SIKKAWIT AMORNNORACHAI
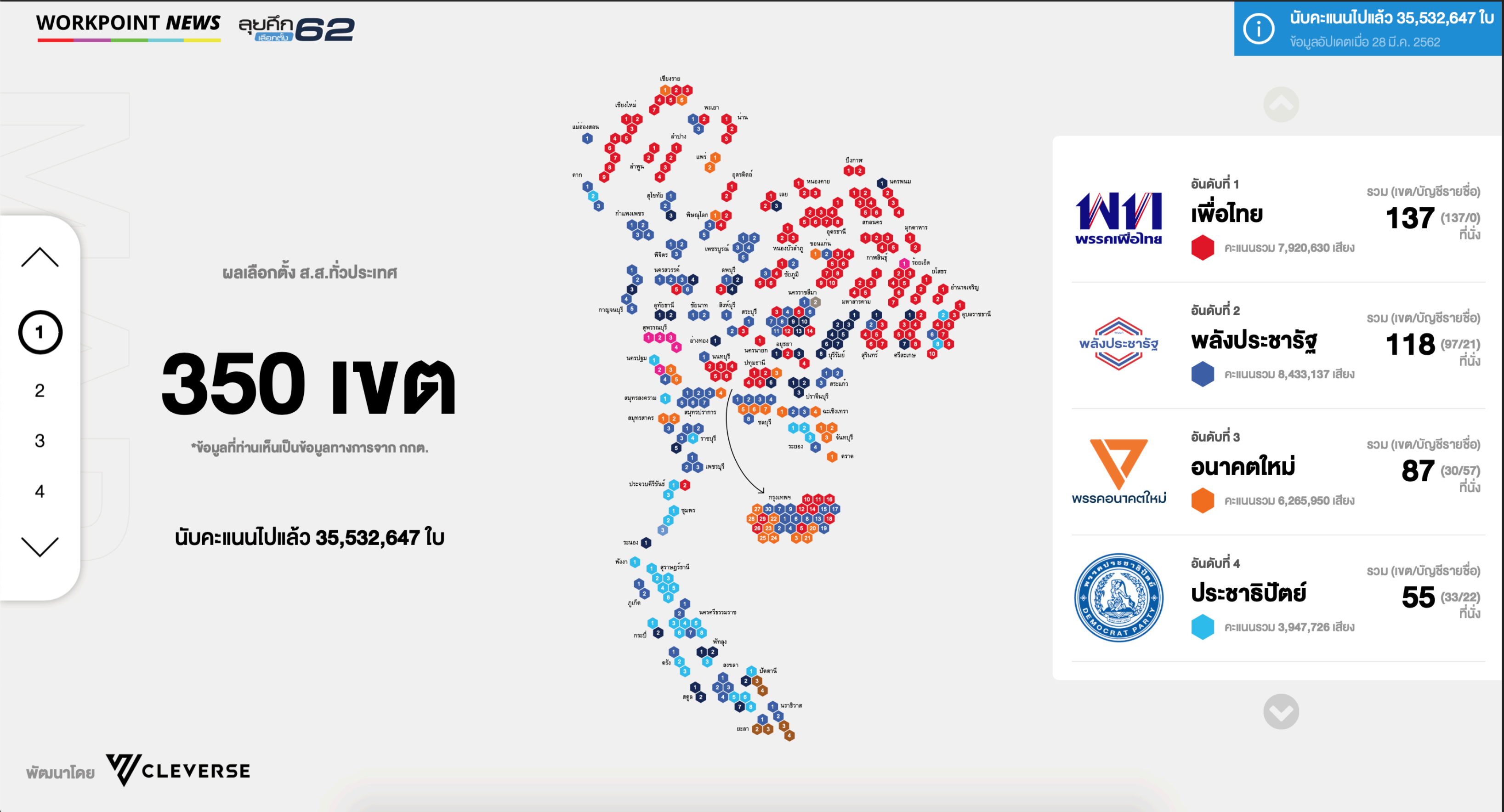
THE TRACKING SKILLS OF THAI LABOUR MARKET USING UNSUPERVISED MACHINE LEARNING WITH INTERNET DATA

OUTLINE

▸ Introduction

▸ Data Collecting Processes

▸ Techniques in this study

▸ Labour Demand Data from *Jobtopgun*

▸ Labour Pricing Data from *Adecco*

▸ Summary

DATA VISUALISER - A BRAND NEW JOB

https://vote.workpointnews.com

INTRODUCTION

- In the era of disruption, a lot of jobs are out of the market, but a lot of new jobs are created.

- Data in Labour Market is rich in text, but it is not fully utilised.

- The first paper using text analysis in studying Thai Labour Market

- A lot of drawbacks in traditional study techniques

COMPARING 2 WAYS OF STUDYING

Topics	Traditional Study	Text Mining
Source of Data	Labour Force Survey	Online Dataset
Availability	Require Permission	Public Data
Cost to access	Not Free	Free
Time to get access	Few weeks	Immediately
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

APPLICATION OF TEXT LABOUR DATA

▸ Realtime labour demand over time

▸ Measuring set of skills in a job

▸ Skills clustering based on skill set

▸ Planning curriculum for necessary skill sets

▸ Analysing wage based on skill set

THE INTERNET DATA

▸ Labour Demand

   ▸ Jobtopgun Job Postings Data

   ▸ One of the largest Job Posting in Thailand

   ▸ Monthly scraping from December 2018 to March 2019

   ▸ About 8,000 to 13,000 jobs per month

THE INTERNET DATA

▸ Labour Wage

  ▸ Adecco Thailand Salary Guide 2019

  ▸ Only source for public available data

  ▸ 766 jobs

IN THE DATA

▸ Jobtopgun
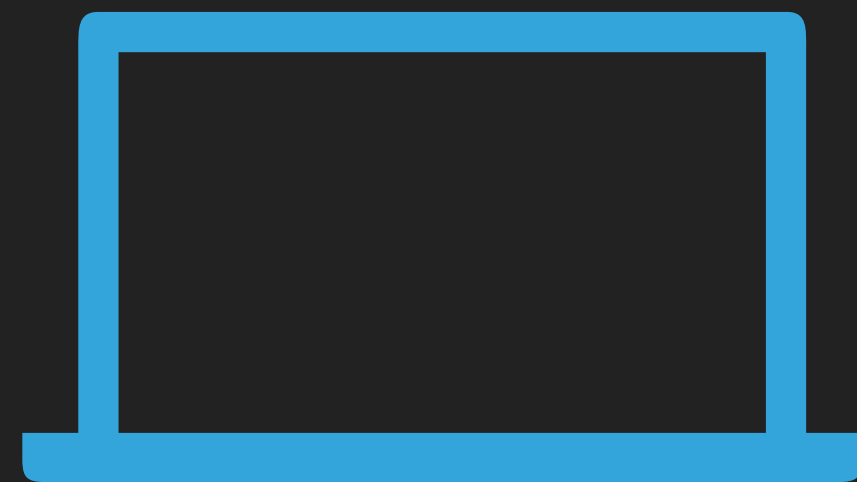
  ▸ Position Name

  ▸ Job Description

▸ Adecco

  ▸ Position Name

  ▸ Job Description

  ▸ Salary

  ▸ Year of Experience

DATA COLLECTING PROCESS

Internet
Data

Python
Scraping

NLP - Text Processing

Data Analytics

Joptopgun
Adecco

Collecting Data
from Web

Bag of Words,
Translating, Word2Vec
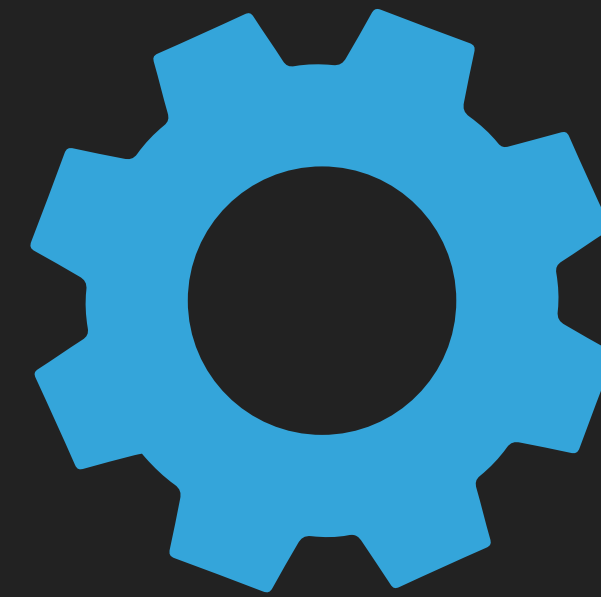
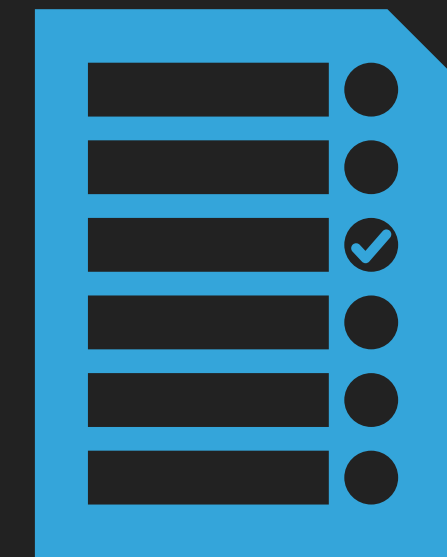Visualisation,
Clustering,
Regression

NLP - TEXT PROCESSING

Translation
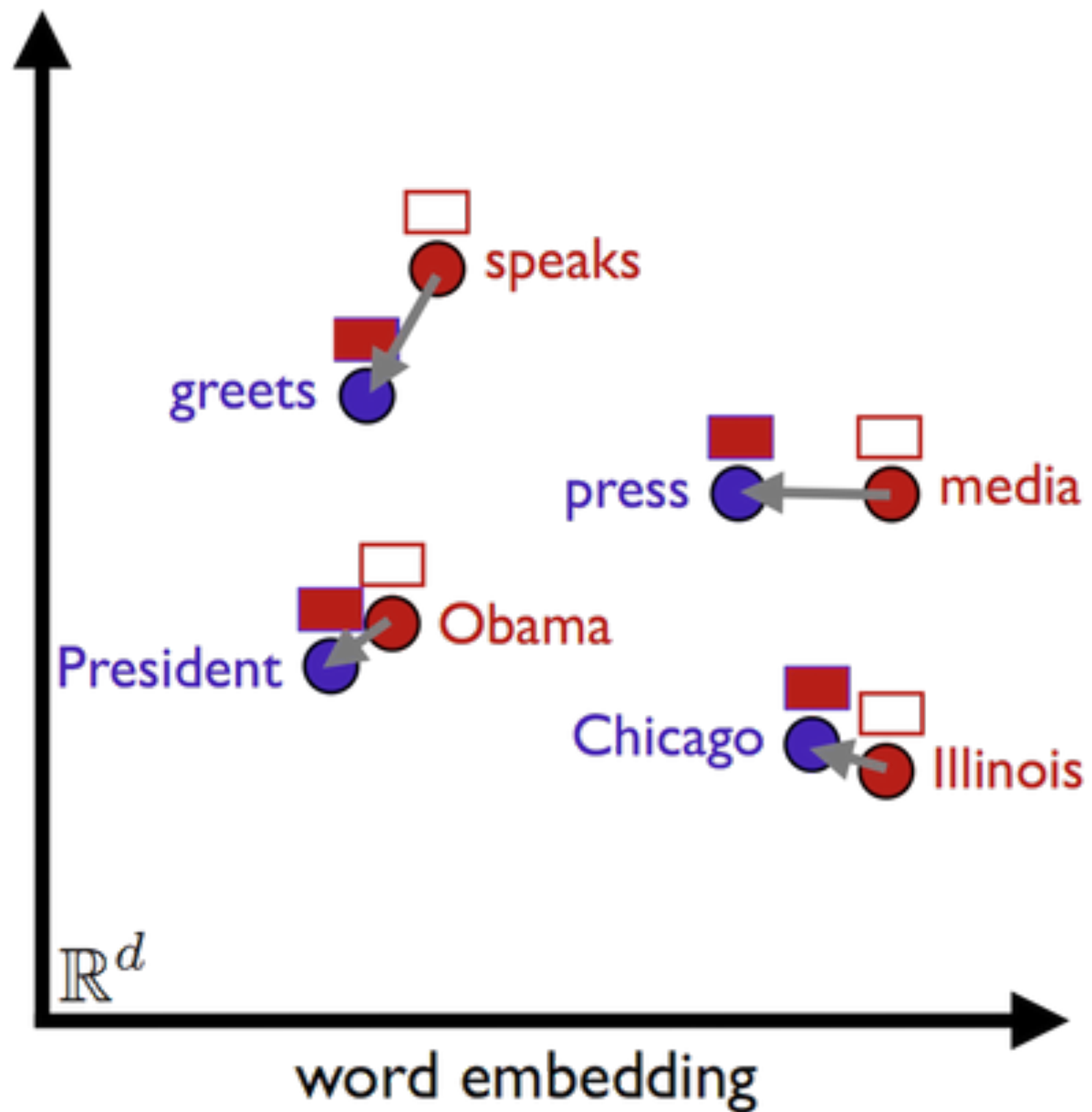
Bag of Words

Word2Vec

k-mean
Clustering

TRANSLATION

▸ Job description is a required skill in a job

▸ Data in Job Description are in Thai and English

▸ Converting all Thai Text into English for analysing process

▸ Using Bag of Words Technique, we do not consider the grammatical order in the translation result

▸ Google Translate Service

BAG OF WORDS
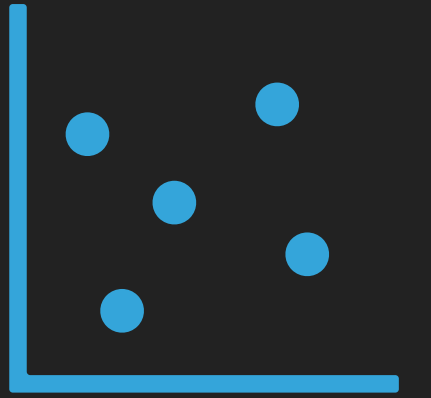
▸ Considering each word in each job description then collect into a bag

▸ Does not consider the order of words, consider the occurrence of words

▸ "John likes to watch movies. Mary likes movies too."
== {"John":1, "likes":2, "to":1, "watch":1, "movies":2, "Mary":1, "too":1}

WORD2VEC

https://www.kernix.com/blog/similarity-measure-of-textual-documents_p12

WORD2VEC

▸ Converting words into vectors

$$cat = [0.2, 0.4, 0.5, ..., 0.6]$$

▸ Using Google News Model to convert words in job postings into vectors

▸ Nearness of vector based on the meaning of 2 vectors

▸ In each job description, I calculate the mean of each word vector

WORD2VEC

▸ Example: King + Women = Queen
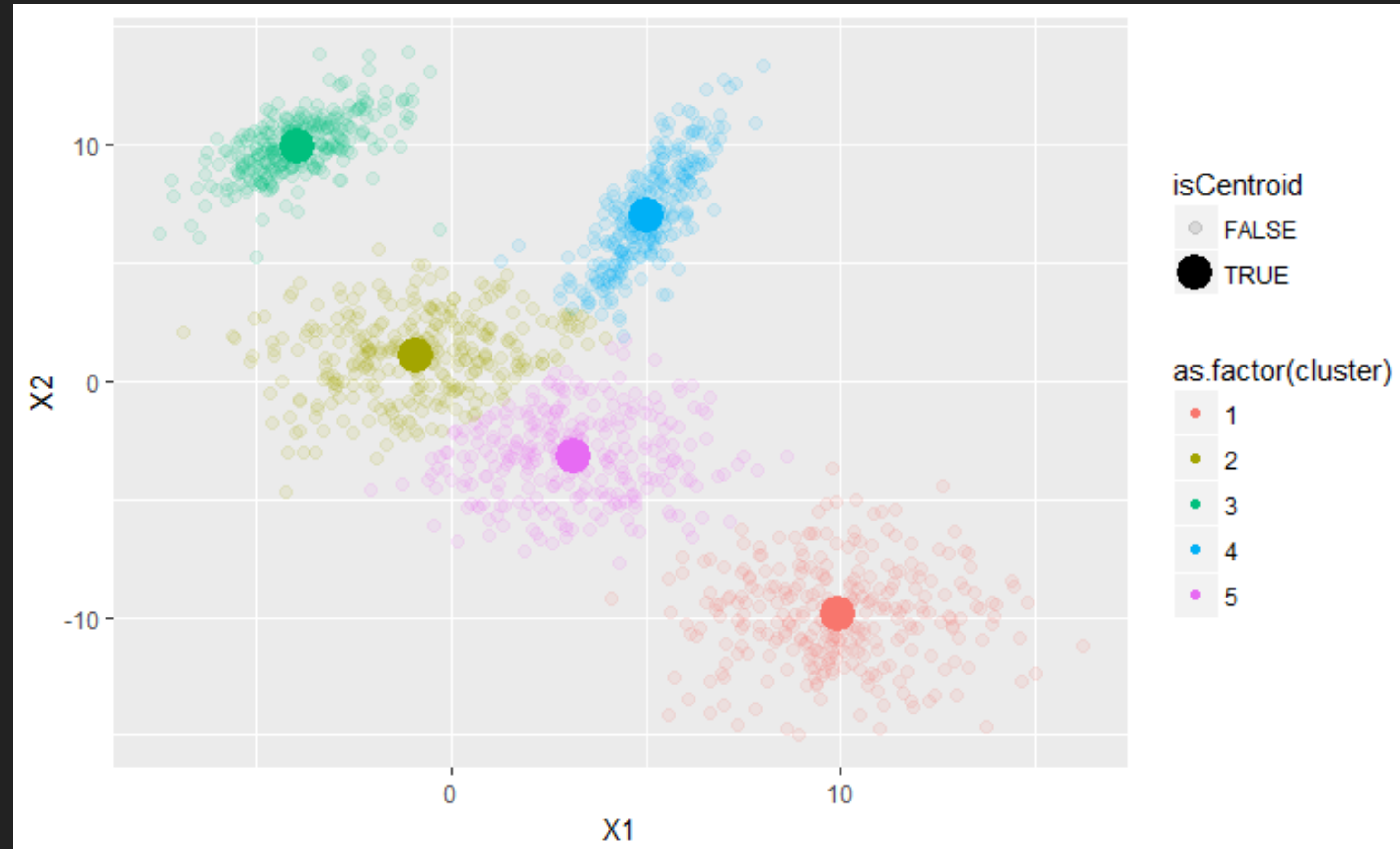
$$king = [3,4,2,6]$$

$$woman = [1,4,2,7]$$

$$queen = [4,8,4,13]$$

K-MEANS

▸ Clustering job by required skill set in the job description

▸ Unsupervised Technique

▸ Machine Learning technique for clustering analysis

▸ K groups of clustering

▸ Cluster by the distance of each vectors

K-MEANS

http://enhancedatascience.com/2017/10/24/machine-learning-explained-kmeans/

JOBTOPGUN

LABOUR DEMAND

IN THE DATA — JOBTOPGUN

DATA FROM JOBTOPGUN

Job Position	Job Description
MIS Support (annual contract)	Supporting documentation, including documentation of the audit department to present the data and verify the accuracy of the information in the system. Including analysis for the department to prepare weekly reports. Monthly coordination with other agencies as assigned.

METHODOLOGY

▸ First, I use all job description vectors in December 2018 and k-means technique to cluster job types.

▸ In next 3-month data, I use the previous cluster point to cluster job description vectors. Cluster of group is selected from the least distance from job description to each cluster.

▸ Then, I use the cluster data to visualise data

OBJECTIVES & VISUALISATION

▸ Cluster Topic Name and most occurrence word

▸ Wordcloud - "words and skills" in each cluster

▸ These techniques to find the skills set in each job description

GROUP OF SKILLS

0: 'customers services',

1: 'engineering',

2: 'sales',

3: 'production',

4: 'accounting and finance',

5: 'language works',

6: 'office works',

7: 'marketing'

SKILL CLUSTERING BY K-MEANS (8 GROUPS)

Main Skills		Dec 18	Jan 19	Feb 19	Mar 19
Group 0:	customers services	1,446	1,118	639	865
Group 1:	engineering	935	759	374	477
Group 2:	sales	1,925	1,590	1,005	1,198
Group 3:	production	2,954	2,578	2,042	2,182
Group 4:	accounting and	1,189	981	704	794
Group 5:	language works	814	652	475	477
Group 6:	office works	2,202	1,930	1,434	1,585
Group 7:	marketing	2,099	1,810	1,608	1,707
Total		13,564	11,418	8,281	9,285

WORDCLOUD – GROUP 1 (ENGINEERING), DEC 2018

WORDCLOUD – GROUP 4 (ACCOUNTING & FINANCE), DEC 2018

Group 4 accounting and finance

CLUSTERING RESULT - DEC 2018

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

SUMMARY — LABOUR DEMAND

▸ Understand the basic set skills in each cluster

ADECCO SALARY GUIDE

SKILL PRICING

OBJECTIVE

▸ There are 8 main skills in Thai labour market

▸ The main question is the revenue earned from each skill

▸ I want to check that which skill set create a lot of wage based on Adecco
Thailand Salary Guide

▸ I use the linear regression technique to solve this problem

IN THE DATA — ADECCO

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

WHAT TO DO
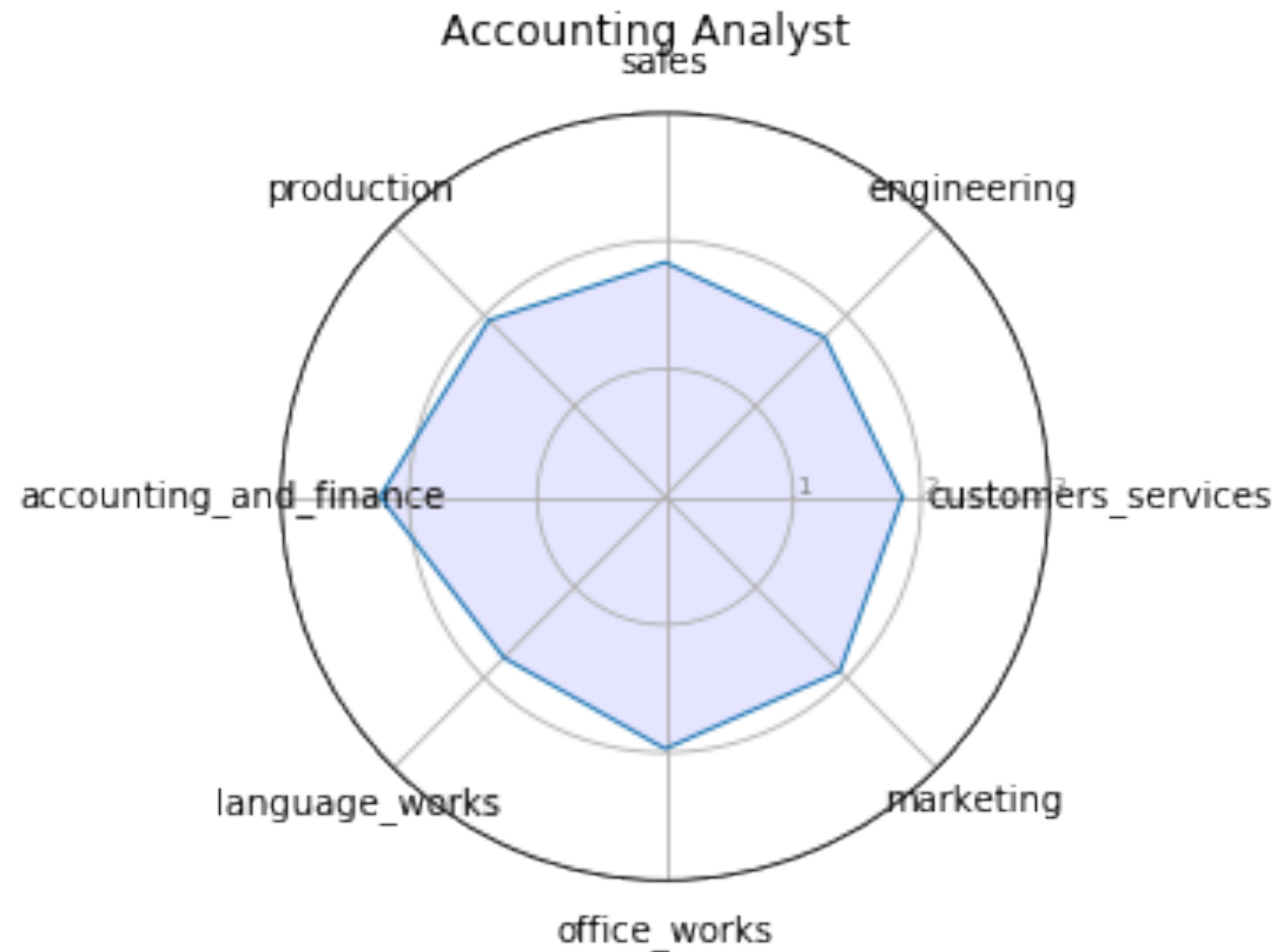
▸ In this part, I use the Adecco Salary Guide 2019 Thailand to analyse the regression equation

▸ Data that depended on work experience and the distance of job description vector to the cluster point that I calculated in the earlier part.

▸ I construct the regression equation

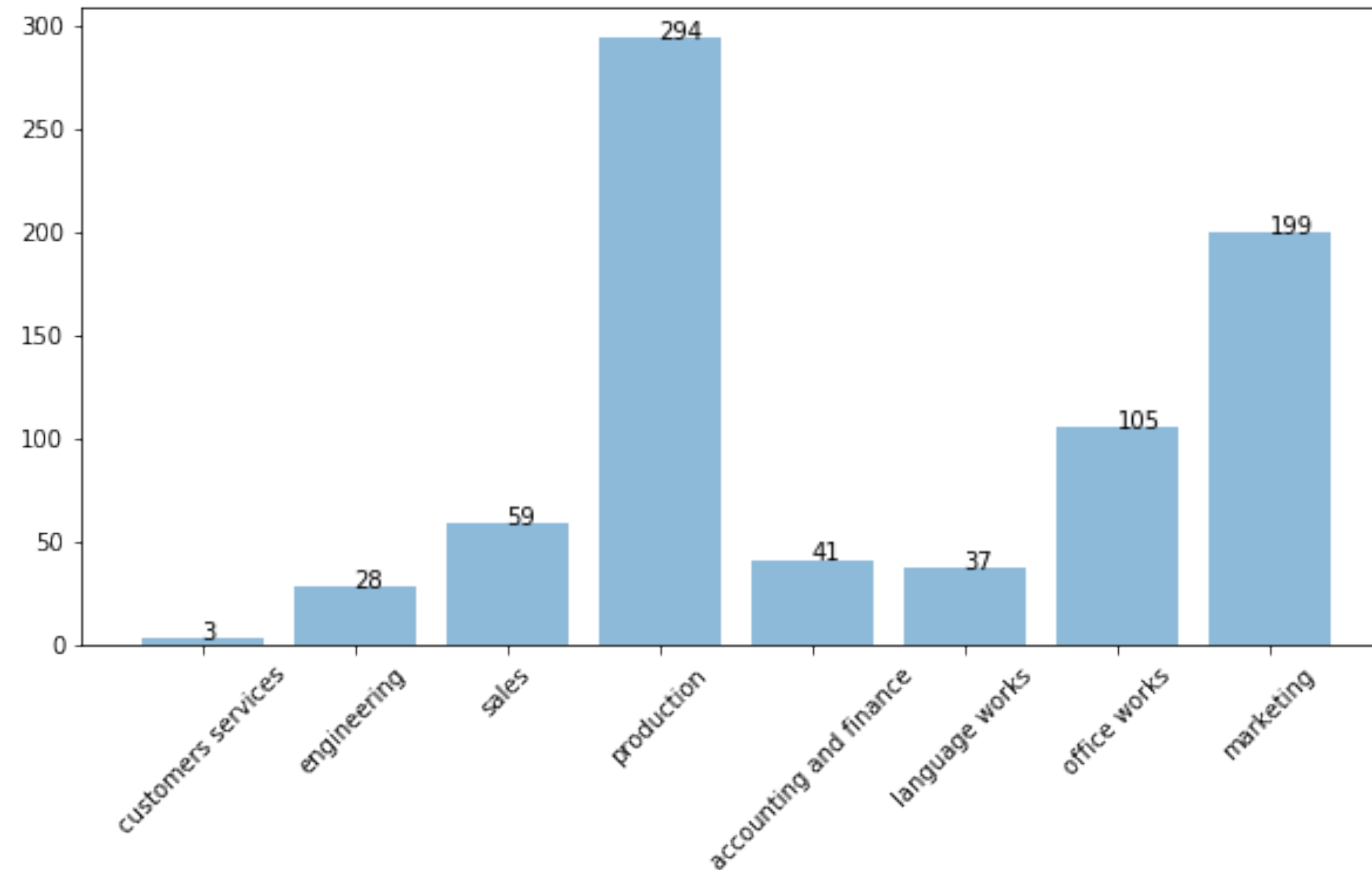$$\ln(wage) = f(dist_0, \ldots, dist_7, D_{newgrad}, D_{senior})$$

IN THE DATA - ADECCO

Job Position	Job Description	Wage
Accounting Analyst	Perform duties related to fixed-asset accounting, bank reconciliation, cost allocation, account closing, financial accounting and balance sheet accounts reconciliation. Review and monitor accounting controls to ensure accurate and timely financial records.	40,000

SPIDER PLOT OF JOB BASED ON SKILLS

MAIN SKILLS IN ADECCO JOBS

REGRESSION - WAGE ON SKILLS

▸ Assumptions:

    ▸ Set of skills determine wage

▸ Skills clustered from job descriptions

▸ Using the distance of datapoint to cluster vectors as the regressor

▸ $\ln(wage) = \beta_0 + \beta_1 Skill_1 + \ldots + \beta_i Skill_i + \beta_{newgrad} D_{newgrad} + \beta_{senior} D_{senior}$

REGRESSION – WAGE ON SKILLS

$$\ln(wage) = f(dist_1, \ldots, dist_8, D_{newgrad}, D_{senior})$$

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.

RESULT

▸ The marketing skill set obtain most percentage change wage from all observations (24.2476)

▸ The customer services skill set obtain least percentage change wage from all observations (-15.7890)

▸ Ranking based on coef column in the regression table

RESULT

▸ In this paper, the controlled group is the 1-5 years of working experience group

▸ The dummy variable for no working experience wage is reducing 52 percent from the controlled group

▸ The dummy variable for more than 5 years of experience wage is increasing about 70 percent from the controlled group

LIMITATIONS

▸ In this paper, I only use 2 data sources to tracking skill set.

▸ In the labour wage part, by economics theory, there are more dependent variables that determined wage, for example, age, education level, etc.

SUMMARY

▸ Internet data & text analytics

▸ Finding main skills required for labour market

▸ Finding price and importance of skills

FEEL FREE
TO ASK ME

Made on a Mac, and Python

LICENCE