Performing Language modeling on job profiles

Gaurav Gade, Siddharth Parmar, Anubha Gupta

Introduction:

Language modeling, text analysis and semantic analysis are powerful tools to analyze large chunks of textual data. Writing a proper job requirement is a complex and detailed process and keeps evolving according to the needs of the company. Most jobseekers have a hard time finding the best position that matches their interests. Especially since job requirements can vary significantly even though the job titles are the same or similar. In today's modern internet era, there are several applicants per job role. According to Glassdoor [1], each corporate job in America attracts 250 resumes. Only 4-6 are selected and 1 candidate is offered the job. Most of these resumes go through the initial step of resume parsing where the resume content is parsed and analyzed by a text mining algorithm that scores every candidate based on criteria set by the company. As part of this study, we are going to analyze a subset of 19,000 job postings to determine which keywords are most frequently occurring in the job requirement. We will investigate the relationship between these top keywords which are usually a combination of management, technical and inter-personal skills. Lastly, we will perform a job similarity analysis and perform topic modeling to view patterns between different job roles.

Why does this topic matter?

In today's digital age, job applications have increasingly become online, and data driven. Most companies use some level of computational text parsing of resumes to shortlist candidates before a recruiter physically looks at the resume. There may be significant variance between different job postings even when the job title may be the same. Hence, it is important for the job seeker to be able to identify roles that are most relevant to his or her skill set. Analyzing job postings will help us to reveal interesting patterns between the job data. We hope to detect similarities in job requirements and find out strongly correlated skills. We have used topic modeling, similarity models, word to vector and network modeling to analyze these postings. The LDA visualizations have been used to identify common keywords across the postings for the top 5 themes. Universities can utilize the inferences from this study to better plan coursework so that students will receive most of the skills required to meet the most frequency sought after job roles. The study results may be of huge interest to professional job-related companies like Indeed.com and LinkedIn. As part of this research study, we are performing text analysis and language modeling on job profiles

Related Background:

While applying for jobs online, we noticed a huge variance in job description even though the job titles were similar. Our research study aims to analyze these 2000 postings using text modeling techniques and linguistic models like vector space, clustering and network models to determine which job descriptions are semantically similar. This is an efficient method for job search.

While we were studying and hunting for jobs in college, it was difficult to gauge which skills are hot in demand for a given job title. By using these linguistic models, we can determine the hottest skills that are currently required in the job market for a specific job. So, college students as well as universities can customize their curriculum accordingly. Previous literature review has been shared as citations in references. [2] The team discusses about German job advertisements and applies word2vec modeling to create semantic models.

Research Question

Our research aims to semantically compare job descriptions to determine the most relevant / similar job(s) for a specific job search. We seek to find patterns between themes using topic modeling. We aim to analyze entity-relationships using network modeling to find correlation between job requirements. Lastly, we seek to build a classifier model that can predict whether the job is an IT job or not. The dataset was sourced from Kaggle which had job descriptions from 2004 to 2015 based out of Armenia. We have used a subset in the year 2015. The columns used for analysis are Job title, Job requirement (detailed text) and IT (if IT job or not). [2]

To help the job seekers on creating resumes is another important factor that has been addressed. Often users with great resumes do not even make it to the interview as they get filtered out for not using the right keywords. We ran a topic modeling algorithm that picks out most common keywords related to a job profile. We also ran a network model that allows us to look for most commonly paired words used in job descriptions. For example: Design and develop. This model also provided further insights to the skills one must have for a particular role.

Analysis

Network Modeling.

We performed Network modeling on the data to find out the most common paired words used in the job description

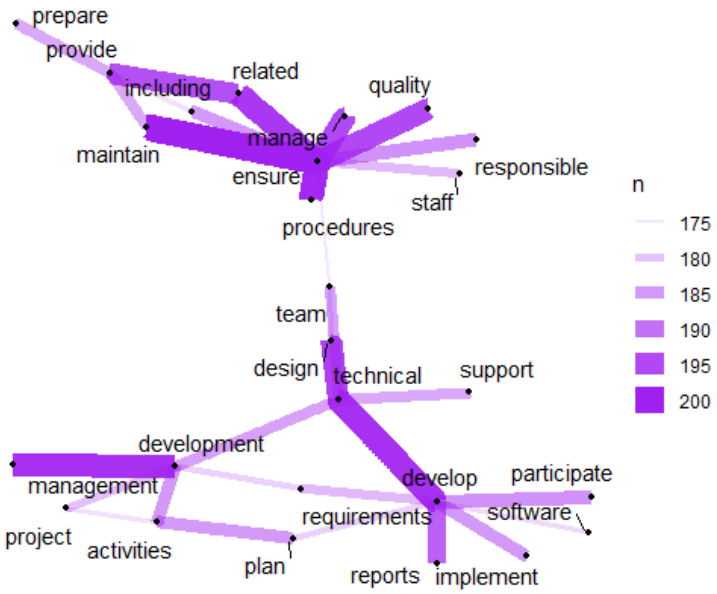


Fig1.1 – Network Plot showing relationship between key terms used in IT job postings.

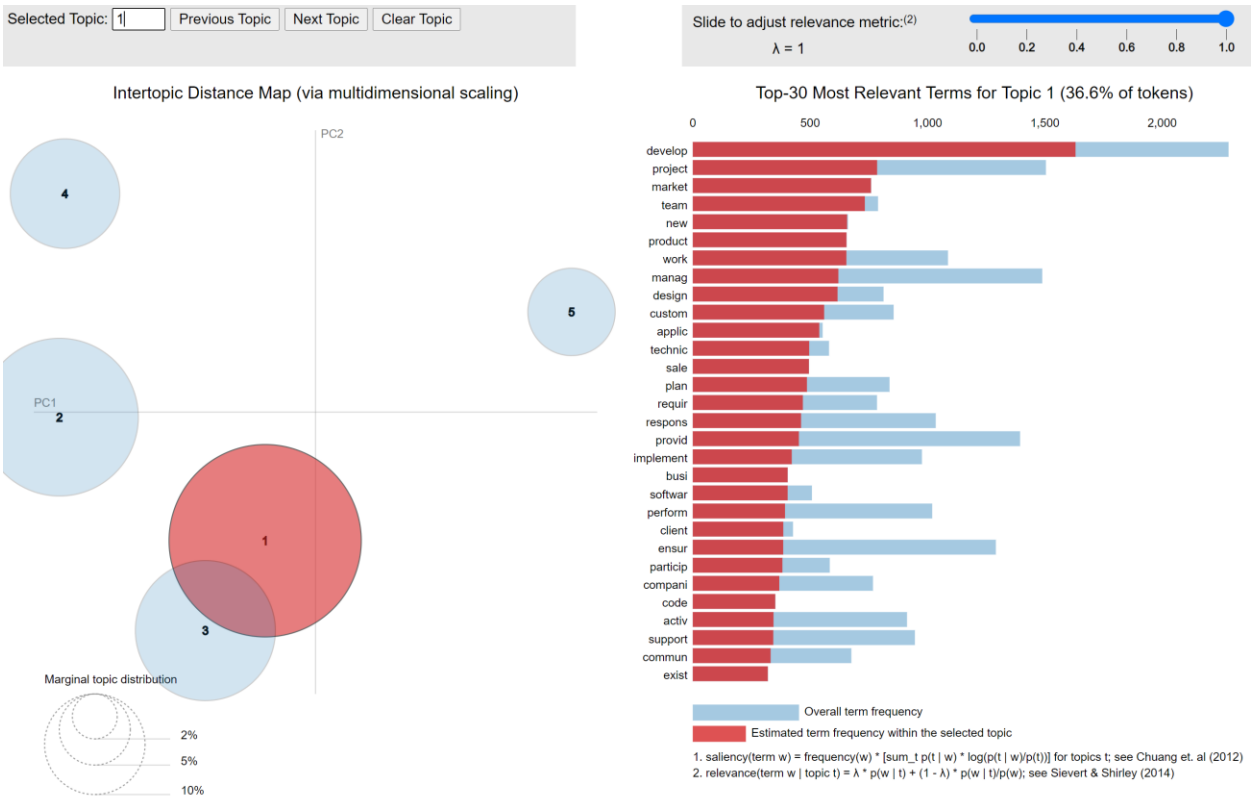
Interpretation:

The network plot shows us major words such as design, technical and develop that other related branched words such as participate, software reports. This helps us understand that the technical design and develop functions in a job description also require the skills of reporting and working in teams. It also needs focus on implementation and most of these are paired with software.

Similarly, ensure, quality, procedures are the branched words mostly paired with manage. For a managerial role, using keywords such as ensure quality and maintain procedures with help boost the probability of job seekers getting filtered for a job applied.

Topic Modeling

We have used the gensim topic modeling on the job posting data to evaluate which topics are most prominent based on the text column of Job Requirement.



Interpretation:

The LDA visualization plot shows the 5 most prominent topics that have been created based on IT job postings. The blue charts show the overall term frequency and the red graph show the estimated term frequency.

The 5 themes reveal the following:

Theme 1: test, develop, media, website, material, design, content, public, work, use. These seem to be related to designing web pages and web-site development for public work and government projects.

Theme 2: project, active, support, implement, including, develop, program, coordination, organ, provide. These words seem to indicate higher level program co-ordination and implementation,

Theme 3: management, ensure, provide, control, intern, legal, system, operations, security, information. These words seem to be related to legalities, compliance and security. May also include network security and operations.

Theme 4: report, account, preparation, financial, custom, response, perform, make, data, tax. These words seem to include tax filing, tax compliance, financial reporting and accounting.

Theme 5: develop, project, market, team, new, product, work, manage, design, custom. These words seem to include product development, product management, design projects and customizations.

Similarity between job descriptions.

We create a corpus vectorizing the words and used cosine similarity algorithm to determine the most similar job descriptions with job title as an input to the model. Following examples are shown for 2 such job titles.

Interpretation:

The job similarity module is extremely useful to quickly find out similar job roles based on the job requirement, roles and responsibility rather than just the job titles. As we know, job titles can vary significantly and by running this module, we can get the most similar roles even though the job titles may be different.

In this case, we tried to find out job titles like 'CFO' and 'Software Engineer'. For CFO, we found that the job titles similar to CFO based on job requirement were:

1. Finance Analyst
2. Chief Financial Officer (CFO)
3. Auditor
4. Senior Auditor
5. Chief Financial Officer (CFO)
6. Senior Finance Specialist
7. Chief Accountant
8. Financial Controller

For 'Software engineer', the titles similar were:

1. DevOps Software Engineer
2. Senior Java Developer
3. Java Developer
4. Java Developer
5. Software Engineer
6. Java Developer
7. Front End Developer
8. UX Designer
9. Java Developer
10. Web UI Developer

Word2vec model:

The word2vec is a neural network similar to a skip gram model that allows creation of meaningful models that understand word embedding from large bodies of texts.

```
#print out results
print('accuracy %s' % accuracy_score(y_pred, y_test))

## accuracy 0.837465564738292
```

```
print(classification_report(y_test, y_pred,target_names=my_tags))

##              precision    recall  f1-score   support
##
##        TRUE       0.84      0.98      0.90       280
##       FALSE       0.83      0.36      0.50        83
##
##    accuracy                           0.84       363
##   macro avg       0.84      0.67      0.70       363
## weighted avg       0.84      0.84      0.81        36
```

Interpretation:

In our case we have a huge list of job requirements that can been categorized as IT and NON-IT jobs. This logistic regression model allows us to predict the category. We trained 80% of the dataset we had and 20% of the dataset was tested. The output shows 86% accuracy of the model.

Discussion

The accuracy of our model is about 86% which is based on Logistic regression classifier. The data that we have concentrated on for our analysis is all 2015 data and is based on Armenian job market data.

We analyzed related literature and found some interesting research papers that discuss their approaches and findings for similar real world problems. The research paper 'Document-based Recommender System for Job Postings using Dense Representations'[3]discusses the usage of dense vector representations to enhance a large-scale job recommendation system and tries to rank German job advertisements regarding their similarity with the best results obtained by combining job titles with full-text job descriptions. The research paper 'Job Recommendation From Semantic Similarity of LinkedIn Users' Skills'[4] tries to find out relationships between jobs and people skills making use of data from LinkedIn users' public profiles. The authors have applied Latent Semantic Analysis (LSA), and hierarchical clustering of job positions to build a job recommendation system. Another paper 'An Information-Geometric Approach to Document Retrieval and Categorization' tries to develop from first information-geometric principles a general method for learning the similarity between text documents and derived a canonical similarity function - known as the Fisher kernel. [5] Another paper "Methods and Metrics for Cold-Start Recommendations" discusses how the authors developed a method for recommending items that combines content and collaborative data under a single probabilistic framework using a naive Bayes classifier on the cold-start problem, to obtain deeper understanding of the performance characteristics of recommender systems. [6]

Conclusion and Recommendations

The online job market is a good indicator of overall demand for labor in the local economy. Based on this research, we understood the demand for certain professions, job titles, or industries and how we can help universities with curriculum development. We can identify skills that are most frequently required by employers, and how the distribution of necessary skills

changes over time. This would help us make recommendations to job seekers and employers and find similar jobs based on the ones that you have previously applied for.

Citations/References:

1. 2020 HR Statistics: Job Search, Hiring, Recruiting & Interviews. (2020). Https://Zety.Com/Blog/Hr-Statistics#:~:Text=On%20average%2C%20each%20corporate%20job,One%20will%20get%20the%20job. https://zety.com/blog/hr-statistics#:~:text=On%20average%2C%20each%20corporate%20job,one%20will%20get%20the%20job.
2. Hab, M. (2016). Online Job Postings. Retrieved 2020, from https://www.kaggle.com/madhab/jobposts/notebooks
3. Elsafty, A. E., Reidl, M. R., & Beimann, C. B. (2017). Document-based Recommender System for Job Postings using Dense Representations. Document-Based Recommender System for Job Postings Using Dense Representations, 216–224. https://aclweb.org/anthology/N18-3027.pdf
4. 2016

   Giacomo Domeniconi, Gianluca Moro, Andrea Pagliarani, Karin Pasini, Roberto Pasolini

   10.5220/0005702302700277

5. Hoffman, T. H. (n.d.). Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization. Https://Proceedings.Neurips.Cc/. https://proceedings.neurips.cc/paper/1999/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf
6. Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and metrics for cold-start recommendations. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’02, pages 253–260, Tampere, Finland