Senior Research

The tracking skills of Thai labour market using unsupervised machine learning with internet data

Sikkawit Amornnorachai

5845622029

Advisor

Tanapong Potipiti, Ph.D., Assoc. Prof.

May 2019

Senior Research submitted in a partial fulfilment of the requirements

for the Bachelor of Economics

Faculty of Economics, Chulalongkorn University

Bangkok, Thailand

Academic Year 2018

The tracking skills of Thai labour market using unsupervised machine learning with internet data

Sikkawit Amornnorachai

This paper uses internet data in Thai labour market by scraping job portal websites and using natural language processing technique to extract the meaning of words appeared in job description data. This paper also uses the job description data that posted in job portal websites to cluster the job by skills and finding which skill sets gain more or less wage than the market wage data by using linear regression technique.

This paper finds that jobs in Thai labour market can be clustered by skill set into 8 groups as followings – customers service, engineering, sales, production, accounting and finance, language works, office works, and marketing. The marketing skill set gain most wage with respect to other groups, meanwhile, the customer service skill set gain least wage with respect to other groups. Moreover, labour with more working experience can acquire more wage than the newly graduated group.

All codes and documentations are under the MIT Licence and they can be accessed at github.com/sikawit/EconSR.

Keywords: Thai labour market, machine learning, unsupervised clustering, web scraping, Natural Language Processing, Word2Vec, Bag of Words

1. Introduction and Literature Review

In the era of disruption by technologies, a lot of jobs in the labour market are disappearing. However, a lot of new jobs are presented in the market. For example, the pager operator was disappeared in 2000s, while a brand-new job like data scientist was created in 2010s.

In the traditional way, to study the labour market, a lot of researchers use the Labour Force Survey to track the status of labour market and the dynamic of the data in this market. However, there are a lot of drawbacks in this traditional study, which is shown in Table 1. The Labour Force Survey (LFS) data are quarterly lagged and monthly collected from the survey respondents which are not reflected the instant labour market status. Moreover, to access LFS, permission is required to access data from the National Statistical Office (NSO). Moreover, the data from NSO is not free of charge.

Moreover, the type of jobs and the job description in LFS are predefined by human which does not capture the dynamic of jobs in this disruptive era that a lot of new jobs are created and disappeared. To capture the dynamics of jobs, it is better to use the real-time job posting datasets from job portal sites.

Because of these limitations of the traditional study with LFS, text mining technique is selected to attain all online job postings dataset from the Internet Employment websites which are instantly and totally free to access. Even there is a study to find the insight in Thai labour market, but this paper uses the Google Trends data as a proxy to study in the labour market. (Nuarpear Warn Lekfuangfu, 2016) Thus, there is no paper using Thai online datasets from job portal sites to track Thai labour market.

Table 1, The difference between 2 studying methods

Topics	**Traditional Study**	**Text Mining**
Source of Data	Labour Force Survey	Online Dataset
Availability	Require Permission	Public Data
Time to get access	Few weeks	Immediately
Cost to access	Not Free	Free
Frequency of Data	Monthly	Real Time
Lagging Period	Quarterly	Real Time
Type of Jobs	Predefined by human	Adaptive by actual data
Job Description	Predefined by human	Dynamic

From the text mining approach, there are a lot of applications that policymakers can use data from the Internet to find the insight from labour market data. For example, tracking real time labour demand, measuring a set of skill embedded in each job in the market, clustering jobs on skill set, planning the education curriculum to serve the market demand, analyzing wage by skill set.

In this paper, I use the Natural Language Processing (NLP), machine learning, and web scraping techniques to answer 2 main questions. First, how can we cluster the job by skills from appearing words in the job description. Second, which skill set gain higher (or lower) wage than the average wage in the Thai Labour market.

This paper is divided into 8 parts, introduction and literature review (as mentioned above), dataset in this paper, data pre-processing, empirical results, limitation, summary, code licence, and acknowledgement.

2. Dataset

In this study, two online datasets are selected for scraping for labour demand and labour wage. The labour demand data are collected from Jobtopgun, one of the largest Thai online job posting website. In this dataset, there are composed of the job position name, and job description. Data from Jobtopgun are monthly scraped from December 2018 to March

2019. The detail of data on Jobtopgun website is shown in Figure 1 and the number of observations from Jobtopgun is shown in Table 2.

Figure 1: Data in Jobtopgun site

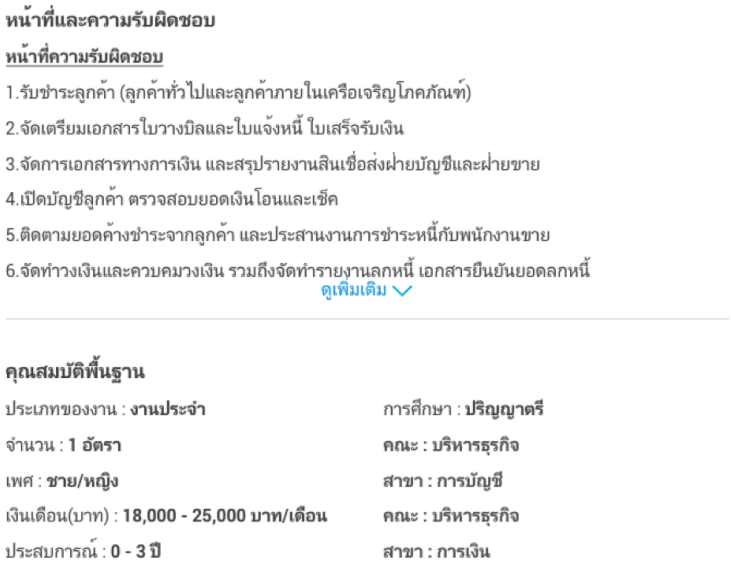


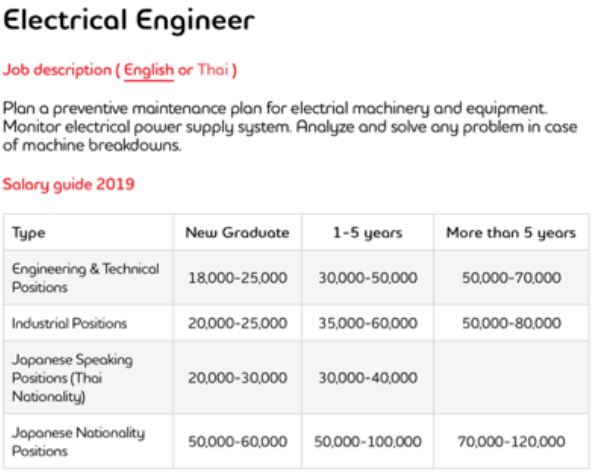
Table 2, The number of observations of Jobtopgun dataset

Period	Number of observations
December 2018	13,564
January 2019	11,418
February 2019	8,281
March 2019	9,285

The labour wage dataset is collected from Adecco Thailand Salary Guide 2019. The salary data are provided by Adecco, a global human resource consultant. In this dataset, there are 766 observations based on job description and work experience.

In this dataset, midpoint value in the wage's range is used for calculating wage in the regression equation. Moreover, only the data in the first row of salary table in the website is considered as a selected wage because data in this row are mostly related with the job description. Data on Jobtopgun website is shown in **Error! Reference source not found.**.

Figure 2, Data in Adecco Thailand Salary Guide 2019 site



Electrical Engineer

Job description (English or Thai)

Plan a preventive maintenance plan for electrial machinery and equipment. Monitor electrical power supply system. Analyze and solve any problem in case of machine breakdowns.

Salary guide 2019

Type	New Graduate	1-5 years	More than 5 years
Engineering & Technical Positions	18,000-25,000	30,000-50,000	50,000-70,000
Industrial Positions	20,000-25,000	35,000-60,000	50,000-80,000
Japanese Speaking Positions (Thai Nationality)	20,000-30,000	30,000-40,000	
Japanese Nationality Positions	50,000-60,000	50,000-100,000	70,000-120,000

3. Data Pre-Processing

Due to text mining technique, this study uses Python for scraping job posting data from Jobtopgun and Adecco. Data in both datasets, the labour demand and labour pricing, are rich of text. The objective on this paper is converting text into computable object for extracting meaning of skill set that embedded in text by using the Natural Language Processing (NLP). In the NLP session, translation, Bag of Words and Word2Vec techniques are selected for converting text into a computable object and apply k-means clustering, a machine learning technique to cluster the computed text data.

a. Translation

Scraped data from Jobtopgun and Adecco are in Thai and English. For analysing purposes and limitation of NLP techniques in Thai language. All job descriptions and job position names are translated into English by using Google Translate service.

b. Bag of Words

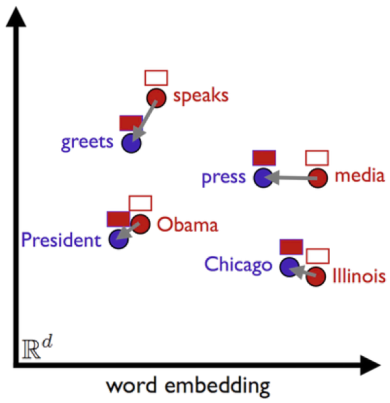
Bag of Words technique is the method that considers the occurrence of word in each sentence, but do not consider the order of words. For example, the sentence “John likes to

watch action movies. Mary likes movies too," can be written as a dictionary, a computable object, by {"John":1, "likes":2, "to":1, "watch":1, "movies":2, "Mary":1, "too":1}. This dictionary counts the occurrence of each word in this sentence. In this paper, each job description is considered as a bag of appearing words in the job description sentence.

c. Word2Vec

Word2Vec is a computational method that converts words that appeared in the job descriptions into vector, a mathematical object. The nearness of vectors depended on the meaning of the word vectors. The more nearness between vectors means the same meanings between word vectors. This concept is illustrated in Figure 3. The Word2Vec technique is chosen for all job descriptions to find all required skills in each job. In this paper, Google News pre-trained model is chosen for converting word into vectors (Google Inc., 2019). Each word vector in the model has 300 dimensions. In each job description, the weighted average method is selected to calculated of all word vectors (exclude stop words, words that do not have meaning in the sentence) that appeared in the job description, which is transformed into bag of word (section b). Calculated vectors are used for computation and clustering process in next sections.

Figure 3, Example of Word2Vec nearness of word vectors (Kernix, 2016)

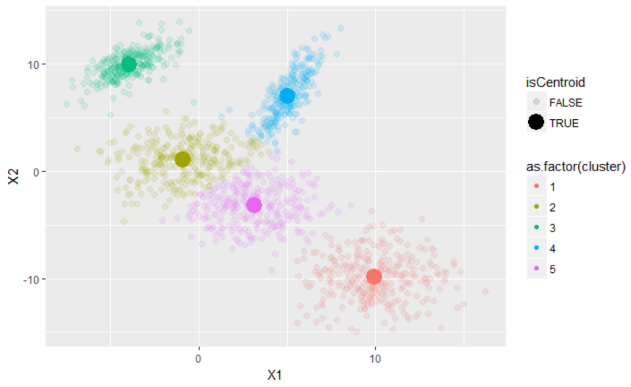


d. k-means clustering

In the clustering process, k-means technique is selected for clustering each job description vector. Figure 4 illustrated how k-means clustering vectors. In Figure 4, each small

dot represents vector in 2 dimensions X1, X2. K-means is the process of clustering all vectors in the 2-dimensional space by finding centroid points, in this case, there are 5 cluster points represented by different colours.

Figure 4, k-means clustering process (Antoine Guillot, 2017)



The concept of k-means is applied in this paper by clustering all job description vectors based on the meaning of words in job descriptions that appeared in vectors; however, the dimension of the word vector in the model is 300, so it is not possible to plot job description vectors.

This paper mainly uses a job description in the data pre-processing process because job descriptions reflect the required skillsets more precisely than job position names. Moreover, some different job descriptions use the same job position names.

4. Empirical Results

In this paper, the result is divided into 2 parts based on the different dataset, labour demand data from Jobtopgun and labour pricing data from Adecco.

a. Jobtopgun, the labour demand data

In this part, data from Jobtopgun are used to finding the insight of the labour market dataset into 3 ways - classification of job types from monthly data, number of jobs in each cluster overtime and WordCloud visualisation to show the skills that appeared in each cluster of jobs.

In the classification section, the December 2018 dataset of job description, the first scraped dataset, is clustered into 8 job clusters. The centroid of each cluster is calculated by the arithmetic mean method. These centroids in the December 2018 data is fixed for clustering process for next 3 datasets, January – March 2019 datasets. The reason to fix these centroids is to find the dynamic of labour demand. The clustering process for datasets in 2019 is choosing the minimum distance from each job description vector to each cluster point.

To visualise job clustering in each cluster, appeared words in job description vectors are used to show the most common words in each job cluster. In this paper, most common 10 words, except stop words, in each cluster are shown on Table 3. Judgment is used to identify the name of each cluster point based on the common words that appeared on each group. Names of each cluster point are in Table 4.

Table 3, Most 10 occurrence words in each cluster group

Group 0:	work, check, customers, care, customer, company, clean, assigned car service
Group 1:	maintenance equipment repair electrical control systems system work installation machine
Group 2:	customers sales customer company products new service product care assigned
Group 3:	control production quality work system project process design management plan
Group 4:	accounting tax check financial documents prepare accounts report assigned company
Group 5:	japanese design thai work assigned support media thailand company english
Group 6:	work assigned company management documents duties training control tasks coordinate
Group 7:	sales marketing company business customer customers new team management plan

Table 4, Job types based on clustering process

Group 0:	customers services
Group 1:	engineering
Group 2:	sales
Group 3:	production
Group 4:	accounting and finance
Group 5:	language works
Group 6:	office works
Group 7:	marketing

After naming and clustering job from job descriptions, the number of jobs in each cluster are on Table 5.

Table 5, The number of jobs based on clustering process

	Cluster Name	Dec 18	Jan 19	Feb 19	Mar 19
Group 0:	customers services	1446	1118	639	865
Group 1:	engineering	935	759	374	477
Group 2:	sales	1925	1590	1005	1198
Group 3:	production	2954	2578	2042	2182
Group 4:	accounting and finance	1189	981	704	794
Group 5:	language works	814	652	475	477
Group 6:	office works	2202	1930	1434	1585
Group 7:	marketing	2099	1810	1608	1707
	Total	13564	11418	8281	9285

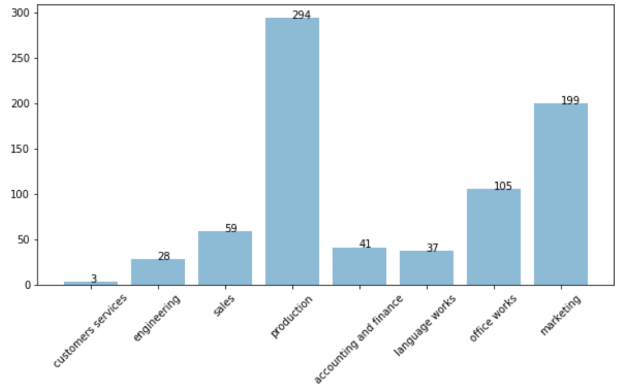
In this paper, WordCloud chart, the visualisation of all words in each cluster based on the frequency of words in each cluster, is selected to illustrate. The larger size of the word means the more frequent of words in the cluster. In Figure 5, showing the WordCloud plot for cluster group 3, “production” jobs. The word “system” is the most appeared word in this cluster due to the largest size in this chart. In this figure show that the main skills in the “production” group is the word in the WordCloud plot.

Figure 5, The main skill set in word cloud plot for "production" jobs in December 2018.

b. Adecco Salary Guide

In this part, job centroid points in part a. are used to cluster all jobs in Adecco data. The norm of each job description vector to the cluster point is calculated. These distance data (norm) are used to calculate wage based on skill set from each job description. Numbers of jobs based on 8 clusters on part a are shown in Figure 6.

Figure 6, The number of jobs based on 8 clusters



With 766 observations based on job descriptions and work experience, the regression equation is set by the assumption that wage is depended on two main parts, the skillset of each job and the experience of labour.

First, the skillset of each job, after converting each job description into job vector. The distance from job description vector to each cluster point is considered as the skillset required on each job. The more distance from each job description vector to the cluster point means the farness from each observation to the cluster point. The cluster points are defined as the skillset that used in that job cluster. For example, a job that has a low level of distance to the cluster point means that the job requires a high level of skill in the cluster point.

Second, the experience of labour, the dummy variable vector that shows work experience. Because Adecco divided jobs into 3 groups, new graduate level, 1-5 years of experience level and more than 5 years of experience level. This paper name these variables “newgrad”, “junior”, and “senior” respectively.

Based on the assumption that stated above, the regression equation is set based on the log-linear Ordinary Linear Regression (OLS) to find the relationship of the percentage

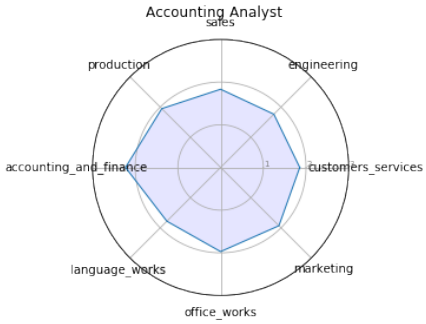
change in distance of the job description vector to each cluster point (norm of a vector) as shown in Equation 1.

Equation 1, Log-linear model of wage and skill in labour pricing data

$$\ln(wage) = f(dist_0, \ldots, dist_7, D_{newgrad}, D_{senior})$$

To illustrate this concept, the chart of the level of required skill set in each job cluster which equals 1 – (the normalise distance between job vector to the cluster point)^0.10. In Figure 7 show the skill set of Accounting Analyst based on 8 types of skills, based on the first clustering in Jobtopgun dataset.

Figure 7, The skill set plot of Accounting Analyst based on main skills



In Figure 7, the “Accounting Analyst” requires most of “accounting_and_finance” skill sets due to the least distance between this job vector to “accounting_and_finance” cluster that means the most radius of “accounting_and_finance” in the spider chart in Figure 7.

The regression result is shown in Figure 8.

Figure 8, The summation of OLS Regression

Dep. Variable:	wage	**R-squared:**	0.694
Model:	OLS	**Adj. R-squared:**	0.690
Method:	Least Squares	**F-statistic:**	171.4
Date:	Sun, 12 May 2019	**Prob (F-statistic):**	1.14e-186
Time:	00:22:34	**Log-Likelihood:**	-325.32
No. Observations:	766	**AIC:**	672.6
Df Residuals:	755	**BIC:**	723.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.0472	0.747	14.789	0.000	9.581	12.514
customers_services	-15.7890	3.032	-5.208	0.000	-21.741	-9.837
engineering	10.5283	2.250	4.678	0.000	6.110	14.946
sales	-7.8899	2.257	-3.495	0.001	-12.321	-3.458
production	-10.1945	2.465	-4.136	0.000	-15.034	-5.355
accounting_and_finance	3.6898	1.524	2.422	0.016	0.699	6.681
language_works	-2.9823	1.789	-1.667	0.096	-6.494	0.530
office_works	2.4608	2.308	1.066	0.287	-2.070	6.992
marketing	24.7426	3.062	8.082	0.000	18.732	30.753
newgrad	-0.5284	0.039	-13.704	0.000	-0.604	-0.453
senior	0.7092	0.031	22.678	0.000	0.648	0.771

Omnibus:	56.065	**Durbin-Watson:**	1.535
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	67.712
Skew:	0.666	**Prob(JB):**	1.98e-15
Kurtosis:	3.587	**Cond. No.**	376.

From the regression result, the coefficient beta represents the percentage change for the skill set in the cluster of jobs has more or less wage than the average wage of all observations. More magnitude of beta means the percentage change of the distance between job and cluster point (more distance from centroid point means less of skill set required in that job type). Because of the inverse relationship between the distance from job description vector and the required skill, -1 is multiplied to the distance in this regression for the ease of interpreting beta coefficient as shown in Equation 2.

Equation 2, The coefficient interpretation

$$\beta = \frac{d\ln(wage)}{d(distance)}$$

The regression equation is shown in Equation 3. Skills in this regression equation are the negative of distance from job description vectors to each cluster point on Table 4.

Equation 3, The regression equation on wage and skills

$$\begin{aligned}\ln(wage) = {} & 11.0472 - 15.7890skill_0 + 10.5283skill_1 \\ & - 7.8899skill_2 - 10.1945skill_3 + 3.6898skill_4 \\ & - 2.9823skill_5 + 2.4608skill_6 + 24.7426skill_7 \\ & - 0.5284D_{newgrad} + 0.7092D_{senior}\end{aligned}$$

The result is shown in 2 parts. First, the relationship between the skill set and wage. The result shows that the coefficient in marketing group is the highest coefficient, 24.7426, means a job that have the marketing skill set gains most wage than other jobs with other skills. Meanwhile, the "customers_services" gains least wage than other jobs with other skills. The coefficient in this part is -15.7890 which is the lowest coefficient.

In the dummy variable analysis, the controlled group is the job that requires at least 1 year but less than 5 years of experiences. From the coefficient in "newgrad" – a group that the work experience less than 5 years, it shows that the wage is reduced by 52.84% with respect to the controlled group. In the "senior" group – a group that requires more than 5 years of work experiences, it shows that the wage is increased by 70.92% with respect to the controlled group.

5. Limitation

This paper only uses 2 sources for tracking Thai labour demand and labour pricing which might not cover all of Thai labour market. Adecco Thailand Salary Guide is the only public data that have salary data in Thai labour market; however, jobs in Adecco does not fully cover all jobs in the market.

By economic theory, in labour pricing part, there are a lot of factors that determined wage, for example, age, education level. This paper only uses data which provided from

scraping technique, distance between job description vector to cluster point and dummy variable data.

6. Summary

The key technique that uses in this paper is Word2Vec, a branch of Natural Language Processing (NLP) by converting all texts into mathematical objects and using them in machine learning part. After converting text, especially in the job description, k-means clustering is chosen to cluster all jobs by searching hidden skills which embedded in the job description. The skills in each job type is found by using NLP techniques. Moreover, the econometric method is selected to find that how each skill set earned wages. The regression equation shows that the marketing skill sets gain the most wage with respect to other job skills while as the customer service skill sets gain the least wage with respect to other job skills and a lot of experience in working has a positive relationship with salary.

7. Code Licence

All codes and documentations are under the MIT Licence and can be accessed at github.com/sikawit/EconSR.

8. Acknowledgement

I express my sincere respect and gratitude to my advisor, Tanapong Potipiti, Ph.D, Assoc. Prof. who has given his valuable supports and suggestions to successfully complete this senior research. I also owe special thanks to Wasawat Somno, ThoughtWorks Thailand for inspiring me in computer programming. I would like to thank my family, who always support and trust me for 4 years at the Faculty of Economics, Chulalongkorn University.

Bibliography

Adecco Group Thailand. (2019). *Adecco Thailand Salary Guide 2019*. Retrieved from https://adecco.co.th/salary-guide/2019/00004

Antoine Guillot. (2017, October 27). *Machine Learning Explained: Kmeans*. Retrieved from Enhance Data Science: http://enhancedatascience.com/2017/10/24/machine-learning-explained-kmeans/

Google Inc. (2019). *Google word2vec*. Retrieved from https://code.google.com/archive/p/word2vec/

Kernix. (2016, September 26). Retrieved from Similarity measure of textual documents: https://www.kernix.com/blog/similarity-measure-of-textual-documents_p12

Nuarpear Warn Lekfuangfu, V. N. (2016). *Labour Market Insights: the Power of Internet-Based Data.* Bangkok: Puey Ungphakorn Institute for Economic Research, Bank of Thailand.

Top Gun Co. Ltd. (2019). *Joptopgun*. Retrieved from https://www.joptopgun.com

Appendix: MIT Licence