



**Wee Kim Wee School of Communication and Information**

18S1-SCI- H6751: Text and Web Mining

---

## **Group Project Report**

### **Automated Job - Resume Matching Solution**

Luo Shuang (G1702502A)

Wang Sixin(G1801764B)

Xia Xiaolong(G1801412A)

Zhao Hengrui(G1801739D)

Date: 28 October 2018

Nanyang Technological University

## **1. Introduction and Objective**

According to a 2015 study on job seeking behavior by Pew Research Center, 79% of the job seekers utilized the online resources for their most recent employment (Aaron ,2015). This study result suggests that the online job boards become the major channel for job seekers in the digital era. However, another finding in the study indicates that most of the job seekers fail to match their experiences with the job requirements and spend hours on job board to apply job which is not seen to be suitable (Aaron, 2015). Additionally, Dr. John Sullivan conducted a similar research in 2013 which highlighted some interesting aspects: on average, 250 resumes are received for each job opening by the major organizations, more than 50% of the resumes does not meet the minimum requirement (John, 2013). This means the time our recruiter spends on these 50% of the resumes for each job is wasted. From both candidate and recruiter's points of view, the phenomenon may suggest that the traditional online job board does not seem to simplify the job application process or reduce the effort required from both parties. With this challenge getting bigger and bigger, the demand to automate the resume - job matching process is getting increased as well. For instance, the content - based recommendation system (CBR) is introduced to analyze the job description to identify the potential area of interest to the job seekers (Shiqiang et al., 2016). To apply the concept in Singapore local context, our team has conducted a text mining project based on the data acquired from the major online job board in Singapore. The primary objective of this project is to create a machine learning model to accelerate the job - resume matching process. The detail of the text mining methodology and results are presented in the following sections.

## **2. Data Extraction**

To support effective model training, the team crawls down and exports 453 job postings (IT software related) and 84 resumes (IT software related) from the selected local online job boards. Currently there are quite a few online job boards operating in Singapore, for example, "JobStreet.com.sg", "JobsCentral.com.sg", "JobsDB.com.sg", "Monster.com.sg", etc. According to the company background, "JobStreet.com.sg" was founded in 1997 and focus on the market in Southeast Asia (JobStreet, 2018). The total number of active posting under JobStreet.com.sg is 61,616 as of November 2018, the volume is much larger than the posting from the other online Job Boards. Additionally, some of the major local companies such as NCS, Capitaland, and Nanyang Technological University are using "JobStreet.com.sg" as their recruitment platform. Based on its strong local presence, the team decides to extract job posting from "JobStreet.com.sg".

Every online job board is having its own standard, the way it organizes the resume and job

posting is highly correlated. In order to minimize the correlation and improve the generalizability of the machine learned model, the resumes are exported from a different job board. According to our research, another online job board “Indeed” provides a free online resume repository, thus we export 84 resumes from “Indeed” to support the model training.

## 2.1 Data Extraction Technique

The tool we use to extract the job posting from “JobStreet.com.sg” is a web - based application called “import.io”. This tool allows users to extract information from multiple web pages at one go. The build-in web crawler can detect the URL embedded under the web pages, and drill into all URLs to extract information automatically (import.co, 2018). The advantage of this tool is no installation required, and no additional coding knowledge required as well. All we need to do is to specify the URL of web page, and the application starts extracting the information automatically. The “web extractor” screen in “import.io” is shown in Figure 1.

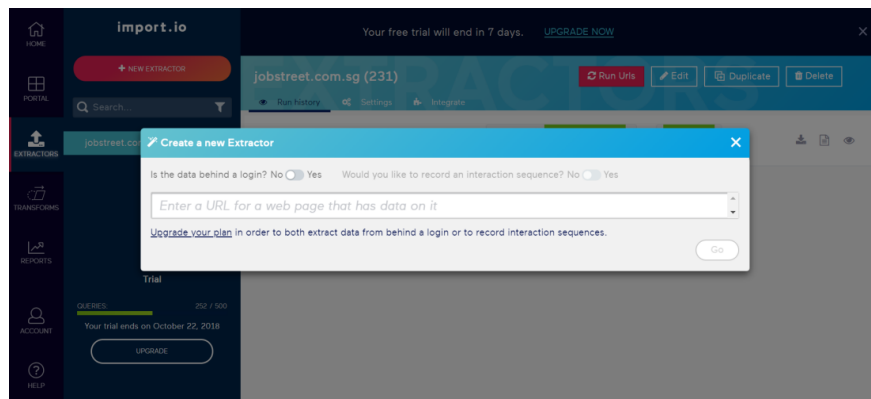


Figure 1. User interface of “import.io”: specify the URL to start extraction.

The preview of the data extraction is shown in Figure 2, the application is able to detect the company name, company logo, position detail, job description and requirement. The extracted information is well organized in a table format and associated with column header. The output file is ready for download at end of the extraction process. Users can choose to download in Excel, CSV or NDJSON format.

Row	Hidden Image	Position Link	Company Name	Job Location	Login View	List Un Styled	Job D
1		Functional Subject Matt...	Oracle Corporation Sin...	West	Login to view salary	As FLEXCUBE Univers...	5 n
2		Functional Subject Matt...	Oracle Corporation Sin...	West	Login to view salary	As FLEXCUBE Univers...	5 n
3		Cloud Customer Incide...	Oracle Corporation Sin...	West	Login to view salary	Do you like being part o...	17
4		AVP / Senior Associate, ...	DBS Bank Ltd	Singapore - singapore		Business Function Grou...	29
5		App Developer - Exp wi...	RecruPlus Consulting ...	Central	Login to view salary	Our client, a premium I...	35
6		Data Governance Cons...	Manpower Staffing Serv...	Central	Login to view salary	Conducive environment...	46
7		Assistant Manager / Ma...	Ministry of Health	Central	Login to view salary	ResponsibilitiesYou will ...	55
8		Android App Developer ...	PropNex Realty Pte Ltd	Toa Payoh (Central)	Login to view salary	Android App Developer ...	1 h
9		Assistant / Deputy / Ma...	Land Transport Authorit...	Central	Login to view salary	Under the Bus Contract...	1 h
10		IOS App Developer (Co...	PropNex Realty Pte Ltd	Toa Payoh (Central)	Login to view salary	IOS App Developer (Co...	1 h
11		Senior Executive (Food ...	SATS	Singapore - SATS Infig...		About usCOMPANY IN...	2 h

Figure 2. The preview page of the data extraction, user can select the columns for download, and arrange the column sequence by drag and drop the column header.

## 2.2 Extract and Analyze Raw Data

The primary object of our project is to assign resumes with the most suitable job posting(s), the essential information required for the mining process includes the job title, job description and job requirement from job posting, and the working experience, education, skills from resume.

We analyse the raw data crawled by import.io, here are some characteristics of raw dataset:

- Company background, working culture and other non-core information are mixed up with job description and job requirement.
- Amounts of HTML/CSS/JavaScripts codes are embedded within the text(both job postings and resumes).
- Word overlaps are obvious among all job postings, such as “job”, “role”, “position”, which may interfere performance of accurate clustering.
- There are certain word overlaps between job postings and resumes, which is possible for using resumes as test set.

Such characteristics help us clean the raw data, so as to improve the accuracy of subsequent processes. Python is used to implement cleansing; the sample python code is show in Figure 3.

- Remove some irrelevant information from the exported job posting data. For instance, the company background, working culture as well as HTML/CSS/JavaScripts codes.
- Remove escape characters and other signifiers.
- Conduct a word dictionary for future use in RapidMiner to remove highly-overlapped words.

```
df = pd.read_excel('IT Jobs in Singapore.xlsx')
titles = []
desc = []
for indexs in df.index:
    str_old = df.loc[indexs].values[1]
    str_old = str_old.replace("\n", " ").replace("\t", " ").replace("&nbsp", " ")
    if str_old.rfind("WORK LOCATION") != -1: # if not find, return -1
        str_new = str_old[:str_old.rfind("WORK LOCATION")]
        df.loc[indexs].values[1] = str_new
        titles.append(df.loc[indexs].values[0])
        desc.append(str_new)
    elif str_old.rfind("RECRUITMENT FIRM SNAPSHOT") != -1:
        str_new = str_old[:str_old.rfind("RECRUITMENT FIRM SNAPSHOT")]
        df.loc[indexs].values[1] = str_new
        titles.append(df.loc[indexs].values[0])
        desc.append(str_new)
    elif str_old.rfind("COMPANY SNAPSHOT") != -1:
        str_new = str_old[:str_old.rfind("COMPANY SNAPSHOT")]
        df.loc[indexs].values[1] = str_new
        titles.append(df.loc[indexs].values[0])
        desc.append(str_new)
    else:
        str_new = str_old
        df.loc[indexs].values[1] = str_new
        titles.append(df.loc[indexs].values[0])
        desc.append(str_new)
df.to_excel("IT Jobs in Singapore_v2.xlsx", index=False)
```

Figure 3. Remove irrelevant information from exported job posting dataset by Python.

## 2.3 Datasets Characteristic after Cleansing

Job requirement is merged into job description since there are a lot of overlaps between these two features. The final dataset for the exported job posting includes 2 columns, which are job title and job description, the sample is shown in Figure 4.

title	description
Assistant IT Project Manager	job responsibilities overall responsible for project life cycle which includes planning, organizing, leading, controlling and etc liaise with
Business & Data Analyst (Up to \$7k) ref: Naz	1) analyse business process ;support mapping & reviewing of existing business processes to identify areas of improvements, achieving
[NO EXP OK] Programming Assistant (SQL / C# / VBA)	► up to \$52.5k + aws + vb!!!► company currently expanding with exciting projects - great opportunity for career growth and prospect.►
CRM Executive - EAST / Up to \$2,500 (No Experience)	independently respond and manage feedback and enquiries received from members that involves but are not limited to registration, r
Software / Web Developer (UP \$3.5K / North / 5 C 5 working days (8.30am – 5.30pm) 5 woodlands 5 basic up \$3,500 + vb 5 electronics it industry; design of web portal to present data a	
Software / Performance Test Engineer (US MNC / ► us mnc, top industry leader in secure file sharing and enterprise cloud solutions ► no experience is needed, comprehensive training )	
Project Manager & IT Operations (Pre-Sales)	our client, a leading company in it solutions & business processing outsourcing services is looking out for a project manager (it solutions
Data Migration Specialist (7 months contract)	the data migration specialist leads the organisation's systems transformation project data migration work stream. he/she will be respon
Service Delivery Manager	: ensure the delivery of cost effective and quality services maintain regular open communication with assigned customers and engine
Java Developer (1 year agency contract) *GOOD C	*good career prospect*we are looking for a java developer (1 year agency contract) to perform the following job scope: ;building java a
Java Developers (Senior & Junior)	• birthday benefits • flexi & fun working environment • central - near mrt : develop and enhance in-house crm ap
Test Engineer	your new company you will be working at a global leading bank in singapore. your new role you will be part of the fm ecommerce syste
Initial Margin Project Manager (contract)	responsibilities:prepare and track project budgetprepare and track detail project plan/activitiesprepare and organize project steering cc
System Engineer (LAN / UAT / SCCM / MSSQL / PII)	• 5 day work week• career progression• attractive salary packageinterested applicants can send your resume to supreme.tammytan@gr
Salesforce Cloud Engineer	job propose and manage implementation of salesforce commerce cloud and other solutions that meet business needs develop sales
IT - Applications Domain Expert (Passenger Service)	you will be responsible for the following:implement revenue management and inventory solutions on a passenger servicing system(ps
Java Developer	• proficient in server programming using ejb • experience in java messaging services jms • experience in intelliJ and eclipse

Figure 4. The sample of the job posting dataset, the total is 453 rows.

The sample dataset for the exported resume is shown in Figure 5. Candidate's experience, education and skills are merged into a single column as “description”. Another column is “title”, indicating title of what position the applicants want to apply.

title	description
DATA SCIENTIST	pravin manickarajdata scientistsingapore-microsoft certified data scientist with 3+ years of exp
Process Analyst	process analystwoodlands-work experienceprocess analystservion global solutions - india-octo
Senior Business Analyst	senior business analystsenior business analyst - sungard financial solutionscity hall-work experi
Servicedesk Analyst	servicedesk analystservicedesk analyst - holcim east asia business service center b.v-to contribu
Programmer/Business Analyst	software developer and business analyst with more than 14 years of successful experience in de
QA Tester/ Analyst	summary:almost 4 years of diversified experience in the various client/server, web environmen
Data Architect/Data modeller	to be part of a growing and professional organisation and sharpen my data architect/data mode
Programmer Analyst	to obtain an information technology position within an affluent company who担 past, current
Programmer Analyst	university of wisconsin – green bay• bachelor of science in information science, may 2014o dea
Web analyst	web analystweb analystsingapore-work experienceweb analystdecision science agency-septeml
Security Analyst IBM ISERIES	work with ibm iseries computers.over 30 years experience using ibm computers of all sizes.i hav

Figure 5. The sample of the final resume dataset, the total is 84 rows.

## 3. Text Mining Tool

The primary text mining tool used in our project is “RapidMiner”, developed by Rapid-I, GmbH of Dortmund, Germany. It is a GUI based data mining software which specialized in text mining, machine learning, data preparation, and predictive analytics (RapidMiner, 2018). Compare to many other data mining software, “RapidMiner” is much easier to install, configure and use (Matthew, 2012). The capability of flexible data handling and fast model training suggest a good fit for our project.

## 4. Creating Text Mining Process

### 4.1. High Level Process Flow

Our team applied the combination of text clustering and text classification methods to train the model, the high-level process flow is shown in Figure 6. The process is broken down into the

below 4 major steps.

1. Cluster and label each position in the unlabelled job posting dataset by k-Means clustering. (try out several clustering methods and choose the best-performed one)
2. In order to ensure efficiency of later classifier, tuning work is applied to the clustered labels.
3. Use labelled job posting dataset as training dataset to train a classification model.
4. Use the labelled resume dataset to test the model and evaluate its performance.

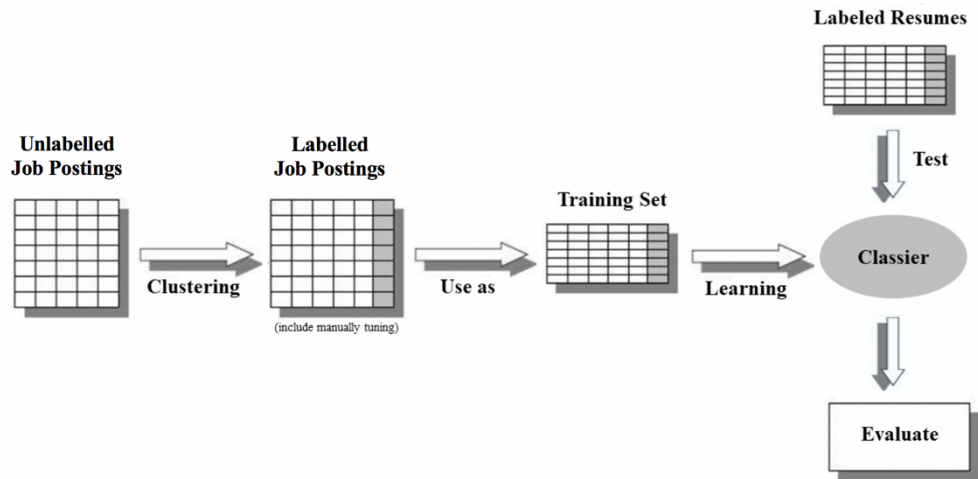


Figure 6. High-level process flow for clustering & classification to predict and evaluate unlabelled data.

## 4.2. Importing and Preprocessing Unlabelled Data for Clustering

We create a new local repository, “Job-Resume Matching” and initiate the process by importing the data file into “RapidMiner”, as shown in Figure 7. For job description dataset, both title and description are set as text

ExampleSet (453 examples, 0 special attributes, 2 regular attributes) Filter (453 / 453 examples)

Ro...	title	description
1	Assistant IT Project Manager	job responsibilities overall responsible for project life cycle which includes plan...
2	Business & Data Analyst (Up to \$7k) ref: Naz	1) analyse business process ;suppoirt mapping & reviewing of existing business p...
3	[NO EXP OK] Programming Assistant (SQL / C#...	► up to s\$2.5k + aws + vb!!!► company currently expanding with exciting proje...
4	CRM Executive – EAST / Up to \$2,500 (No Ex...	:independently respond and manage feedback and enquiries received from mem...
5	Software / Web Developer (UP \$3.5K / North ...	▣ 5 working days (8.30am – 5.30pm)▣ woodlands▣ basic up \$3,500 + vb▣ ele...
6	Software / Performance Test Engineer (US MN...	►us mnc, top industry leader in secure file sharing and enterprise cloud solutions...
7	Project Manager & IT Operations (Pre-Sales)	our client, a leading company in it solutions & business processing outsourcing ser...
8	Data Migration Specialist (7 months contract)	the data migration specialist leads the organisation's systems transformation proj...
9	Service Delivery Manager	: ensure the delivery of cost effective and quality services maintain regular ope...
10	Java Developer (1 year agency contract) *GOO...	*good career prospect*we are looking for a java developer (1 year agency contra...

Figure 7. The job posting dataset is imported into “RapidMiner”.

### 4.3. Preprocessing Data

According to previous analysis and subsequent need, we conduct preprocessing before going into clustering. The subprocess below is added in ‘Process Document from Data’ (Figure 8):

- Using ‘Tokenize’ with non letters to split the document into words.
- Using ‘Tokenize’ with English to conduct part of speech (POS) tagger.
- With ‘Filter Stopwords(English)’, we can remove common English stopwords from the text.
- Wordlist conducted in step2.2 is imported to ‘Filter Stopwords(Dictionary)’ as dictionary.
- Adding ‘Filter Tokens(lengths)’ and setting parameters to 2 and 25, words with 1 character or more than 25 characters are removed.
- ‘Stem(Porter)’ stems the text content and applies an iterative, rule-based replacement of word suffixes, to reduce the length of words.
- ‘Transform Cases’ is added to transform all characters to lower cases.
- With ‘Generate n-grams’ operator (max length set to 2), a series of consecutive tokens of 2 words are created.

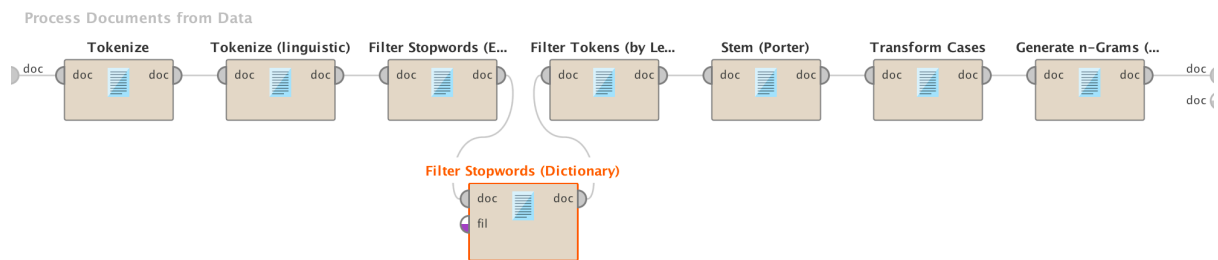


Figure 8. Subprocess of preprocessing the unlabelled dataset.

### 4.4. Clustering the Unlabelled Job Posting Dataset

The objective of this step is to assign label to each job posting through clustering process. We try for 3 clustering methods: k-Means clustering, EM clustering and DBSCAN. After comparing the clustered labels and original text content of each job posting, we found that k-Means made best balance of performance and efficiency. The detail of the clustering process is shown in Figure 9.

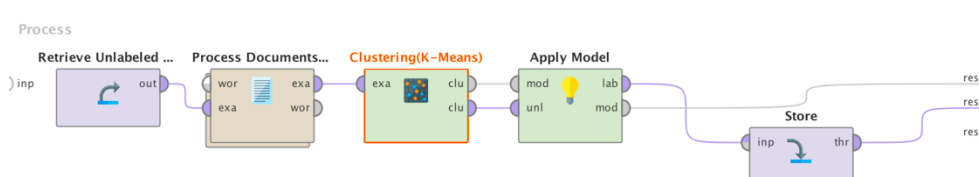


Figure 9. The clustering process to assign label to job postings dataset based on the similarity.



Furthermore, we drill down into k-Means and try to further improve clustering performance. (Performance is based on the comparison between clustered outcome and later tuning outcome.) We predefine different number of clusters and find out k=4 brings best outcome: looking into the top15 most frequent words of each cluster, overlaps are the least when k=4 (Figure 10). When k=5, overlaps are obvious between 2 developer positions; when k=3, most of analyst positions are clustered with developer positions.

Attribute	cluster_0 ↓	Attribute	cluster_1 ↓	Attribute	cluster_2 ↓	Attribute	cluster_3 ↓
test	0.455	data	0.213	project	0.372	develop	0.239
develop	0.106	busi	0.126	manag	0.274	applic	0.136
engin	0.100	manag	0.108	project_manag	0.122	java	0.125
softwar	0.082	system	0.100	develop	0.077	design	0.104
perform	0.063	develop	0.078	busi	0.071	softwar	0.098
design	0.062	support	0.064	applic	0.061	system	0.075
system	0.052	inform	0.053	ensur	0.057	technolog	0.055
manag	0.045	analyt	0.053	implement	0.054	test	0.054
product	0.042	solut	0.052	inform	0.053	program	0.052
applic	0.042	technolog	0.050	plan	0.048	implement	0.049
test_engin	0.041	work	0.047	system	0.048	code	0.048
autom	0.039	process	0.044	commun	0.046	web	0.048
appli	0.037	project	0.043	process	0.044	candid	0.043
comput	0.036	applic	0.043	technolog	0.044	busi	0.042
pleas	0.035	servic	0.041	includ	0.042	appli	0.042

Figure 10. The clustering process to assign label to the job postings dataset based on the similarity.

At the end of the clustering process, we can still see some inappropriately clustered observation, which is unavoidable. So, we apply fine tuning on them, transform cluster number (i.e. ‘Cluster\_3’) to nominal name (i.e. ‘developer’) and conduct a modified labelled job posting dataset. It is stored as training dataset for the subsequent steps, shown in Figure 11.


Row No.	label		Ro... ↑	label
1	cluster_2	<p><b>Fine tuning cluster label</b></p>  <p><b>Transform label name</b></p>	1	project
2	cluster_1		2	data
3	cluster_1		3	data
4	cluster_1		4	data
5	cluster_3		5	data
6	cluster_0		6	test
7	cluster_2		7	project
8	cluster_1		8	data
9	cluster_2		9	data
10	cluster_3		10	developer

Figure 11. The labelled job description dataset is stored at the end of the clustering process.



Comparing initial cluster labels and tuned cluster labels, we calculate accuracy, precision, recall and F-1 score to evaluate performance each clustering process that we conducted in the previous section.

Clustering Method	k value	Measurement	Accuracy
k-Means	4	Cosine Similarity	84.54%
k-Means	4	Euclidean Distance	83.44%

Figure 12. Results of clustering with 2 different measurements

With the same attribute  $k=4$ , we compare 2 different numerical measurement types: Euclidean Distance and Cosine Similarity. It turns out latter one yields better performance. At last, we adopted the tuned outcome of clustering process with  $k\text{-Means}=4$  (cosine similarity) as training data in later classification. Performance of each cluster is shown as Figure 13.

	'test' (cluster_0)	'data' (cluster_1)	'project' (cluster2)	'developer' (cluster_3)
<b>Class recall</b>	97.40%	67.02%	97.06%	96.51%
<b>Class precision</b>	91.46%	100%	86.84%	63.36%
<b>F1-score</b>	94.34%	80.25%	91.67%	76.50%

Figure 13. Results of  $k\text{-Means}$  clustering ( $k=4$ , cosine similarity)

#### 4.5. Training Classification Model with Labelled Job Postings Dataset

In the training process, we use the training data file generated from the previous step to perform classification based on the label assigned by clustering process. We apply several algorithms to train the model and use the 10-Fold Cross-Validation to evaluate the prediction accuracy (Figure 14, Figure 15)

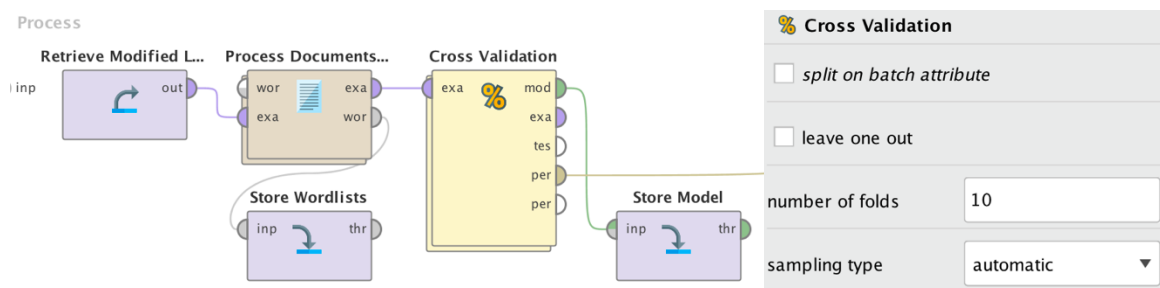


Figure 14. The training process for the classifier 10-Fold Cross-Validation with Decision Tree

Algorithm	Training Accuracy
Naïve Bayes	82.98% +/- 5.12%
Decision Tree	89.83% +/- 4.70%
Boosting Tree	94.03% +/- 4.98%
Gradient Boosted Tree	94.91% +/- 3.86%
Logistic Regression	80.56% +/- 4.84%
Linear SVM	84.08% +/- 5.00%
K-NN (K=3)	88.27% +/- 6.07%

Figure 15. Training Results of Different Algorithms

Trees algorithms show highest training accuracy, but we view the high accuracy as signs of overfitting. Further pruning is needed, in order to reduce the size of decision trees and to improve predictive accuracy by the reduction of overfitting. (Predictive accuracy shown in Figure 15.)

K-NN is relatively better than the others in the comparison results. Because both clustering and classification are based on similarity, and K-NN is to measure similarity among documents. This makes K-Nearest Neighbors algorithm an appropriate for this scenario.

#### 4.6. Applying Resume Dataset as Test Set to Model

The resume dataset mentioned in step2.3 is imported and cluster labels are added on manually. The test process is built using the resume dataset we preprocessed in the previous step; the detail of testing process is shown in Figure 16. We also verify the testing performance in the next step and adjust parameters of operators to improve the testing accuracy (Figure 17).

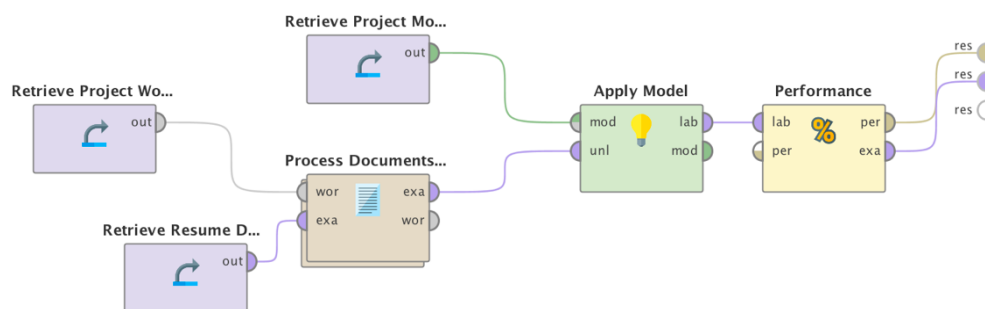


Figure 16. The testing process for the classifier

<b>Algorithm</b>	<b>Testing Accuracy</b>
Naïve Bayes	73.81%
Decision Tree	77.38%
Boosting Tree	82.14%
Gradient Boosted Tree	77.38%
Logistic Regression	82.14%
Linear SVM	70.24%
K-NN (K=3)	83.33%

*Figure 17. Testing results of different algorithms*

K-NN yields rather good training accuracy and testing accuracy. So, we continue to adjust parameters of K-NN operator. It turns out when K=5, both training and testing performance reach best outcomes (Figure 18).

<b>Algorithm</b>	<b>K value</b>	<b>Training Accuracy</b>	<b>Testing Accuracy</b>
K-NN	3	88.27% +/- 6.07%	83.33%
K-NN	5	89.15% +/- 6.07%	83.33%
K-NN	7	88.27% +/- 6.07%	83.33%

*Figure 18. Results of K-NN*

## 5. Discussion and Limitation

Based on results mentioned above, we conclude some interesting points:

- Records labelled as ‘data’ have tendency to be clustered as ‘developer’, which means the boundary lines are not so definite between ‘data’ and ‘developer’ and indicates developer positions and data analyst positions have more commons in skill requirements and work details. In terms of job recommendation, data analyst position may also be suggested to candidates who are initially interested in developer only.
- When job titles are included in the text, clustering performance is greatly improved. It suggested title context are highly distinguishable and essential for job clustering.
- Though elementary, k-Means still works best among all clustering methods. It balances performance and time efficiency.
- Filter words dictionary is important to this case. Job posting contents have amounts of same/overlaps, so do resumes. So, using filter words dictionary helps to accelerate clustering efficiency.

During the project development, our team discovered some limitations and challenges:

- The sample size is relatively smaller than the total records of job website, which may not be sufficient to cluster and train a model with generalization strength.
- A big challenge is the text length, for instance, text becomes quite lengthy after merged from the job requirement and job description in job posting dataset. Lengthy contents are time-consuming and memory-consuming when we run clustering, which forces us to cut down initial number of job postings. The behavior may suggest that in order to generalize the process in a much bigger database, we may need more resources to support the computation.
- Applying decision trees improves results modestly. However, decision trees suffered from overfitting to the data(Weiss et al., 1999). Tree algorithms in our study perform best in training process but they get overfitting. Further pruning is not applied to adjust performance.

## Reference

- [1] Aaron, S. (2015). Searching for Work in the Digital Era, Pew Research Center, November 2015. Retrieved on October 15, 2018, from the website: [http://www.pewresearch.org/wp-content/uploads/sites/9/2015/11/PI\\_2015-11-19-Internet-and-Job-Seeking\\_FINAL.pdf](http://www.pewresearch.org/wp-content/uploads/sites/9/2015/11/PI_2015-11-19-Internet-and-Job-Seeking_FINAL.pdf)
- [2] John, S. (2013). Why You Can't Get A Job ... Recruiting Explained By the Numbers. Retrieved on October 15, 2018, from the website: <https://www.ere.net/why-you-cant-get-a-job-recruiting-explained-by-the-numbers/>
- [3] Guo, S., Alamudun, F., & Hammond, T. (2016). Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications*, 60, 169-182.
- [4] JobStreet. (2018). About Us. Retrieved on October 16, 2018, from the website: <https://www.jobstreet.com.sg/en/about-us/>
- [5] Import.co. (2018) Import.io Extract: get structured data from web pages. Retrieved on October 16, 2018, from the website: <https://www.import.io/builder/data-extraction/>
- [6] RapidMiner. (2018). RapidMiner Platform Lightning Fast Data Science Platform. Retrieved on October 16, 2018, from the website: <https://rapidminer.com/products/>
- [7] North, M. (2012). Data mining for the masses (pp. 91-100). Athens: Global Text Project.
- [8] Wowczko, I. A. (2015). Skills and Vacancy Analysis with Data Mining Techniques. *Informatics* 2, 4 (2015), 31.
- [9] Patel, B., Kakuste, V., & Eirinaki, M. (2017, April). CaPaR: A Career Path Recommendation Framework. In *Big Data Computing Service and Applications (BigDataService)*, 2017 IEEE Third International Conference on (pp. 23-30). IEEE.
- [10] Kwartler, T. (2017). Text mining in practice with R. John Wiley & Sons.
- [11] Weiss, S. M., Apte, C., Damerau, F. J., Johnson, D. E., Oles, F. J., Goetz, T., & Hampp, T. (1999). Maximizing text-mining performance. *IEEE Intelligent Systems and their applications*, 14(4), 63-69.