香港中文大學

計算機科學及工程學系

Department of Computer Science and Engineering,

The Chinese University of Hong Kong

Title

Artificial Intelligence Supported Online Matching Platform

Name(s)

Lo Chun Hei 1155077843

Supervised By

Prof. LEUNG Kwong-Sak

Table of Contents

ACKNOWLEDGEMENT

I would like to thank Professor Leung and Mr Sunny Lai for their expert advice throughout the project. I would also like to thank Mr Qiu Yao and Miss Lyu Si Ning for their collaborative effort in developing the webpages for the job matching system.

0. ABSTRACT

This report serves to outline and compare different approaches to the task of CV-JD (job description) matching, explains the technical issue involved and describes an approach to the task. It focuses on the use of natural language processing in facilitating matching of jobs and resumes on the semantic level. In particular, document modelling methods are applied to a set of CVs and JDs to quantitatively compare documents and their suitability against each other. A hybrid method is specifically developed for the CV-JD matching scenario.

1. INTRODUCTION

A company's success depends much on the talents it recruited; and one's successful career starts with a right job. That is why the quality of recruitment and job seeking processes are crucial to both parties. From Forbes 2018, it is found that more than half of talent acquisition leaders consider it to be most difficult to identifying the right candidates from a large applicant pool of recruitment processes. Particularly, this is traditionally done by hand without support of computers. It is reported that talent acquisition professionals spend nearly one-third of their work week sourcing candidates for a single job role. On the other side, a job seeker may also face too much choice of jobs when they do job searching online. Therefore, here comes the need of a bridge between job seekers and job providers.

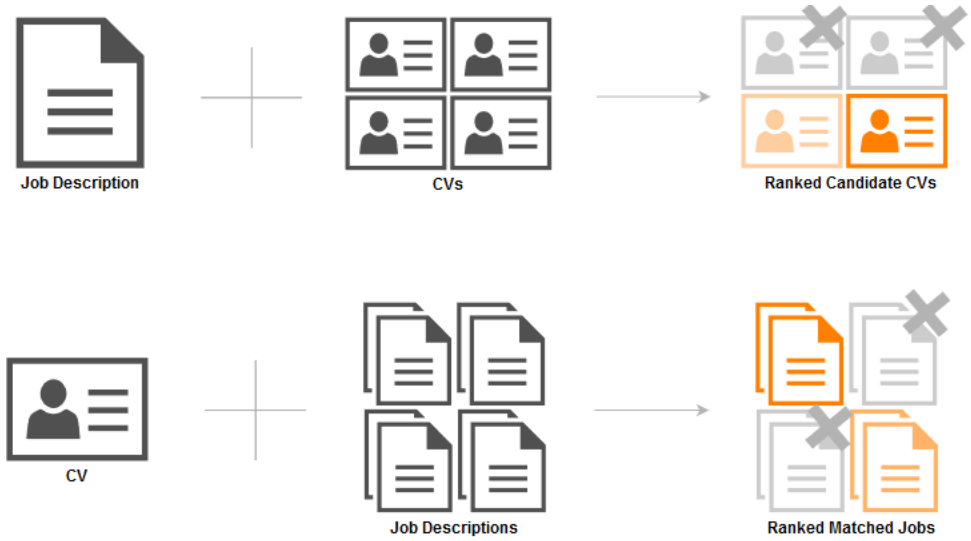


Figure 1: Job Matching Process

With vast information of jobs and applications, an automated solution for matching suitable jobs and candidates to each other is crucial for the benefits of both. To job providers, with automated screening on candidates' profiles, background and qualification of suitable candidates would be identified and those who do not fit are filtered at ease at the first stage of recruitment process, hence greatly reduce effort of checking. Furthermore, companies may also do headhunting conveniently among a sea of available resumes. They can rely on machines to help screen out the preferred candidate before looking into all of them with extra effort. To job seekers, automated job matching would guide them to the jobs that matches their qualifications and skills according to their submitted profiles.

The development of job matching system and algorithms would be essential in achieving the benefits mentioned. When most of the information of a job candidate and a role is in the form of text, one key area of such matching system is to extract relevant information from the documents. To make natural language comprehendible by machines, we rely much on natural language processing (NLP) to convert human understandable text into data form of machines.

- **Contributions**

In summary, this report presents the following contributions: (1) Over 60,000 semi-structured CV and raw JD documents are scrapped; (2) document-level and word-level representations are learnt using the scrapped data; and (3) a hybrid matching algorithm that uses the learnt representations is developed for comprehensive CV-JD matching.

2. SYSTEM DESIGN SPECIFICATION

2.1. Data Flow Diagram

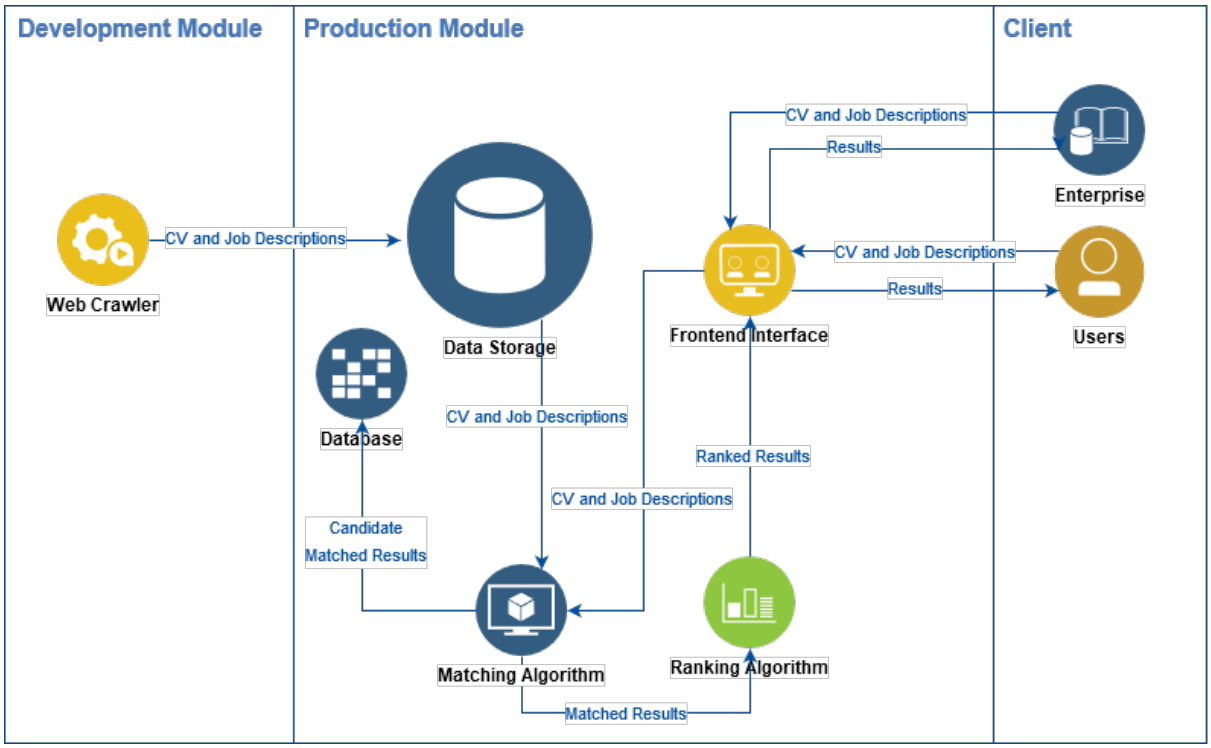


Figure 3: Dataflow Diagram of the system

The data flow of the system is done between three main components, namely Development Module, Production Module and Client. A data storage system will hold the files of job descriptions and CVs. The sources of these documents are either from the web or provided by the clients, who are enterprise users and individual users. Within the production module, analysis will be performed and different algorithms will be applied to produce matched job results.

2.2. Description of Components

The system is developed collaboratively with two Msc students. My work focuses on retrieving job description and CVs data from the internet and the matching algorithm. The following descriptions shall only focus on my part of work.

2.2.1. Web Crawler

Real CVs and job descriptions data are needed for two reasons: for evaluating the performance of our job matching model; and for training the matching and ranking algorithms before any users actually submit data to our platform, or when the amount is too little or too sparse.

To obtain substantial amount of data, publicly available online resources are sought by parsing html into structured formatted data. About 40000 publicly available CVs are obtained from Indeed.com using a web crawler. The CVs are semi-structured with work experience, skills, education and additional information, whilst JDs are raw text. About 20000 job

descriptions are also obtained from Indeed.com, with a majority of jobs from the information technology industry and the remaining comprises from Accounting/Finance, Upper Management/Consulting, etc.

2.2.2. Matching Algorithm

The matching algorithm is described in details in Section 3.2.2.

2.3. User Interface Design

2.3.1. Index Page

The index page provides the basic functionality of account registration and login function.

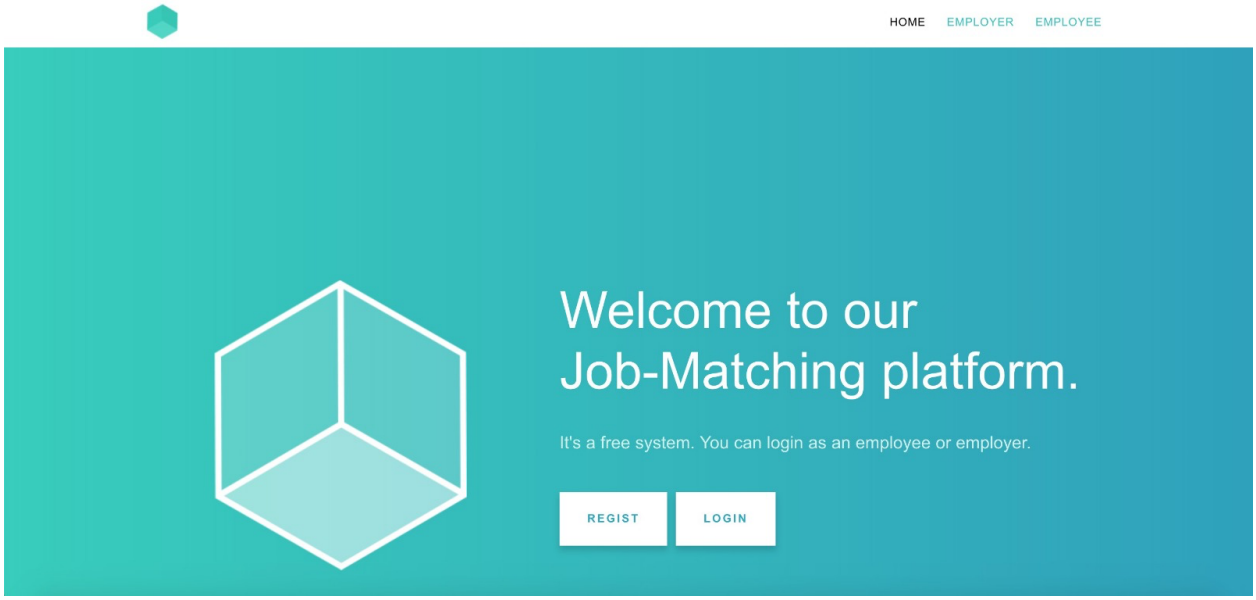


Figure 8: Index Page

2.3.2. Job Description Upload Page

This page provides the function of uploading text files of job description.

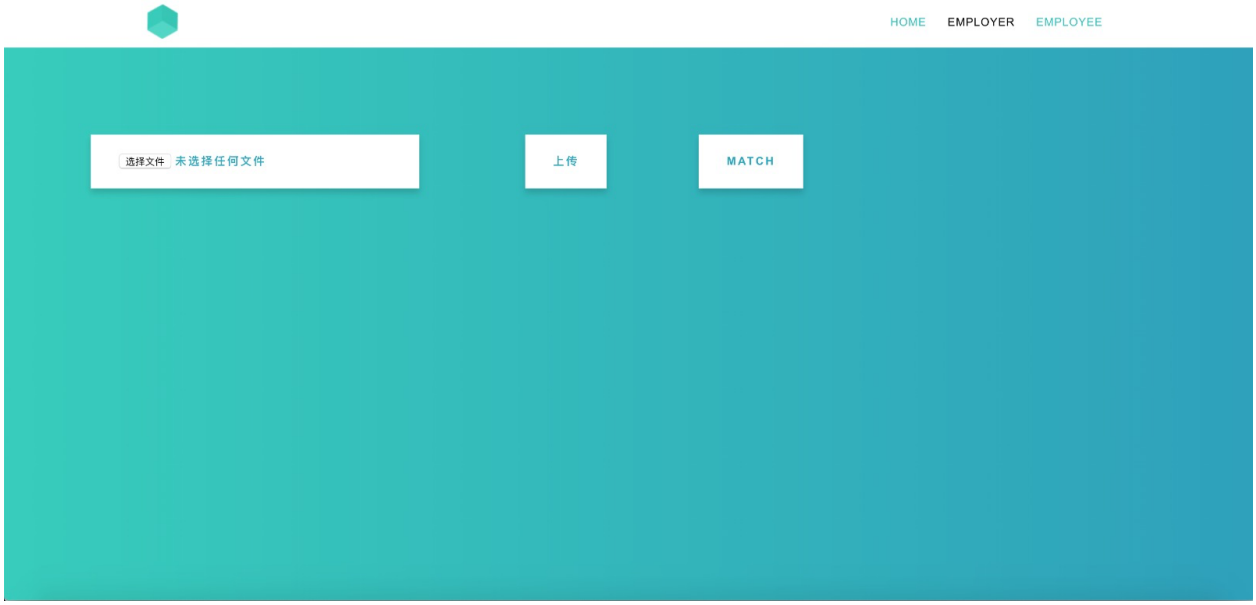


Figure 9: Job Description Upload Page

2.3.3. Radar Chart of Matched Results

A radar chart is displayed upon successful matching between a job and a job seeker. It illustrates the quality of matching based on different matching criteria.

Figure 10: Radar chart of matched results

3. CV-JD MATCHING ALGORITHM

3.1. Related Work on CV-JD Matching

Various models have been developed for matching CVs and JDs. Some related work is studied briefly here for reference, but not in deep, as the settings, data and objective of each work is different, and the problem outcome is subjective to a large extent.

3.1.1. Using Deep Learning To Extract Knowledge From Job Descriptions

In this project, a deep learning model is presented to extract knowledge implicitly from 10 million vacancies of recruitment data. The data set was split into a set of 8.5 million for training and a set of 1.5 million for testing.

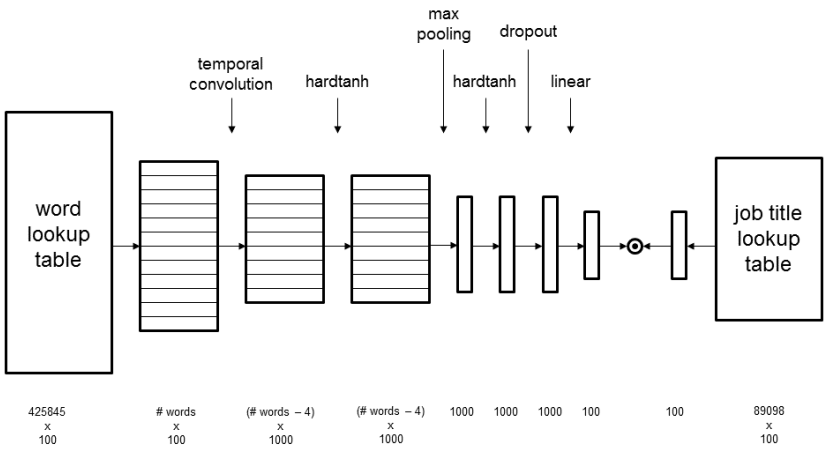


Figure 14: Model architecture

The model they used is a convolutional neural network (CNN), which would accept bag-of-words input embeddings of the 425,845 most frequent words that exists in the job description and trained using learning-to-rank approach with pairwise hinge loss, given the label of job title corresponding to the job description, which is the embeddings of the job title.

The trained model possessed the ability to predict the most possible job titles given the job description. However, it only extracts knowledge from the final output layer and does not handle job matching problem.

3.1.2. Matching Resumes to Jobs via Deep Siamese Network

This paper leverages Siamese convolutional network for identifying matching job description and CV pairs. It also experiments with different representations of documents and compares them to their proposed approach. However, their training data consists only of about a total of 5000 CVs and JDs, and their training data is labelled by human in loose manners.

3.2. My Approach to Job-CV Matching

3.2.1. Problem Definitions

There are several issues regarding the matching problem, and one of them would be that the quality of matching results is subjective. Employer and applicants may add in extra matching criteria and personally preference of suitable candidates/jobs, which could not be totally reflected from CVs and JDs. Therefore, the matching algorithm should mainly focus on the

similarity between the contents of two text documents. Particularly, an optimal matching of a CV-JD pair should have the work experience, skills and additional information of the CV match with the work descriptions and requirements of a JD.

Formally, the problem to be solved is: given a set of CVs and a set of JDs, return a ranked list of CVs (JDs) that best match a JD (CV).

3.2.2. Data

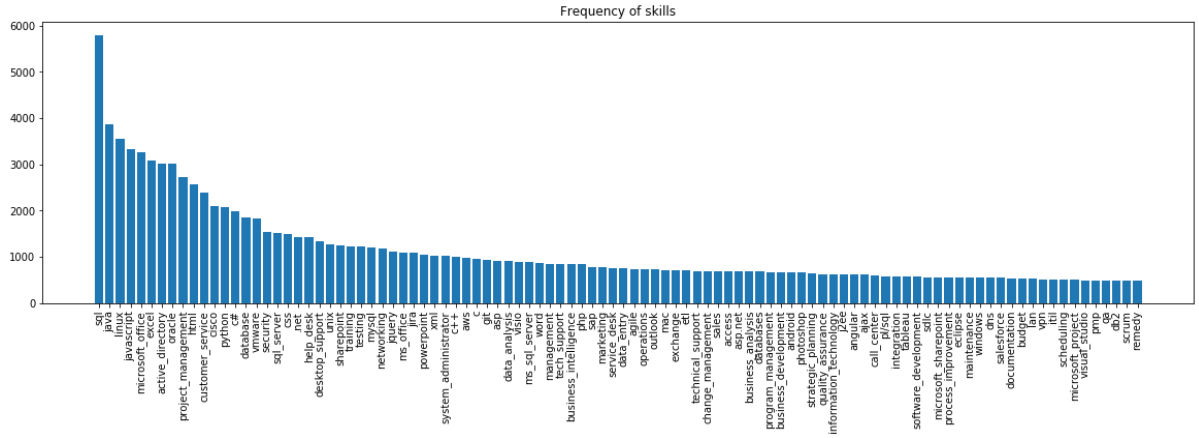


Figure 15: Skill distribution of crawled CVs

The data to be used is the same as those described in section 2.2.1 (see appendix 6.3, 6.4). For the evaluation purpose, each CV is divided into two sections, namely the work experience of current job and the work experience excluding current job plus additional information (see appendix 6.5). They will be called generated test data hereafter. Figure 15 shows the number of job seekers possessing different skills. Most of the skills belong to the IT industry.

3.2.3. Hybrid Matching Algorithm

The key information contained in a CV that is crucial for employers includes the main body of work experience and additional information, the skills and their respective number of years of experiences and current job title. Therefore, it is essential to take every of these into account when deciding for suitable positions.

I propose to use vector space model for modelling CVs and JDs since it is easy to understand, allows swift computation of different measures and thus quantitative comparisons among vectors.

The process of generating document vectors is as follows:

- **Pre-processing of CVs and JDs**

Much pre-processing for CVs and JDs has to be done before applying any transform:

1. JDs from the same company often contains same paragraphs that describes the company or equal opportunity employer declaration, but not the job itself, which therefore could add noise to the job descriptions. Hence, they are removed by

removing the longest possible longest common substring of each JD with other JDs from the same company.

2. Tokenize each document so that each document is represented as a list of tokens
3. Extract bigrams using the simple data-driven approach [1]:

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i, w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

The metric is higher when two words appearing together more than proportional to that of their individual occurrence. Two words scoring above a certain threshold will form a bigram as a result.

4. Semantically unmeaningful stopwords like 'is', 'using', etc. are removed from documents.

- **Vectorizing Documents**

Vector representation of document has been a hot research topic, and many methods have been invented to produce document embeddings. In this project, two of them are experimented with and studied.

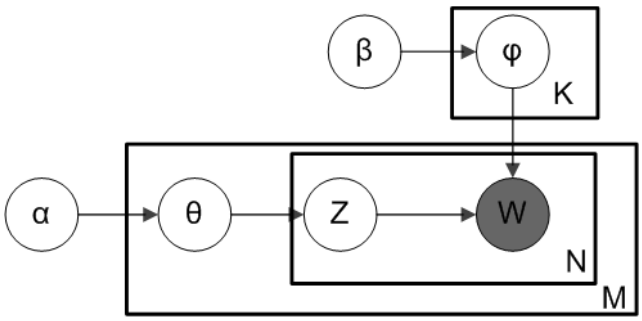


Figure 16: Plate model of LDA

One of the ways to vectorize documents is applying Latent Dirichlet Allocation (LDA) [11]. LDA is a generative probabilistic model for collections of discrete data such as text corpora. The three-level hierarchical Bayesian model treats each word of a collection as a finite mixture over an underlying set of topics, and each topic is modeled as an infinite mixture over an underlying set of topic probabilities. Therefore, a document can be explicitly represented as its topic distribution.

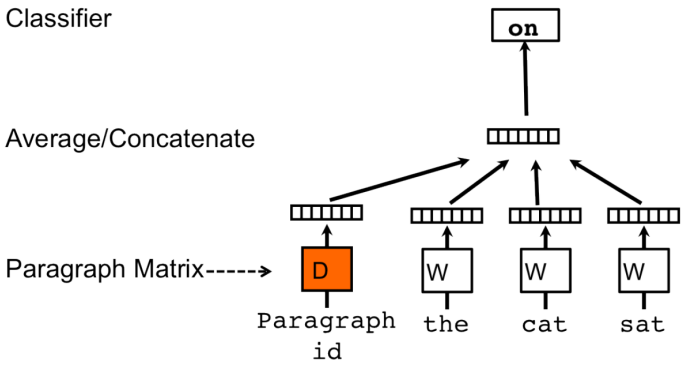


Figure 17: Learning document vector in doc2vec

Another current approach on document modelling is using *Distributed Representations of Sentences and Documents* (as known as Doc2vec) [12]. The model uses distributional information of text and train document vectors in an unsupervised manner. As shown in figure 17, the concatenation or average of the paragraph vector with a context of words is used to predict the following word.

The trained models (LDA/Doc2vec) will generate document vectors $e_d(\mathrm{CV}_i)$ for CV_i and $e_d(\mathrm{JD}_j)$ for JD_j.

- **Vectorizing Job Titles and Skills**

Apart from document vectors, LDA and Doc2vec also generate embeddings for token in the same embedding space: in LDA, each token is modelled as a distribution over the set of topics in; and each token in Doc2vec is trained simultaneously when training for paragraph vectors. Thanks to the two models, job titles and skills may also be embedded to the same semantic space as the documents.

For document i,

The embedding of current job title of a CV could be given by:

$$e_t(\mathrm{CV}_i) = \mathrm{avg}\big(e_{token}(w)\big) \text{ for } w \in tokensOf(jobTitle_i)$$
$$= \frac{\sum_{w \in tokensOf(jobTitle_i)} e_{token}(w)}{|tokensOf(jobTitle_i)|}$$

where:

$tokensOf(jobTitle_i)$ is the tokens extracted from the job title (after removal of stopwords and extraction of bigram as described in section 3.2.3. Pre-processing of CVs and JDs), and

$e_{token}(w)$ is the embedding of token (unigram or bigram) w (if there exists).

For CV i, the embedding of the skills possessed by the candidate could then be given by:

$$e_s(\mathrm{CV}_i) = \frac{\sum_{skill_j \in skills_i} e_{token}(skill_j) \times year_j}{\sum_{years_j \in years_i} year_j}$$

where:

$skills_i$ and $years_i$ are the set of names of skills and respective number of years of experience possessed by the candidate i,

$e_{token}(skill_j)$ is the embedding of skill name $skill_j$ (if there exists),

$year_j$ is the number of years of experience of $skill_j$.

In other words, the embedding of the skills of the candidate is the weighted (by number of years of experience) sum of the respective skill name embeddings.

A wide range of matching could be performed through manipulation on document, job title and skills vectors:

- **CV-JD Matching Based on Document Vectors**

This is the basic functionality of the CV-JD matching system, where CVs and JDs that share similar textual content would be matched against each other. The degree of similarity could be quantified by measuring the cosine similarity, of CV-JD pairs in the embedding space, i.e. by

$$sim_d(\mathrm{CV}_i, \mathrm{JD}_j) = \frac{e_d(\mathrm{CV}_i) \cdot e_d(\mathrm{JD}_j)}{|e_d(\mathrm{CV}_i)||e_d(\mathrm{JD}_j)|}$$

- **CV-JD Matching Based on Document, Job Title and Skills Vectors**

Skills of a candidate is also a crucial information to employer, and they may be profiled with skills vectors. Weighted-summing of skills provide information to the overall 'direction' of the skills of a candidate.

Supplementing to the document content, job title vector of the current job of a CV can also be compared against that of the job title of a JD.

A simple workaround for integrating the three measures would be to consider the weighted sum of the three similarities:

$$sim_{d,t,s}(\mathrm{CV}_i, \mathrm{JD}_j) = \alpha sim_d(\mathrm{CV}_i, \mathrm{JD}_j) + \beta sim_t(\mathrm{CV}_i, \mathrm{JD}_j) + \gamma sim_s(\mathrm{CV}_i, \mathrm{JD}_j)$$

where:

$$sim_s(\mathrm{CV}_i, \mathrm{JD}_j) = \frac{e_s(\mathrm{CV}_i) \cdot e_d(\mathrm{JD}_j)}{|e_s(\mathrm{CV}_i)||e_d(\mathrm{JD}_j)|} \; or \; |e_s(\mathrm{CV}_i) - e_d(\mathrm{JD}_j)|$$

$$sim_t(\mathrm{CV}_i, \mathrm{JD}_j) = \frac{e_t(\mathrm{CV}_i) \cdot e_t(\mathrm{JD}_j)}{|e_t(\mathrm{CV}_i)||e_t(\mathrm{JD}_j)|} \; or \; |e_t(\mathrm{CV}_i) - e_t(\mathrm{JD}_j)|$$

However, it leads to information loss about the years of experience, e.g. a person having 1-year 'java' and 1-year 'SQL' experience would have the same skills vectors as that of having 10-year 'java' and 10-year 'SQL'. Therefore, an extra term could be added to account for experience level:

$$\overline{year}(\mathrm{CV}_i) = \frac{\sum_{year_i \in years_i} year_i}{|skills_i|}$$

Finally, a hybrid measure for similarity between CV-JD pair is given by:

$$sim_{d,t,s,y}(\mathrm{CV}_i, \mathrm{JD}_j) = \alpha sim_d(\mathrm{CV}_i, \mathrm{JD}_j) + \beta sim_t(\mathrm{CV}_i, \mathrm{JD}_j) + \gamma \left(sim_s(\mathrm{CV}_i, \mathrm{JD}_j) + \delta \overline{year}(\mathrm{CV}_i)\right)$$

where weight parameters α, β, γ and δ can be tuned according to the preference of matching, to consider more on the textual similarity between documents or the matching of skills and emphasis of experience.

Finally, for a CV_i, we may find its best matched JDs by:

$$\arg\min_{JD_j} \left(sim_{d,t,s,y}(\mathrm{CV}_i, \mathrm{JD}_j)\right)$$

Similarly, for a JD_i, we may find its best matched JDs by:

$$\arg\min_{CV_i}\left(sim_{d,t,s,y}\left(\mathrm{CV}_i, \mathrm{JD}_j\right)\right)$$

- **Suggestion of JD Based on Document Vectors (and Job Title Vectors)**

Apart from CV-JD matching, having documents in vectors provide handy tools for suggestion of JD. For example, it becomes easy to search for other JDs similar to a query JD.

- **Suggestion of CV Based on Document Vectors (and Job Title and Skill Vectors)**

Similarly, an employer may search for other suitable candidates once they find one. Similar candidates to a query CV may be returned upon request.

3.2.4. Experiments and Evaluations

In training a LDA model, over 100,000 documents including 20,000 JDs, 40,000 CVs (without latest work experience of latest job) and another part of the 40,000 CVs (work description of current job only), are fed to the model. The implementation by Gensim is used, with number of topics equal to 30 and learning of asymmetric prior from the corpus ($\alpha = 'auto'$).

word_0	word_1	word_2	word_3	word_4	word_5	word_6	word_7	word_8	word_9	...	coherence score
services	enterprise	development	solutions	infrastructure	technology	cloud	systems	architecture	management	...	-0.811097
web	application	data	developed	server	sql	net	javascript	asp_net	services	...	-0.959353
project	business	requirements	team	development	process	user	agile	design	functional	...	-0.967748
management	provide	support	ensure	manage	develop	business	processes	maintain	service	...	-0.996901
project	management	managed	projects	program	team	process	developed	provided	budget	...	-1.008459
java	application	developed	web	spring	services	involved	framework	data	xml	...	-1.035806
support	windows	server	software	network	users	user	servers	desktop	hardware	...	-1.037761
data	reports	application	developed	applications	oracle	systems	project	new	support	...	-1.041885
software	systems	hardware	equipment	network	maintenance	support	installation	troubleshooting	problems	...	-1.046228
servers	server	linux	environment	storage	windows	administration	configuration	systems	unix	...	-1.082731
team	business	skills	teams	technology	solutions	product	role	drive	help	...	-1.133721
data	sql	reports	business	created	developed	etl	involved	tables	informatica	...	-1.139412
test	testing	cases	scripts	automation	performed	qa	automated	defects	functional	...	-1.171782
microsoft	sharepoint	server	ms	sql	custom	web	site	created	windows	...	-1.196664
systems	skills	engineering	software	development	technical	requirements	design	related	team	...	-1.320835
business	sales	product	market	financial	marketing	new	customer	revenue	strategy	...	-1.336174
network	cisco	switches	routers	configured	security	troubleshooting	configuration	ip	firewalls	...	-1.356935
customer	support	customers	service	issues	technical	calls	provide	products	resolution	...	-1.359070
oracle	database	sql	databases	performance	data	server	production	tuning	pl	...	-1.359844
inventory	daily	maintained	assisted	information	orders	customers	customer	reports	provided	...	-1.364292
data	analysis	analytics	research	models	tools	reporting	statistical	machine_learning	business	...	-1.492977
health	care	staff	office	information	medical	state	department	services	duties	...	-1.496763
aws	cloud	build	python	jenkins	code	tools	applications	automation	deployment	...	-1.517550
design	engineering	test	manufacturing	control	product	production	equipment	electrical	flight	...	-1.532229
sales	new	company	training	team	responsible	clients	development	business	hr	...	-1.557334
training	web	created	content	developed	development	website	designed	students	learning	...	-1.592071
security	systems	information	support	program	operations	requirements	management	government	training	...	-1.598776
software	development	mobile	code	android	embedded	design	developed	devices	hardware	...	-1.620130
marketing	design	com	digital	content	media	product	web	online	website	...	-1.783620
network	center	data	new	voice	wireless	site	services	project	service	...	-1.896533

Table 18: LDA topic distribution after training on all documents

For the Doc2vec model, the same documents are supplied as the training corpus. The dimensions of document vector are set to 150, with context window of size 10 using PV-DM. Word vectors are simultaneously trained in the skip-gram fashion. In this report, the CV-JD matching results shown are only generated based on Doc2vec representations for simplicity.

As mentioned, it is not possible to perform rigorous evaluation on the matching results, as they are subjective. What is more is that even a person currently works for company A, it does not mean that company A suits his background (excluding company A's experience) most. Therefore, several evaluation methods are presented here so that we may analyse the performance of the matching algorithm both quantitatively and qualitatively.

- **Evaluation on Generated Test Data**

The generated test data serves to provide employment history so that evaluation can be performed on the modelling of document vectors.

There are over 40,000 test data pairs in the generated test data. To understand the correlation between similarities of vectors and actual employment history, the same number of CV pairs (see Appendix 6.5) are randomly sampled and the frequencies of similarity scores are plotted.

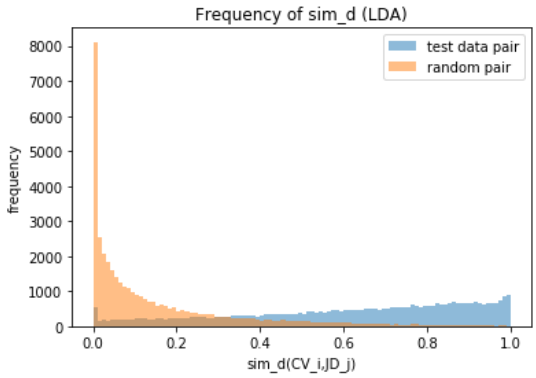
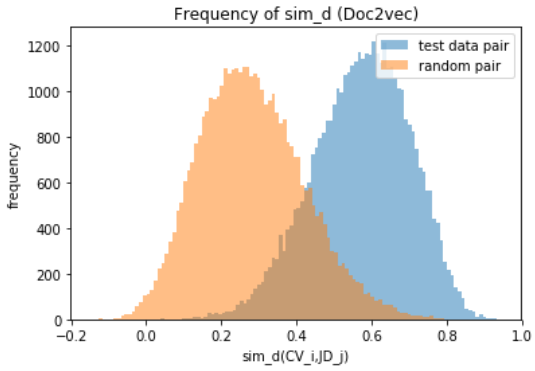
Using $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$	
	

Table 19: Plot of frequencies of $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$, where $e_d(\mathrm{CV}_i)$ and $e_d(\mathrm{JD}_i)$ are generated with LDA or Doc2vec respectively.

Table 19 shows that both document vectors similarities/differences measures operated on the document vectors generated by both LDA and Doc2vec $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$ are able to distinguish test data pairs from random CV pairs. Particularly, they are better separated by operating on LDA document vectors, due to the sparsity of LDA document vectors, i.e. many documents have a sharp distribution on topics and some topics' components are zero.

Using $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$

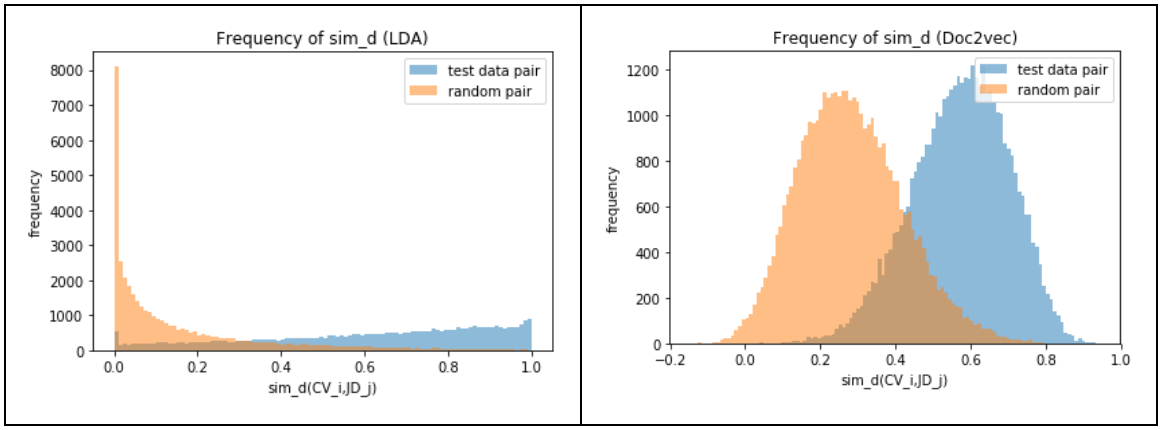


Table 19: Plot of frequencies of $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$, where $e_d(\mathrm{CV}_i)$ and $e_d(\mathrm{JD}_i)$ are generated with LDA or Doc2vec respectively.

Table 19 shows that both document vectors similarities/differences measures operated on the document vectors generated by both LDA and Doc2vec $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$ are able to distinguish test data pairs from random CV pairs. Particularly, they are better separated by operating on LDA document vectors, due to the sparsity of LDA document vectors, i.e. many documents have a sharp distribution on topics and some topics' components are zero.

Using $sim_{d,t,s,y}(\mathrm{CV}_i, \mathrm{JD}_j)$

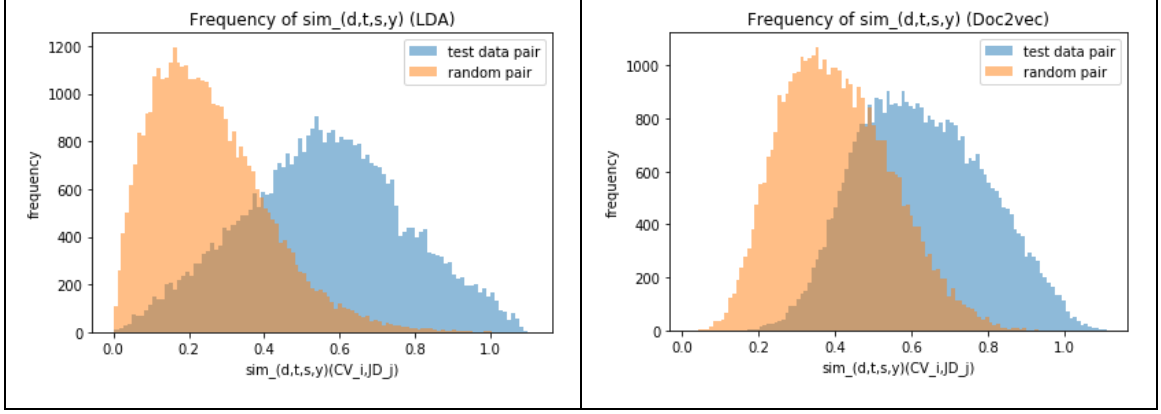


Table 20: Plot of frequencies of $sim_{d,t,s,y}(\mathrm{CV}_i, \mathrm{JD}_j)$, where all e_d, e_t and e_s are generated with LDA or Doc2vec, with $(\alpha, \beta, \gamma, \delta) = (0.5, 0.25, 0.25, 0.1)$.

Table 20 shows that $sim_{d,t,s,y}$ is still able to distinguish test data pairs from random ones, despite the noise from weighted-summing up random variables. However, the distribution of test data pairs and random pairs get closer, which means it is harder to distinguish good pairs from bad ones. One reason could be that for a CV, the job title of the second last job may have dissimilar vector representations from the title of the latest job, as one may not stay in the same position or field all the time. Nevertheless, it is still reasonable to match any JD of a title that is similar to that of the latest job for a candidate CV.

- **Evaluation on Matching a query CV with JDs**

To experiment with actual JDs, we tested matching one CV of a senior Android developer based on different similarity metrics as defined above. The results are as shown:

<table>
<tr><td colspan="2">Using $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$</td></tr>
<tr><td>Query CV</td><td>Returned JDs</td></tr>
<tr><td rowspan="3">Senior Android Developer

Description: Amazon is an American electronic commerce and cloud computing company based in Seattle, Washington. Worked with a team to develop an application which could track the components involved in manufacturing the product by scanning the QR code used on a product. The application provides the user details about participating product's origins, including the manufacturing date and location Every item with a Transparency label includes a unique code that can be used to track product information. This helps the users understand the authenticity of the product they purchased. App Link: https://play.google.com/store/apps/details?id=com.amazon.aba.application&hl=en_US Responsibilities: ● Part of the team which has developed and deployed the Amazon transparency application for the android platform. ● Worked with the team through all phases of Software Development from requirement analysis to deployment of the application. Currently working on adding enhancements and scaling the application horizontally. ● The application is developed using the MVC architecture in order to make the application more flexible and easier for enhancements. ● Details of the products has been provided to the user by utilizing several layouts and UI components such as Dialog Boxes, Gallery, Spinner, ImageSwitcher, Action Bar, Navigation Control and Alert boxes. ● ImageSwitcher was used to traverse between the components and material design was used to provide Floating Action Button, App Bar and Navigation Drawer. ● Handled the data storage and requests locally in the application using SQLite and in the back-end using PostgreSQL. ● User login credentials is secured to login into the amazon.com portal using the Amazon SDK for Android. ● Details of the product components are stored in the AWS RDS and PostgreSQL was used to perform operations on the data available in the database. ● Amazon SNS (Simple Notification Service) is being used to send out push notifications generated in the application. ● Unit testing is done through Robotium and Genymotion emulator has been used to test the application. ● Application was developed efficiently using Agile and SCRUM methodology with bi-weekly sprints. Environment: Android Studio, Android SDK, SQLite, Realm, JDK, Eclipse IDE, JavaScript, RESTful services, Logcat, Gradle and GIT.

(previous work experience) …

Skills:
java – 4 years
mysql – 4 years
android – 2 years
eclipse – 2 years
testing – 2 years</td><td>Software Engineer, Mobile, Android NDK @ Facebook

(company's description) …
RESPONSIBILITIES
Build client side mobile infrastructure in C++
Debug app wide issues across Java and C++
Improve app wide performance and efficiency
Develop tools for debugging, instrumenting and shipping native code on Android
…</td></tr>
<tr><td>Instagram - Software Engineer, Android (Stories) @ Facebook

(company's description) …
RESPONSIBILITIES
Work closely with our product and design teams to build and maintain features for the Stories Product.
Build reusable Android Software components to support the stories format through Instagram.
Collaborate with cross-functional teams such as design and data to improve the Stories experience.
MINIMUM QUALIFICATIONS
Knowledge of object-oriented software development
Experience of building Android applications in Java using Android SDK
Mobile application development experience at the user interface and system levels
…</td></tr>
<tr><td>Android Software Engineer @ Facebook

(company's description) …
RESPONSIBILITIES
Work closely with our product team to customize the Portal experience for the Android platform
Prototype new and redesign features
Contribute best-in-class programming skills to develop highly innovative, consumer-facing hardware products
MINIMUM QUALIFICATIONS
2+ years of object-oriented software development experience
Experience in understanding code bases, including API design techniques
Experience with Java language and related frameworks
Experience with caching mechanisms
Coding experience with Java and Android SDK
Knowledge in UI design principles and making apps work
PREFERRED QUALIFICATIONS
2+ years of experience building Android applications in Java using Android SDK</td></tr>
</table>

Table 21: top 3 CVs retrieved using $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$

This simple measure takes only document content into account, but it is enough for an Android developer to be matched with Android software engineer jobs.

<table>
<tr><td colspan="2">Using $sim_s(\mathbf{CV}_i, \mathbf{JD}_j)$</td></tr>
<tr><td>Query CV</td><td>Returned JDs</td></tr>
<tr><td rowspan="4">Senior Android Developer

Description: Amazon is an American electronic commerce and cloud computing company based in Seattle, Washington. Worked with a team to develop an application which could track the components involved in manufacturing the product by scanning the QR code used on a product. The application provides the user details about participating product's origins, including the manufacturing date and location Every item with a Transparency label includes a unique code that can be used to track product information. This helps the users understand the authenticity of the product they purchased. App Link: https://play.google.com/store/apps/details?id=com.amazon.aba.application&hl=en_US
Responsibilities: ● Part of the team which has developed and deployed the Amazon transparency application for the android platform. ● Worked with the team through all phases of Software Development from requirement analysis to deployment of the application. Currently working on adding enhancements and scaling the application horizontally. ● The application is developed using the MVC architecture in order to make the application more flexible and easier for enhancements. ● Details of the products has been provided to the user by utilizing several layouts and UI components such as Dialog Boxes, Gallery, Spinner, ImageSwitcher, Action Bar, Navigation Control and Alert boxes. ● ImageSwitcher was used to traverse between the components and material design was used to provide Floating Action Button, App Bar and Navigation Drawer. ● Handled the data storage and requests locally in the application using SQLite and in the back-end using PostgreSQL. ● User login credentials is secured to login into the amazon.com portal using the Amazon SDK for Android. ● Details of the product components are stored in the AWS RDS and PostgreSQL was used to perform operations on the data available in the database. ● Amazon SNS (Simple Notification Service) is being used to send out push notifications generated in the application. ● Unit testing is done through Robotium and Genymotion emulator has been used to test the application. ● Application was developed efficiently using Agile and SCRUM methodology with bi-weekly sprints. Environment: Android Studio, Android SDK, SQLite, Realm, JDK,</td><td>Android NDK Engineer @ Facebook

(company’s description) …
RESPONSIBILITIES
Build client side mobile infrastructure in C++
Debug app wide issues across Java and C++
Improve app wide performance and efficiency
Develop tools for debugging, instrumenting and shipping native code on Android
MINIMUM QUALIFICATIONS
B.S. or M.S. Computer Science
2+ years of object-oriented software development experience
2+ years experience in building infrastructure in C++ using the Android NDK
2+ years mobile application development experience (Android or iOS)
Experience in understanding code bases, including API design techniques
Experience in the following technologies:
Experience with C++ language and related frameworks
Experience with Multi-Threading and memory management specific to mobile devices
Experience debugging system issues across languages</td></tr>
<tr><td>Full Stack Developer @ General Dynamics Information Technology

Required Skills: Full stack developer with documented experience developing web-based applications with JAVA, Spring and MySQL. Experienced developing and consuming REST and SOAP web services. Desired Skills: Experience with common JavaScript frameworks, Amazon C2S and cross-domain solutions.
1. Provides application development and technical support for customer websites.
2. Collaborates with graphic artists to develop web page graphics that support interactive, marketing-focused content.
3. Provides technical consultation in new systems development, new package evaluations and enhancements of existing systems.
4. Prepares functional specifications from which programs will be written, then designs, codes, tests, debugs and documents programs.
5. Participates in the technical design, development, testing, implementation and maintenance of website enhancements.
6. Plans, schedules and conducts systems tests, monitors test results, and takes appropriate corrective action.
7. Provides guidance and work leadership to less-experienced staff, and may have supervisory responsibilities.
8. May serve as a technical team or task leader.
9. Maintains current knowledge of relevant technology as assigned.
10. Participates in special projects as required.</td></tr>
<tr><td>Sr. Web Developer- Amazon, Web w/Polygraph @ General Dynamics Information Technology

Develops, codes, deploys and maintains web applications and websites.
Required Skills: Full stack developer with documented experience developing web-based applications with JAVA, Spring and MySQL. Experienced developing and consuming REST and SOAP web services. Desired Skills: Experience with common JavaScript frameworks, Amazon C2S and cross-domain solutions.</td></tr>
</table>

Eclipse IDE, JavaScript, RESTful services, Logcat, Gradle and GIT. (previous work experience) … Skills: java – 4 years mysql – 4 years android – 2 years eclipse – 2 years testing – 2 years	1. Provides advanced application development and technical support for customer websites. 2. May work in a consulting capacity across multiple tasks or contracts. 3. Collaborates with graphic artists to develop web page graphics that support interactive, marketing-focused content. 4. Provides technical consultation in new systems development, new package evaluations and enhancements of existing systems. 5. Prepares functional specifications from which programs will be written, then designs, codes, tests, debugs and documents programs. 6. Participates in the technical design, development, testing, implementation and maintenance of website enhancements. 7. Oversees systems tests and monitors test results and corrective actions. 8. Provides guidance and work leadership to less-experienced staff, and may have supervisory responsibilities. 9. May serve as a technical team or task leader. 10. Maintains current knowledge of relevant technology as assigned. 11. Participates in special projects as required.

Table 22: top 3 JDs retrieved using $sim_s(\mathrm{CV}_i, \mathrm{JD}_j)$

sim_s takes only skills information of a candidate CV into account. Despite having far less tokens than a full document, the skills vectors indeed capture the skills of the candidate and they are matched with reasonable JDs.

Using $sim_{d,t,s,y}(\mathbf{CV}_i, \mathbf{JD}_j)$	
Query CV	Returned JDs
Senior Android Developer Description: Amazon is an American electronic commerce and cloud computing company based in Seattle, Washington. Worked with a team to develop an application which could track the components involved in manufacturing the product by scanning the QR code used on a product. The application provides the user details about participating product's origins, including the manufacturing date and location Every item with a Transparency label includes a unique code that can be used to track product information. This helps the users understand the authenticity of the product they purchased. App Link: https://play.google.com/store/apps/details?id=com.amazon.aba.application&hl=en_US Responsibilities: ● Part of the team which has developed and deployed the Amazon transparency application for the android platform. ● Worked with the team through all phases of Software Development from requirement analysis to deployment of the application. Currently working on adding enhancements and scaling the application horizontally. ● The application is developed using the MVC architecture in order to make the application more flexible and easier for	**Software Engineer, Mobile, Android NDK @ Facebook** Facebook's mission is to give people the power to build community and bring the world closer together. Through our family of apps and services, we're building a different kind of company that connects billions of people around the world, gives them ways to share what matters most to them, and helps bring people closer together. Whether we're creating new products or helping a small business expand its reach, people at Facebook are builders at heart. Our global teams are constantly iterating, solving problems, and working together to empower people around the world to build community and connect in meaningful ways. Together, we can help people build stronger communities — we're just getting started. Every month, more than 1.57 billion people access Facebook using mobile devices from across the world. As our mobile user base grows, we're looking at ways to scale our engineering teams to continue delivering new and innovative products while still moving fast. We're looking for mobile system engineers who are passionate about working on client infrastructure that powers all of the products we build across Android and iOS. You will get an opportunity to influence the way we build mobile apps at Facebook. In this role, you will build core Android infrastructure in C++, build awesome new tools that will lead to better efficiency and work on complex system investigations throughout our mobile stack. RESPONSIBILITIES Build client side mobile infrastructure in C++ Debug app wide issues across Java and C++ Improve app wide performance and efficiency Develop tools for debugging, instrumenting and shipping native code on Android MINIMUM QUALIFICATIONS B.S. or M.S. Computer Science or 2+ years in software development experience 2+ years of object-oriented software development experience 2+ years experience building infrastructure in C++ using the Android NDK 2+ years mobile application development experience (Android or iOS)

enhancements. ● Details of the products has been provided to the user by utilizing several layouts and UI components such as Dialog Boxes, Gallery, Spinner, ImageSwitcher, Action Bar, Navigation Control and Alert boxes. ● ImageSwitcher was used to traverse between the components and material design was used to provide Floating Action Button, App Bar and Navigation Drawer. ● Handled the data storage and requests locally in the application using SQLite and in the back-end using PostgreSQL. ● User login credentials is secured to login into the amazon.com portal using the Amazon SDK for Android. ● Details of the product components are stored in the AWS RDS and PostgreSQL was used to perform operations on the data available in the database. ● Amazon SNS (Simple Notification Service) is being used to send out push notifications generated in the application. ● Unit testing is done through Robotium and Genymotion emulator has been used to test the application. ● Application was developed efficiently using Agile and SCRUM methodology with bi-weekly sprints. Environment: Android Studio, Android SDK, SQLite, Realm, JDK, Eclipse IDE, JavaScript, RESTful services, Logcat, Gradle and GIT. (previous work experience) … Skills: java – 4 years mysql – 4 years android – 2 years eclipse – 2 years testing – 2 years	Experience in understanding code bases, including API design techniques Experience in the following technologies: Experience with C++ language and related frameworks Experience with Multi-Threading and memory management specific to mobile devices Experience debugging system issues across languages **Principal Android Developer @ AT&T** AT&T Entertainment Group provides world class digital delivery of entertainment and sports content. Our engineers are constantly developing and enhancing our leading edge software solutions, delivering an industry-leading customer experience over satellite, mobile, video on demand, and interactive services. We are constantly looking for ways to reduce roadblocks for our engineers so they can do what they do best: deliver world-class entertainment products. We are looking for a Principal Android Engineer. This role will be responsible for fundamental development work of designing/developing/maintaining the flagship android apps for our group and over all platform. This person will also own end to end Android solutions and important KPIs. Requirements #LI-CL1 BS in Computer Science and/or MS in Computer Science degree. Good understanding of CICD 5+ years of Android Application development experience. Strong Java/Kotlin programming language knowledge. Familiarity with Android Application development model. Familiarity with Android Application Architecture, such as MVVM. Familiarity with Android Unit test framework, such as Espresso. Familiarity with SCM tools, such as Git, SVN, Jenkins, etc. Strong problem solving, and analytical skills. Detail oriented, quick study, and self-motivated Ability to work independently in a fast paced, deadline driven environment. Ability to multitask within an environment of rapidly changing priorities. Strong team player and passion for technology Excellent written/verbal communication skills. **Full Stack Developer @ General Dynamics Information Technology** Required Skills: Full stack developer with documented experience developing web-based applications with JAVA, Spring and MySQL. Experienced developing and consuming REST and SOAP web services. Desired Skills: Experience with common JavaScript frameworks, Amazon C2S and cross-domain solutions. 1. Provides application development and technical support for customer websites. 2. Collaborates with graphic artists to develop web page graphics that support interactive, marketing-focused content. 3. Provides technical consultation in new systems development, new package evaluations and enhancements of existing systems. 4. Prepares functional specifications from which programs will be written, then designs, codes, tests, debugs and documents programs. 5. Participates in the technical design, development, testing, implementation and maintenance of website enhancements. 6. Plans, schedules and conducts systems tests, monitors test results, and takes appropriate corrective action. 7. Provides guidance and work leadership to less-experienced staff, and may have supervisory responsibilities. 8. May serve as a technical team or task leader. 9. Maintains current knowledge of relevant technology as assigned. 10. Participates in special projects as required.

Table 23: top 3 JDs retrieved using $sim_{d,t,s,y}(\mathrm{CV}_i, \mathrm{JD}_j)$

Using the hybrid approach, we are able to match the CV of a Senior Android Developer to some more senior positions, which we failed to retrieve using the first two measures. Using parameters setting of $(\alpha, \beta, \gamma, \delta) = (0.4, 0.4, 0.2, 0.1)$, we can retrieve the position of Principal Android Developer as the second JD. Furthermore, the other two positions are obtained consistently as well.

- **Evaluation on Matching a query JD with CVs**

<table>
<tr><td colspan="2">Using $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$</td></tr>
<tr><td>Query JD</td><td>Returned CVs</td></tr>
<tr><td rowspan="2">Sr. Business Intelligence Data Engineer @ Amazon.com Services, Inc.

Bachelor's degree in computer science, engineering, mathematics, or a related technical discipline
5+ years of industry experience in software development, data engineering, business intelligence, data science, or related field with a track record of manipulating, processing, and extracting value from large datasets
Demonstrated strength in data modeling, ETL development, and data warehousing
Solid experience in RDBMS including Redshift, PostgreSQL, and MySQL
Proven skillset in coding automation solutions using Python/Perl/Java
Experience working with AWS solutions – Redshift, S3, DynamoDB, Lambda, Aurora
Experience supporting business intelligence reporting tools - Tableau / QuickSight a plus
Experience using RESTful/OData API services
Knowledge of data management fundamentals and data storage principles
Knowledge of distributed systems as it pertains to data storage and computing
Understand business processes, logical data models and relational database implementations for data analysis
Highly motivated, self-driven, capable of defining own design and test scenarios

Do you want to build a cutting-edge highly scalable data platform and automation solutions using AWS technologies?

We are looking for an experienced, self-driven, analytical, and strategic Data Engineer. In this role, you will be working in a large and complex data warehouse environment. You should be passionate about working with disparate datasets and be someone who loves to bring data together to answer business questions. You should have deep expertise in the creation and management of data automation</td><td>Data Analyst

Key Skills: SQL Data & Relational Modeling Statistical Analysis Excel ETL & SSIS Developed critical metrics and actionable insights on the rapidly changing software base to action the $8.8B split to Micro Focus. Collaborated with teams to power business decisions around complex software challenges. Navigated large data-sets, built and interpreted predictive models, and partnered with IT teams managing applications to understand IT demand signals.

Key Skills: SQL Data Modeling Advanced Analytics Data Cleansing/Transformation ETL Intimately familiarized with software development life-cycle methodology and interpretation, and presented findings through both verbal and written formats. Prepared detailed documents and reports with willingness to question the validity and accuracy of data. Demonstrated ability to multitask and prioritize diverse tasks with a proven ability to meet hard deadlines.

Skills:
sql – 4 years
data analysis – 4 years
excel – 4 years
data managerment – 4 years
reporting – 4 years
sharepoint – 0.5 years</td></tr>
<tr><td>Senior ETL Developer (SAP and Hadoop)

Highly skilled and driven developer with the ability to devise innovative solutions, identify risks, and support new product functionality. Skilled troubleshooter with excellent project management skills; adept at accomplishing milestones while meeting requirements and deadlines. Strong knowledge and comfort working with querying languages, SAP HANA, SAP BO Data Services, and SAP BW. Able to understand, debug and develop unique solutions and provide excellent customer service. Articulate and analytical; excel in both independent and collaborative environments. Valuable team member with extensive SQL knowledge and a strong track record of success.

Earned advancement through several promotions, culminating in present senior programmer analyst role overseeing and delivering enterprise solutions. Manage a 5 member team with developers residing on-site, Mexico, and India.

Skills:
.net – 2 years
apache hadoop ambari – 1 year</td></tr>
</table>

and the proven ability to translate data into meaningful insights through collaboration with analysts and engineers. In this role, you will share ownership of end-to-end development of data engineering solutions to answer complex questions and you'll play an integral role in strategic decision-making. The right candidate will possess excellent business and communication skills, be able to work with business owners to develop and define key business questions, be able to automate data ingestion at scale, onboard and integrate new and existing datasets, and be able to collaborate to analyze data that will answer those questions. In this role, you will have the opportunity to display and develop your skills in the following areas: Manage and automate enterprise scale data warehouse solution within AWS. Create and manage python-based automation solutions using AWS technologies. Prioritize and deliver on ad hoc data automation projects. Troubleshoot and support new and existing coding solutions on Development, Beta and Production platforms following industry best practices. Collect, analyze and present actionable data insights to drive operational support and logistics decisions	sap hana – 4 years sap data services – 4 years sql – 5 years sap bw 7.3-7.5 2 years backoffice associates (cransoft) – 0.5 years git – 1 year redwood cronacle – 2 years agile development – 2 years **Technology Analyst** • Professionally trained in all aspects of software engineering, including the use of Python, SQL, and Java for process automation, software design, information/database management, and enterprise application development • Consults with cross-functional teams to collect and synthesize information from various data sources in order to identify and solve technical issues and mitigate potential risks that could impact critical financial processes • Uses SQL, Unix, Excel, and Python to manage databases, generate business reports, and conduct ad hoc analyses Computer Systems Analyst 2016-2017 - IAM analyst for aerospace sector customers and stakeholders - Provision access following SOD principles for various systems - assist in process improvement and process redesign efforts Skills: java – 2 years linux – 0.5 years python – 2 years sql – 1 year unix – 1 year javascript – 0.5 years

Table 24: top 3 CVs retrieved using $sim_d(\mathrm{CV}_i, \mathrm{JD}_j)$

As shown in table 24, the retrieved CVs are generally of the same field. Although the third returned CV possesses the skills required by the query JD, it only has an average of 1 year of experience among his skills, which there is not very suitable to the query JD position of Sr. Business Intelligence Data Engineer. The hybrid approach can be applied to improve on the results.

Using $sim_s(\mathbf{CV}_i, \mathbf{JD}_j)$	
Query JD	Returned CVs
Sr. Business Intelligence Data Engineer @ Amazon.com Services, Inc. Bachelor's degree in computer science, engineering, mathematics, or a related technical discipline 5+ years of industry experience in software development, data engineering, business intelligence, data science, or related field with a track record of manipulating, processing, and extracting value from large datasets Demonstrated strength in data modeling, ETL development, and data warehousing	**Data Scientist/ Data Science Consultant** Price optimization engine: Designed a price optimization engine which enable to better price, refresh products for customers in different programs. Utilized Non-linear optimization technique (NLOPT) for maximizing revenue while satisfying the constraints. Performed clusters analysis on PID's that covers 80% of total revenue. • Developed a 3 month demand forecast model using ARIMA time-series technique. Performed demand and discount variation analysis across different channels and conducted exploratory data analysis on variables for building model constraints using python. • Built RF and MFG Price correlation analysis, Segment based pricing analysis for demand and inventory mapping. Developed Optimization Models for clusters covering PIDs which contribute to 68% of the total revenue. Expected an incremental revenue of 15% to 20% (~5M). • Build to Max Forecast: Built an analytical model to forecast Build to Max for each product for different theatres. Provided recommendations for building optimal safety stock in order to

Solid experience in RDBMS including Redshift, PostgreSQL, and MySQL Proven skillset in coding automation solutions using Python/Perl/Java Experience working with AWS solutions – Redshift, S3, DynamoDB, Lambda, Aurora Experience supporting business intelligence reporting tools - Tableau / QuickSight a plus Experience using RESTful/OData API services Knowledge of data management fundamentals and data storage principles Knowledge of distributed systems as it pertains to data storage and computing Understand business processes, logical data models and relational database implementations for data analysis Highly motivated, self-driven, capable of defining own design and test scenarios Do you want to build a cutting-edge highly scalable data platform and automation solutions using AWS technologies? We are looking for an experienced, self-driven, analytical, and strategic Data Engineer. In this role, you will be working in a large and complex data warehouse environment. You should be passionate about working with disparate datasets and be someone who loves to bring data together to answer business questions. You should have deep expertise in the creation and management of data automation and the proven ability to translate data into meaningful insights through collaboration with analysts and engineers. In this role, you will share ownership of end-to-end development of data engineering solutions to answer complex questions and you'll play an integral role in strategic decision-making. The right candidate will possess excellent business and communication skills, be able to work with business owners to develop and define key business questions, be able to automate data ingestion at scale, onboard and integrate new and existing datasets, and be able to collaborate to analyze data that will answer those questions. In this role, you will have the opportunity to display and develop your skills in the following areas:	avoid overbuild or under build situations. This model reduced overbuilding of finished goods inventory (FGI) to about 40% compared to as-is model. • Blended lead times for damaged goods processing at different repair sites with demand variability for calculating safety stock. Automated the model predictions for safety stock calculations per business requirements. • Predominantly used Python, PySpark, HDFS, Sqoop, Hive (HQL) and Oracle SQL developer in providing insights to the business users. (previous work experience) … Skills: hive – 1 year machine learning – 3 years python – 2 years sql – 1 year testing – 1 year **Sr. Data Scientist** Responsibilities: • Responsible for performing Machine-learning techniques regression/classification to predict the outcomes. • Performed Ad-hoc reporting/customer profiling, segmentation using R/Python. • Tracked various campaigns, generating customer profiling analysis and data manipulation. • Provided R/SQL programming, with detailed direction, in the execution of data analysis that contributed to the final project deliverables. Responsible for data mining. • Utilized Label Encoders in Python to convert non-numerical significant variables to numerical significant variables to identify their impact on pre-acquisition and post acquisitions by using 2 sample paired t test. • Worked with ETLSQL Server Integration Services (SSIS) for data investigation and mapping to extract data and applied fast parsing and enhanced efficiency by 17%. • Developed Data Science content involving Data Manipulation and Visualization, Web Scraping, Machine Learning, Python programming, SQL, GIT and ETL for DataExtraction. … (previous work experience) … Skills: etl – 6 years extract – 6 years transform – 6 years and load – 6 years hadoop – 6 years machine learning – 6 years sql – 7 years **Data Scientist** • Own TCS Growth Hacking engagement for Microsoft Azure - measure customer upsell, cross-sell, conversion, and retention via controlled experiments in email marketing • Create metrics/business logic, measure results using probability models and confidence intervals and provide business recommendations • Analysis of customer usage trends/behavior/demographics with exploratory data analysis and inference using R; generate and evaluate testable hypotheses for experimentation • Clustering analysis of customer usage/spend using unsupervised method (k-means) leading to segmentation of customer base used for targeting efforts • Develop automated outcome reporting tool atop SQL databases using statistical functions in R, saving hours of processing time per experiment • Develop web-hosted R Shiny application for managing experiments with user-friendly UI/UX, leading to scaled experimentation on Azure products • Manage and onboard new team members to Growth Hacking team; currently managing two BI developer/analysts • Develop Power BI dashboards to visualize experiment characteristics, metrics, outcomes - currently used by hundreds of stakeholders to find successful marketing techniques for implementation • Use and manage Azure tools - SQL Data Warehouse, storage • Create SSIS packages to automate data load to and between

Manage and automate enterprise scale data warehouse solution within AWS. Create and manage python-based automation solutions using AWS technologies. Prioritize and deliver on ad hoc data automation projects. Troubleshoot and support new and existing coding solutions on Development, Beta and Production platforms following industry best practices. Collect, analyze and present actionable data insights to drive operational support and logistics decisions	databases • Train Azure product teams in experimentation methodology • Teach statistics, R, and SQL to both colleagues and clients, leading to larger base of technical proficiency in team (previous work experience) … Skills: data analysis – 0.5 years sql – 0.5 years data mining – 0.5 years data science – 0.5 years data visualization – 0.5 years python – 0.5 years machine learning – 0.5 years hadoop – 0.5 years r – 0.5 years statistical analysis – 7 years

Table 25: top 3 CVs retrieved using $sim_s(\mathrm{CV}_i, \mathrm{JD}_j)$

Table 25 shows the matching results of CVs' skills vectors to a query JD document vector. Similarly, the first and the third CVs returned have most of their skills of less than 3 years, which is unsuitable as compared to a senior position.

Using $sim_{d,t,s,y}(\mathbf{CV}_i, \mathbf{JD}_j)$	
Query JD	Returned CVs
Sr. Business Intelligence Data Engineer @ Amazon.com Services, Inc. Bachelor's degree in computer science, engineering, mathematics, or a related technical discipline 5+ years of industry experience in software development, data engineering, business intelligence, data science, or related field with a track record of manipulating, processing, and extracting value from large datasets Demonstrated strength in data modeling, ETL development, and data warehousing Solid experience in RDBMS including Redshift, PostgreSQL, and MySQL Proven skillset in coding automation solutions using Python/Perl/Java Experience working with AWS solutions – Redshift, S3, DynamoDB, Lambda, Aurora Experience supporting business intelligence reporting tools - Tableau / QuickSight a plus Experience using RESTful/OData API services Knowledge of data management fundamentals and data storage principles	**Sr. Architect** Drive the success of Business Intelligence Center of Excellence (BICoE) initiatives focused on developing, implementing, and supporting new solutions in Automation, Big Data, Data Warehousing, Data Analytics, and Tier-0 On-Premises and Cloud Infrastructures. • Develop and present technical roadmaps, reference architectures, and cost benefit analyses to customers and C-level management stakeholders. FANNIE MAE USA - Sr. Architect - Enterprise Data ~~~~~~~~~~) • Improve the performance of BI and Big Data platforms, applying strong DevOps practices throughout the development of new enterprise systems. • Deliver training programs and speaking engagements in the areas of BI, Data Warehouse, ETL, Data Preparation and Integration. • Deliver customer-centric BI solutions and technical roadmaps by collaborating with customers to evaluate business needs and present appropriate tools based on capabilities, costing, and available options. • Identify opportunities to design and implement transformative Data Warehousing platforms using strong, hands- on skills in programming, scripting, infrastructure development, and ongoing support of new data management environments. • Lead the planning and completion of cloud-first implementations, including architecting and deploying MicroStrategy and Tableau on AWS Cloud for hybrid platforms. … (previous work experience) … Skills: aws – 10+ years bi – 10+ years business intelligence – 10+ years data warehousing – 10+ years solutions – 10+ years
	Data Engineer Analyst

Knowledge of distributed systems as it pertains to data storage and computing Understand business processes, logical data models and relational database implementations for data analysis Highly motivated, self-driven, capable of defining own design and test scenarios Do you want to build a cutting-edge highly scalable data platform and automation solutions using AWS technologies? We are looking for an experienced, self-driven, analytical, and strategic Data Engineer. In this role, you will be working in a large and complex data warehouse environment. You should be passionate about working with disparate datasets and be someone who loves to bring data together to answer business questions. You should have deep expertise in the creation and management of data automation and the proven ability to translate data into meaningful insights through collaboration with analysts and engineers. In this role, you will share ownership of end-to-end development of data engineering solutions to answer complex questions and you'll play an integral role in strategic decision-making. The right candidate will possess excellent business and communication skills, be able to work with business owners to develop and define key business questions, be able to automate data ingestion at scale, onboard and integrate new and existing datasets, and be able to collaborate to analyze data that will answer those questions. In this role, you will have the opportunity to display and develop your skills in the following areas: Manage and automate enterprise scale data warehouse solution within AWS. Create and manage python-based automation solutions using AWS technologies. Prioritize and deliver on ad hoc data automation projects. Troubleshoot and support new and existing coding solutions on Development, Beta and Production platforms following industry best practices.	• Data mapping, modeling, and integration • Analysis, Design, Development, Testing, Implementation, and Maintenance of internal programs/processes • Effectively communicate/present solutions to various stakeholders • Extensive use of Oracle10g/11g, TOAD, SAS Enterprise Guide 6.1, Teradata 14.10, Teradata Studio, MS Excel • Manipulate data to extract desired information • Collaborate with various stakeholders on tests and report set-up required to effectively measure results • Develop analyses and reports that facilitate data-driven decisions, both ad hoc and ongoing • Identify, understand, interpret, and communicate trends/patterns/other findings • Extensive use of MySQL, TOAD, Salesforce, MS Excel • Interrogate data from an array of sources and providers • Analysis, development, and production of transactional reports to understand and provide insight • Prepare specific program deliverables (presentations, reports, graphs, etc.) utilizing database software tools • Drive and support measurement objectives - develop and implement requirements, templates and deliverables • Support development and segmentation of targeting efforts • Extensive use of SAS, Oracle9i/10g (SQL, PL/SQL), SQL Server, Netezza, TOAD, MS Excel, MS PowerPoint • Analysis, Design, Development, Testing, Implementation, and Maintenance of internal programs/processes • Ad Hoc analysis, reporting, troubleshooting • Extensive use of SAS Enterprise Guide 4.1, Oracle 9i, SQL Server, MS Access, MS Excel (previous work experience) … Skills: sql – 10+ years excel – 10+ years power pivot – 10+ years power query – 10+ years power view – 10+ years dax – 10+ years power bi desktop – 10+ years oracle – 10+ years sql server – 10+ years mysql – 10+ years teradata – 10+ years netezza – 10+ years sas – 10+ years sas enterprise guide – 10+ years business intelligence – 0.5 years
	Sr. Big data Engineer Responsibilities: • Responsible for building scalable distributed data solutions using Big Data technologies like Apache Hadoop, MapReduce, Shell Scripting, Hive. • Used Agile (SCRUM) methodologies for Software Development. • Wrote complex Hive queries to extract data from heterogeneous sources (Data Lake) and persist the data into HDFS. • Involved in all phases of data mining, data collection, data cleaning, developing models, validation and visualization. • Designed and develop end to end ETL processing from Oracle to AWS using Amazon S3, EMR, and Spark. • Developed the code to perform Data extractions from Oracle Database and load it into AWS platform using AWS Data Pipeline. • Installed and configured Hadoop ecosystem like HBase, Flume, Pig and Sqoop. • Designed and develop Big Data analytic solutions on a Hadoop-based platform and engage clients in technical discussions. • Developed workflow in Oozie to automate the tasks of loading the data into HDFS and pre-processing with Pig. • Implemented AWS cloud computing platform using S3, RDS, Dynamo DB, Redshift, and Python. • Responsible in loading and transforming huge sets of structured, semi structured and unstructured data. …

Collect, analyze and present actionable data insights to drive operational support and logistics decisions	Skills sql – 10+ years hbase – 10+ years hadoop – 10+ years

Table 26: top 3 CVs retrieved using $sim_{d,t,s,y}(\mathrm{CV}_i, \mathrm{JD}_j)$

In this matching, $(\alpha, \beta, \gamma, \delta) = (0.4, 0.2, 0.4, 0.1)$, i.e. the emphasis on this matching is document similarity and skills matching, but less on the job title similarity. Setting $\delta = 0.1$ increases the importance of years of experience in different skills. As shown in table 26, We may notice that all of the top 3 retrieved CVs possess over 10 years of experience in all/most of the skills relevant to the query JD.

3.2.5. Discussions

The results of CV-JD matching is subjective, but we can still evaluate the matching quality by different metrics. Evaluation results show that vectorizing document contents and also the information of job titles and skills provides comprehensive matching based on a hybrid metric. Moreover, new token knowledge is discovered. For example, the token 'Sr.' and 'Senior' shows to have a highly similar token embeddings after training of both LDA and Doc2vec models on the CV-JD corpus. This is extremely useful for disambiguating among senses of words, but also provide the CV-JD scenario with more information and make quantification of skills and job titles possible.

4. CONCLUSION AND FUTURE WORK

A reliable job matching platform brings automation to recruitment process. Within the system, the matching algorithm is the most important component, and yet is the most challenging part because it is subjective to both companies and applicants whether they suit each other and preferences are not fully reflected from text documents. Although preferences of companies and applicants may be inferred based on employment/recruitment history by recommendation systems, when there is inadequate data or data is sparse, natural language processing is cruicial for extracting information of job descriptions and CVs.

First, the problem of automatic taxonomy induction from text corpora is researched, and an original algorithm is devised with consistent performance over multiple taxonomy settings. Secondly, a comprehensive document matching strategy is developed for matching CVs and JDs against each other. Utilizing vector models, the matching algorithm is able to identify suitable candidates/jobs for a query based on the preference of a user.

For the problem of taxonomy generation, one of the most fundamental problem here is to define “good” taxonomies. Besides, pair-wise classification of hypernyms, domain specific knowledge, disambiguation of senses in a term and machine learning algorithms for the search of optimal graph structure are potential areas to be researched on to improve the quality of induced taxonomies.

Furthermore, there is still a lot of rooms of improvement on CV-JD matching problem. They include but are not limited to the application of taxonomy, inclusion of more features such as education background, locality and company preferences, and collaborative recommendation based on employment histories.

5. REFERENCE

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[2] Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014) 12.

[3] Wang, C.; He, X.; and Zhou, A. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In EMNLP, 1190–1203. ACL

[4] Yang, H. and Callan, J., 2009, August. A metric-based framework for automatic taxonomy induction. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (pp. 271-279). Association for Computational Linguistics.

[5] Bansal, M., Burkett, D., De Melo, G. and Klein, D., 2014. Structured learning for taxonomy induction with belief propagation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1041-1051).

[6] Mao, Y., Ren, X., Shen, J., Gu, X. and Han, J., 2018. End-to-End Reinforcement Learning for Automatic Taxonomy Induction. arXiv preprint arXiv:1805.04044.

[7] Hearst, M.A., 1992, August. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics-Volume 2 (pp. 539-545). Association for Computational Linguistics.

[8] Shwartz, V., Goldberg, Y. and Dagan, I., 2016. Improving hypernymy detection with an integrated path-based and distributional method. arXiv preprint arXiv:1603.06076.

[9] Navigli, R., Velardi, P. and Faralli, S., 2011, July. A graph-based algorithm for inducing lexical taxonomies from scratch. In IJCAI (Vol. 11, pp. 1872-1877).

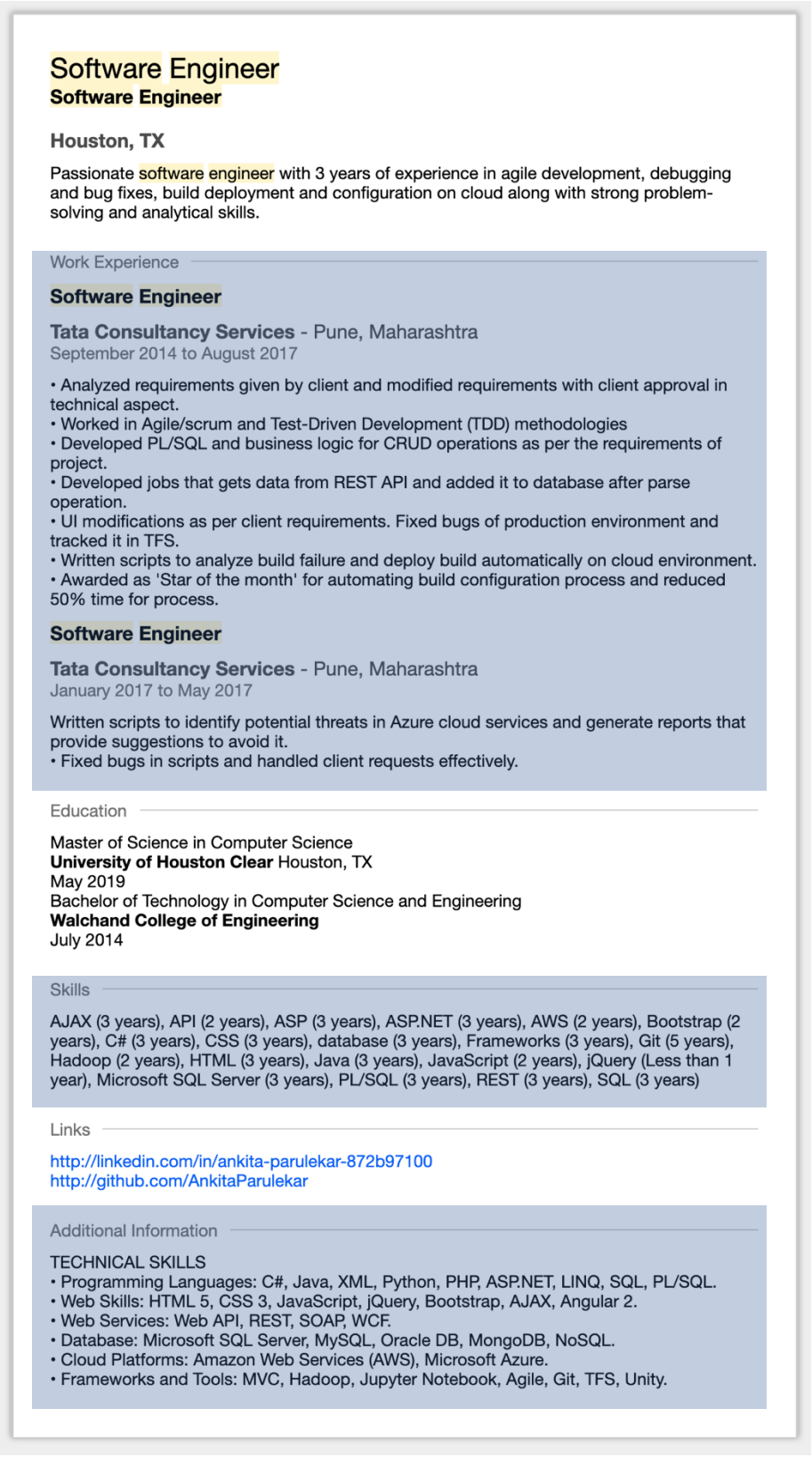
[10] Bordea, G., Lefever, E. and Buitelaar, P., 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 1081-1091).

[11] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

[12] Le, Q. and Mikolov, T., 2014, January. Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).

6. APPENDIX

6.1 CV sample



Software Engineer
Software Engineer

Houston, TX

Passionate software engineer with 3 years of experience in agile development, debugging and bug fixes, build deployment and configuration on cloud along with strong problem-solving and analytical skills.

Work Experience

Software Engineer

Tata Consultancy Services - Pune, Maharashtra
September 2014 to August 2017

• Analyzed requirements given by client and modified requirements with client approval in technical aspect.
• Worked in Agile/scrum and Test-Driven Development (TDD) methodologies
• Developed PL/SQL and business logic for CRUD operations as per the requirements of project.
• Developed jobs that gets data from REST API and added it to database after parse operation.
• UI modifications as per client requirements. Fixed bugs of production environment and tracked it in TFS.
• Written scripts to analyze build failure and deploy build automatically on cloud environment.
• Awarded as 'Star of the month' for automating build configuration process and reduced 50% time for process.

Software Engineer

Tata Consultancy Services - Pune, Maharashtra
January 2017 to May 2017

Written scripts to identify potential threats in Azure cloud services and generate reports that provide suggestions to avoid it.
• Fixed bugs in scripts and handled client requests effectively.

Education

Master of Science in Computer Science
University of Houston Clear Houston, TX
May 2019
Bachelor of Technology in Computer Science and Engineering
Walchand College of Engineering
July 2014

Skills

AJAX (3 years), API (2 years), ASP (3 years), ASP.NET (3 years), AWS (2 years), Bootstrap (2 years), C# (3 years), CSS (3 years), database (3 years), Frameworks (3 years), Git (5 years), Hadoop (2 years), HTML (3 years), Java (3 years), JavaScript (2 years), jQuery (Less than 1 year), Microsoft SQL Server (3 years), PL/SQL (3 years), REST (3 years), SQL (3 years)

Links

http://linkedin.com/in/ankita-parulekar-872b97100
http://github.com/AnkitaParulekar

Additional Information

TECHNICAL SKILLS
• Programming Languages: C#, Java, XML, Python, PHP, ASP.NET, LINQ, SQL, PL/SQL.
• Web Skills: HTML 5, CSS 3, JavaScript, jQuery, Bootstrap, AJAX, Angular 2.
• Web Services: Web API, REST, SOAP, WCF.
• Database: Microsoft SQL Server, MySQL, Oracle DB, MongoDB, NoSQL.
• Cloud Platforms: Amazon Web Services (AWS), Microsoft Azure.
• Frameworks and Tools: MVC, Hadoop, Jupyter Notebook, Agile, Git, TFS, Unity.

CV sample of a software engineer (only blue-coloured regions are considered in our scenario)

6.2 JD sample

Research Software Engineer, Google Brain Google - Mountain View, CA

In school or graduated within last 6 months? We encourage you to apply to openings on the Students Job Site .

Note: By applying to this position your application is automatically submitted to the following locations: Mountain View, CA, USA; New York, NY, USA

Minimum qualifications:
BA/BS degree in Computer Science, related technical field or equivalent practical experience.

Experience with one or more general purpose programming languages including but not limited to: C/C++ or Python

Experience with linear algebra, calculus and statistics

Machine learning experience

Preferred qualifications:
MS or PhD degree in Computer Science, Artificial Intelligence, Machine Learning, or related technical field.

Strong computer systems experience.

GPU programming experience.

Large data analysis and visualization experience.

Exposure to industry or academic research.

Exposure to Deep Learning, Neural Networks, or related fields and a strong interest and desire to learn about them.
About the job
We do research differently here at Google. Our team of Research Scientists aren't cloistered in a secret lab but are embedded throughout the engineering organization, allowing them to setup large-scale tests and deploy promising ideas quickly and broadly. Ideas may come from internal projects as well as from collaborations with research programs at partner universities and technical institutes all over the world. From creating experiments and prototyping implementations to designing new architectures, Research Scientists work on real-world problems including artificial intelligence, data mining, natural language processing, hardware and software performance analysis, improving compilers for mobile platforms, as well as core search and much more. But you stay connected to your research roots as an active contributor to the wider research community by partnering with universities and publishing papers.

You manage individual project priorities, deadlines, and deliverables, adapting to changes and setbacks in order to manage pressures, proving and applying theories through research efforts to develop new and improved products, processes, or technologies.

Research and Machine Intelligence is a high impact team that’s building the next generation of intelligence and language understanding for all Google products. To achieve this, we’re working on projects that utilize the latest techniques in Artificial Intelligence, Machine Learning (including Deep Learning approaches like Google Brain) and Natural Language Understanding. Our work gets used by product teams across Google, including Search, Maps and Google Now.

As a Research Software Engineer in the Google Brain team, you work and collaborate closely with Research Scientists on the team. You have the flexibility to switch projects as our research focus shifts and evolves. We need our engineers to be versatile and passionate about managing new problems.

This role bridges the gap between Software Engineer and Research Scientist. We are looking for great software engineers who also have experience with language understanding and perception (speech, images, video) - as well as improving algorithms. We work with teams across Google to make their products better and make Moonshots possible. In this role, you'll work alongside Research Scientists in the Google Brain team to bring their ideas to life by implementing algorithms, running experiments and building prototypes.
There is always more information out there, and the Research and Machine Intelligence team has a never-ending quest to find it and make it accessible. We're constantly refining our signature search engine to provide better results, and developing offerings like Google Instant, Google Voice Search and Google Image Search to make it faster and more engaging. We're providing users around the world with great search results every day, but at Google, great just isn't good enough. We're just getting started. Responsibilities

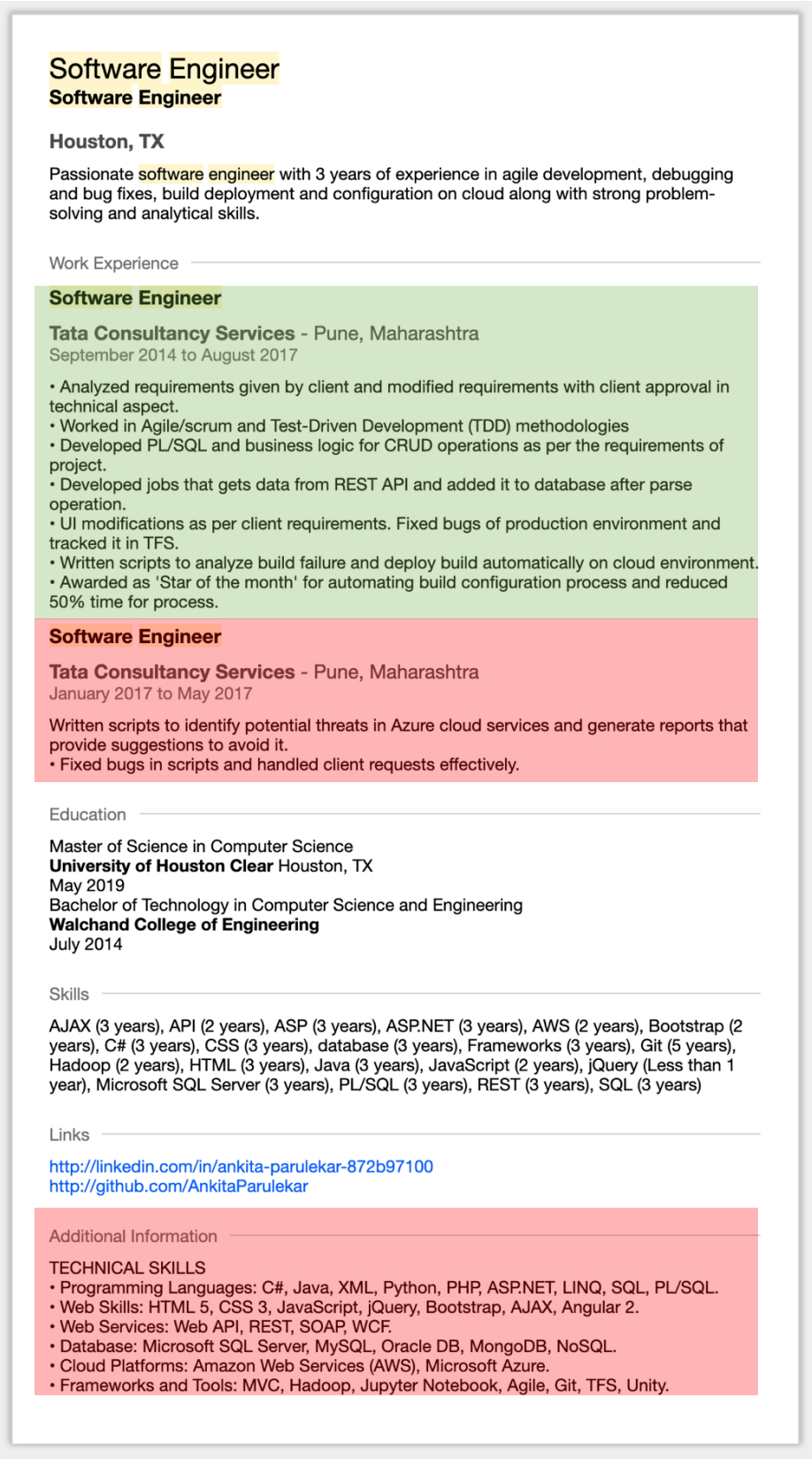
Participate in cutting-edge research in artificial intelligence and machine learning applications.

Develop solutions for real-world, large-scale problems.

At Google, we don’t just accept difference—we celebrate it, we support it, and we thrive on it for the benefit of our employees, our products and our community. Google is proud to be an equal opportunity workplace and is an affirmative action employer. We are committed to equal employment opportunity regardless of race, color, ancestry, religion, sex, national origin, sexual orientation, age, citizenship, marital status, disability, gender identity or Veteran status. We also consider qualified applicants regardless of criminal histories, consistent with legal requirements. See also Google's EEO Policy and EEO is the Law. If you have a disability or special need that requires accommodation, please let us know by completing this form .

JD sample (yellow-coloured region is the longest common substring to be removed, as explained in section 3.2.3. Pre-processing of CVs and JDs)

6.3 CV sample 2



Software Engineer
Software Engineer

Houston, TX

Passionate software engineer with 3 years of experience in agile development, debugging and bug fixes, build deployment and configuration on cloud along with strong problem-solving and analytical skills.

Work Experience

Software Engineer

Tata Consultancy Services - Pune, Maharashtra
September 2014 to August 2017

• Analyzed requirements given by client and modified requirements with client approval in technical aspect.
• Worked in Agile/scrum and Test-Driven Development (TDD) methodologies
• Developed PL/SQL and business logic for CRUD operations as per the requirements of project.
• Developed jobs that gets data from REST API and added it to database after parse operation.
• UI modifications as per client requirements. Fixed bugs of production environment and tracked it in TFS.
• Written scripts to analyze build failure and deploy build automatically on cloud environment.
• Awarded as 'Star of the month' for automating build configuration process and reduced 50% time for process.

Software Engineer

Tata Consultancy Services - Pune, Maharashtra
January 2017 to May 2017

Written scripts to identify potential threats in Azure cloud services and generate reports that provide suggestions to avoid it.
• Fixed bugs in scripts and handled client requests effectively.

Education

Master of Science in Computer Science
University of Houston Clear Houston, TX
May 2019
Bachelor of Technology in Computer Science and Engineering
Walchand College of Engineering
July 2014

Skills

AJAX (3 years), API (2 years), ASP (3 years), ASP.NET (3 years), AWS (2 years), Bootstrap (2 years), C# (3 years), CSS (3 years), database (3 years), Frameworks (3 years), Git (5 years), Hadoop (2 years), HTML (3 years), Java (3 years), JavaScript (2 years), jQuery (Less than 1 year), Microsoft SQL Server (3 years), PL/SQL (3 years), REST (3 years), SQL (3 years)

Links

http://linkedin.com/in/ankita-parulekar-872b97100
http://github.com/AnkitaParulekar

Additional Information

TECHNICAL SKILLS
• Programming Languages: C#, Java, XML, Python, PHP, ASP.NET, LINQ, SQL, PL/SQL.
• Web Skills: HTML 5, CSS 3, JavaScript, jQuery, Bootstrap, AJAX, Angular 2.
• Web Services: Web API, REST, SOAP, WCF.
• Database: Microsoft SQL Server, MySQL, Oracle DB, MongoDB, NoSQL.
• Cloud Platforms: Amazon Web Services (AWS), Microsoft Azure.
• Frameworks and Tools: MVC, Hadoop, Jupyter Notebook, Agile, Git, TFS, Unity.

CV sample of a software engineer (red-coloured regions are past information and green-coloured region is the current work experience; two regions form a test pair)