Detecting New and Emerging Occupations from Online Job Data

MINISTRY OF MANPOWER

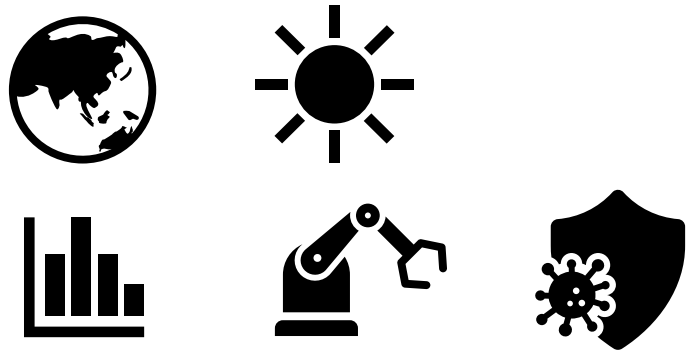
Sim Tze Wei, Daniel

Manpower Research and Statistics Department

INTRODUCTION

Needs of the labour market are constantly evolving due to:

- Globalisation
- Climate Change
- Demographic Shifts
- 4th Industrial Revolution
- Covid 19 Pandemic



Considering rapid technological advancement and availability of large amounts of traditional and non-traditional data sources,

It is crucial to explore new avenues for keeping ISCO up to date

What do we want to achieve?

Identify new and emerging occupations based on data from online job advertisements

Data Preparation

DATA COLLECTION

- Web-scraped general and niche job portals (eg welfare related from Idealist, technology from Nodeflair, Various jobs posting from JobStreet, etc)
- Data sharing partnership arrangement between the International Labour Organization (ILO) and Uruguayan online job board BuscoJobs
- Data ranged from May 2022 to January 2023
- Close to 2 million job postings

DATA CLEANING

- Categorised job posting using *langdetect* library in Python using the first 50 words of the job description
- Only considered online job advertisements in English
- Processed dataset consists of >1.6 million job postings
- Creation of
  - Dataset with distinct job titles and description
  - Dataset with distinct job titles

Data Process

Processed dataset of > 1.6 million job postings (OJA)
(Busco Jobs and Jobstreet)

Supervised Learning Techniques

Data collected through Labour Force and Establishment Surveys (mapped SSOC with ISCO at 4 Digit)

Distinct Job Titles and Job Descriptions from OJA

Split into 80% training and 20% test
Pre-train the BERT to determine the ISCO 4 Digits code for the 20% (Avg F1 score ~ 0.56)

Obtain ISCO 4 Digit with the highest probability through the pre-trained BEST model

Unsupervised Learning Techniques

Distinct Cleaned Job Titles from OJA

Word and Sentence Embedding

(Using Fasttext – continuous bag of word, out of vocabulary and morphology that helps in our use case)
Eg, Ice, Icy, word such as 'Where' to consider 'Whe', 'her', etc

Dimension Reduction and Clustering

UMAP – further improve the performance and accuracy of clustering algorithm
HDBSCAN – Form the Occupation Cluster that are largely similar and obtain Tri-gram

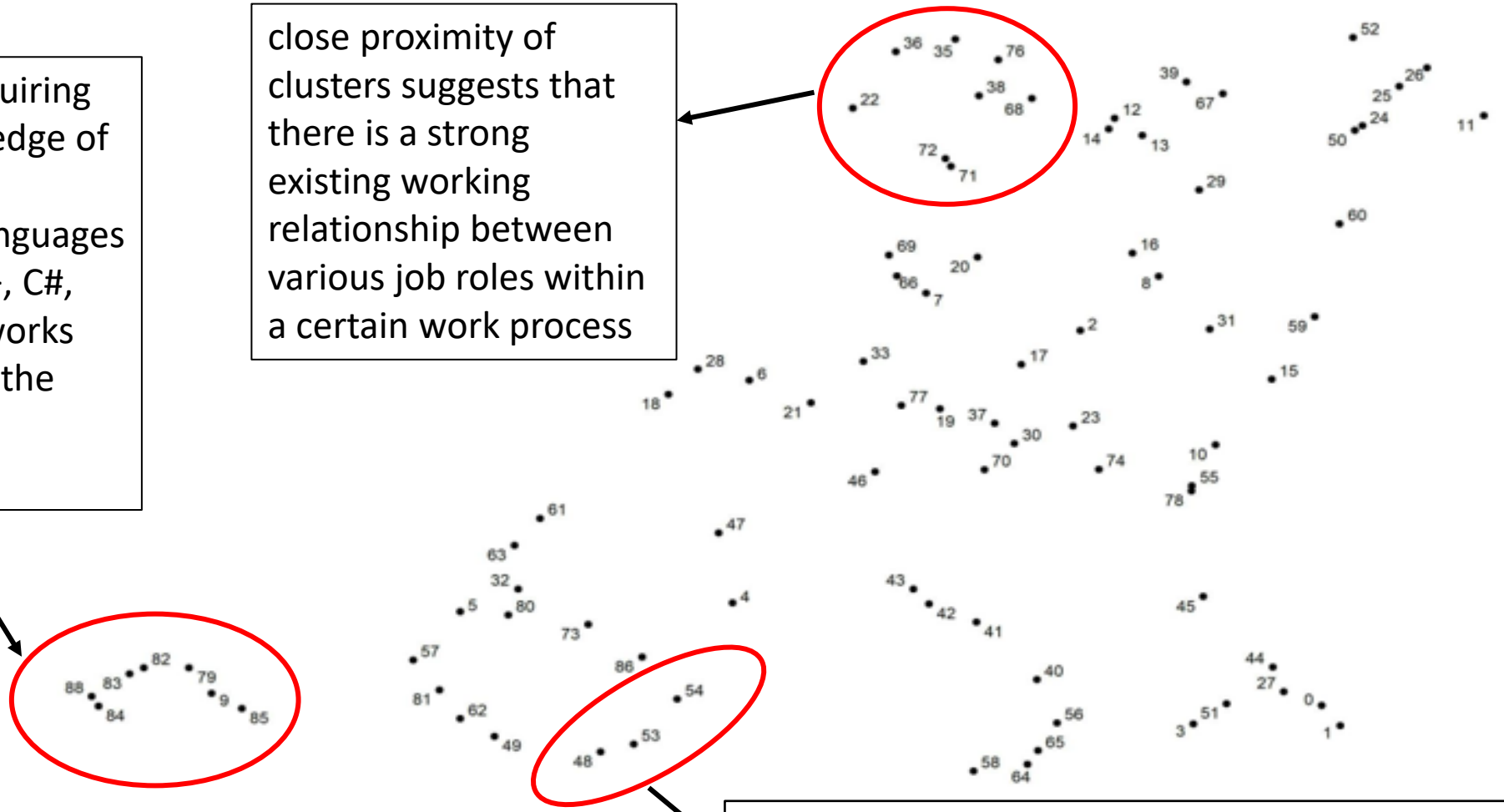
Revealed Comparative Advantage and Average probability

Compute for each tri-gram the average score for RCA
Compare each tri-gram across other clusters and compute the GNA score

Visualising Cluster Centroids – Unsupervised Learning

Occupations requiring technical knowledge of general-purpose programming languages such as PHP, C++, C#, Java and frameworks such as .NET for the development of softwares

close proximity of clusters suggests that there is a strong existing working relationship between various job roles within a certain work process



occupations related to software testing such as user acceptance testing (UAT) and quality assurance testing (QAT)

Data Process – Identifying New and Emerging Occupation Index (NEOI)

Identifying New and Emerging Jobs through
Computation of New and Emerging Occupation Index (NEOI)

New and Emerging Occupations index (NEOI) to blend both techniques

1. Normalised score for all tri-gram within each cluster
2. Normalised score for global cluster
3. Average Probability sub-measures from the occupation index through supervised classification model

Filtering tri-grams >= 85th percentile of NEOI score

25,707 candidates

Filtering candidates with a low CS score in ISCO with a high CS score in at least 2 other NOCs/ROCs

Comparison against NOCs/ROCs to further reduce the number of candidates

288 candidates

FINDINGS

25,707 candidates

Comparison against NOCs/ROCs to further reduce the number of candidates

International Standard Classification of Occupations, 2008 (ISCO-08)

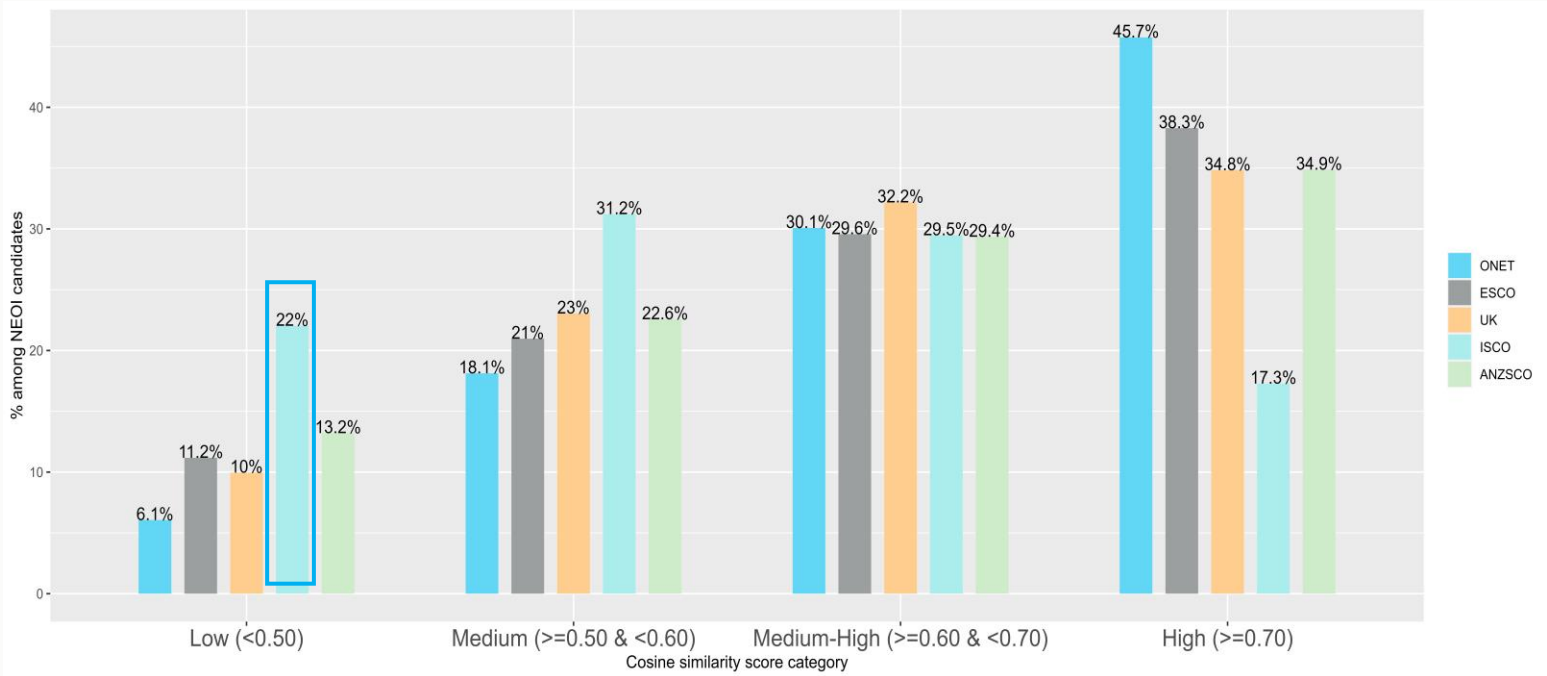
O*NET

ESCO European classification of Skills, Competences, Occupations and Qualifications

Stats NZ Tatauranga Aotearoa

Office for National Statistics

Ensures candidate surfaced is not country nor region specific and has been identified in at least 2 other NOCs/ROCs



- The probability of finding a job title similar to our candidate is higher in some NOC/ROC
- For ISCO, more than 50% of candidates are classified in the low to medium cosine-similarity score (CS score) category
- We focus on the 5,655 (22%) of candidates with a low CS score in ISCO. Filtering those with a high CS score in at least 2 other NOCs/ROCs yielding 288 candidates.

FINDINGS

288 candidates

ICT professionals

38.9%

Business and administration professionals

16.8%

Science and engineering professionals

15.0%

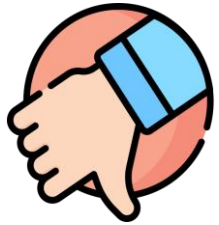
Examples of candidates

- Artificial intelligence product owner
- Cloud network engineer
- Cyber threat hunter
- Customer success advocate
- Scrum master
- User experience designer

EVALUATION OF STUDY

Benefits

- Data from online job postings is **near real-time**, avoiding time gaps associated with traditional data sources such as labour force surveys which are typically conducted quarterly or annually.
- Data from online job postings have **high degree of granularity** made possible by leveraging on technological solutions like the NLP models and interactive user interfaces.
- Data from online job postings contain **rich source of information** on types of skills and tasks that helps in the some related parts of the ISCO analysis



Limitations

- Data from online job postings may cause an **issue of representativeness** where certain sectors and occupations are over/under-represented.
- Some job postings advertised in print media or through word of mouth may also **not be captured**.
- There is a **lack of time series data** as only a snapshot of the existing stock of job advertisements is considered.
- Job titles in the data **do not fully capture one's roles and responsibilities**.

MINISTRY OF MANPOWER

Thank you

www.mom.gov.sg
SGMinistryofManpower
SGMinistryofManpower
MOMsingapore
Ministry of Manpower

A Great Workforce A Great Workplace