in the Job topics over time (4 years sections)

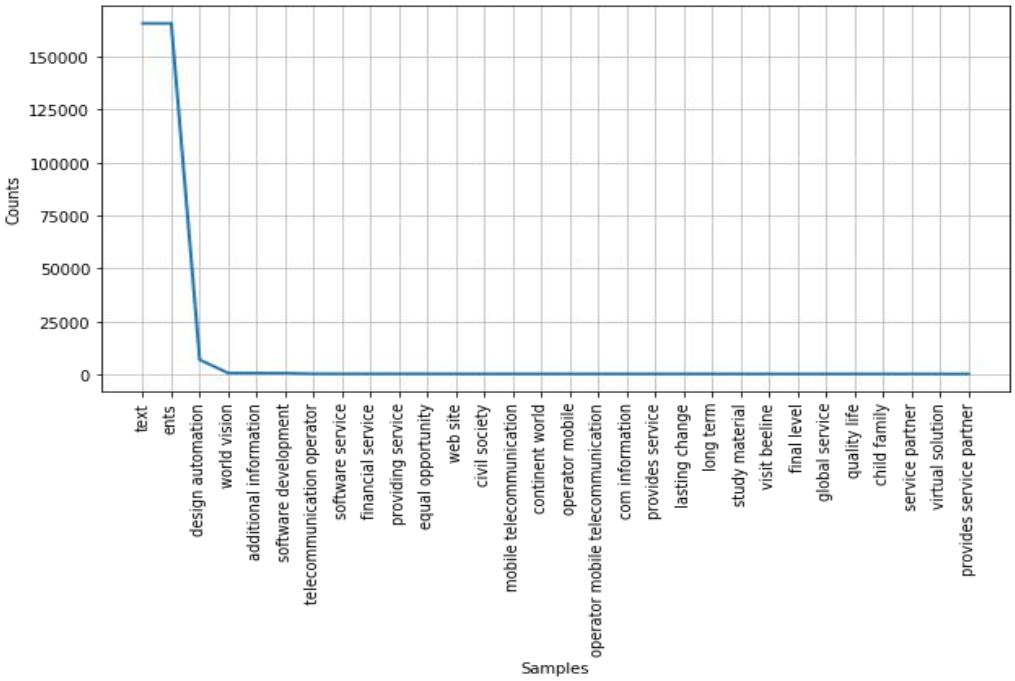


From the above plots,
There is an increasing number of job posts over the years.
The pattern remains the same for the three sections.
Business management and Marketting showing highest increase every year which was lower than Software development initially.
Event management shows declining pattern

11 Company profiles analysis

- Applying Information Extraction on Companies profiles (Pattern Matching)
  A total of 11 patters were defined in order to extract the company profile from the AboutC (About Company) column.

- Defining the patterns and the frequency analysis



These are some top frequent company profiles.

('design automation', 6835),
('world vision', 564),
('additional information', 558),
('software development', 512),
('telecommunication operator', 257),
('software service', 245),
('financial service', 244),
('providing service', 240),
('equal opportunity', 236),
('web site', 221),
('civil society', 220),
('mobile telecommunication', 220)

12 Job Similarity

To identify the job similarity between jobs, four columns were selected. Title, RequiredQual, JobDescription and JobRequiement

- TF-IDF vectorization
  The TF-IDF vectorization was done on each of the above selected four columns.

- Cosine simialrity
  A function was created for finding the similarity between the jobs based on the cosine similarity metrics of all the four columns and returns the top similar columns.

- Analyzing the similar jobs for each job
  Few of the similar jobs were listed below,

Software developer

Job title = software developer
Similar jobs are,

software developer

converse bank cjsc software developer yerevan description design develop softwar
SIMILARITY SCORE: 0.5760695155473291

---

software developer

converse bank cjsc software developer yerevan description develop implement new
SIMILARITY SCORE: 0.5737643542416898

---

software developer

cascade capital holding cjsc software developer yerevan description develop data
SIMILARITY SCORE: 0.5656704358968097

Internship

Job title = full time community connection intern paid internship
Similar jobs are,

teacher full time

military institute mod ra full time yerevan degree computer science information
SIMILARITY SCORE: 0.6824664284763577

---

non paid part full time administrative intern

international research exchange board irex non paid part full time administrativ
SIMILARITY SCORE: 0.67663381892011

---

non paid part full time programmatic intern

international research exchange board irex non paid part full time programmatic
SIMILARITY SCORE: 0.665398959778587

Accountant

Job title = chief accountant finance assistant
Similar jobs are,

chief accountant

moscow state university economics statistic information chief accountant term fu
SIMILARITY SCORE: 0.5551694415320927

---

chief accountant

doubletree hilton hotel yerevan chief accountant term full time duration long te
SIMILARITY SCORE: 0.5516590377191786

---

chief accountant assistant

caparol georgia ltd chief accountant assistant tbilisi georgia description board
SIMILARITY SCORE: 0.5457632645048219

---

13 Findings and Recommendations

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**, 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, 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 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, 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.