Ohio Department of Job and Family Services - Office of Workforce Development
Ohio Education Research Center

Department of Job & Family Services

Labor Market Research: Technician Jobs

Zachary Howes

Background Information

- **U.S. Bureau of Labor Statistics Standard Occupational Classification (SOC) System:** A system that classifies workers in the United States into 867 detailed occupations. This was last updated in 2018.
- **Technician SOC Codes:** A list of 30 SOC codes created by JobsOhio to identify technician jobs that are currently being prioritized in Ohio.
- **Natural Language Processing (NLP):** A subfield of computer science that helps computers to understand, process, and interpret human language.
- **Term Frequency-Inverse Document Frequency (TF-IDF) Matrix:** A table that shows how important each term is in each document by giving higher scores to words that appear often in one document but not in many others.
- **Cosine Similarity (In a TF-IDF Matrix):** A way to measure how similar two documents are by comparing the patterns of their word importance from the TF-IDF Matrix. If the patterns point in a similar direction, the documents are alike.
- **Logistic Regression:** A statistical method that predicts the chance of something binary happening (0 or 1/ yes or no) based on one or more input factors, using a formula that calculates the odds.

Data Source

- **TalentNeuron Data:** A private research company that collects job ad postings from data sources including government reports, research networks, job boards, trade publications, candidate profiles, and proprietary databases. This project utilizes raw posting data as well as the TalentNeuron created Skill & Credential categories where TalentNeuron uses Natural Language Processing to tag skills and credentials onto postings. This is done using taxonomies of over 30,000 skills and over 5,000 credentials.
- This data has not been seasonally adjusted and is a snapshot in time that may not represent long-term trends.

Research Questions

1. What are the most important skills & credentials for technicians?
2. Can the Ohio Department of Job and Family Services (JFS) develop customized natural language processing tools to better extract job ad data compared to available commercial tools?

Technicians Dashboard

A Tableau dashboard was created using data from TalentNeuron's internally created skill list and credential list over the past 18 months (January 1, 2024 – June 30, 2025) for the 30 technician SOC codes. This dashboard is intended to help job seekers interested in these occupations better understand the requirements for these jobs, as well as look at geographic trends to assist them in their job search.

Data Deduplication and Normalization in R

Before this data was uploaded into Tableau, it needed to be deduplicated and normalized. This ensured that repeat postings that TalentNeuron did not remove were caught, so that they did not impact results.

After running the deduplication script in RStudio, the results showed that there were 430 duplicate job postings removed from the dataset, resulting in 29,978 unique job postings for analysis from the date range of January 1, 2024 - June 30, 2025.

The data then had to be normalized using another R script with the skills & credentials columns both individually expanded and then recombined into the dataset based on Unique ID's. This ensured the data was in the easiest and fastest to load format for use in Tableau.

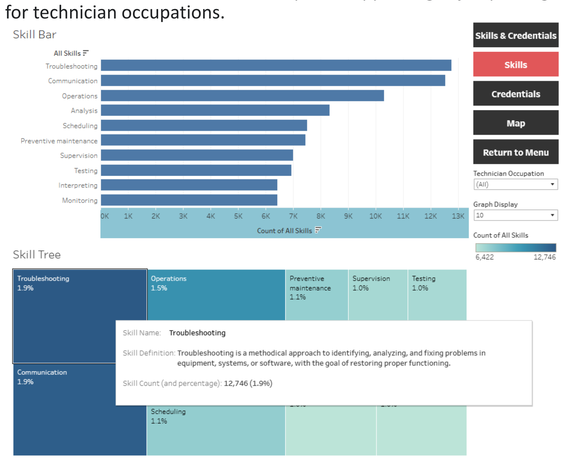
Dashboard Display Information

This dashboard includes a variety of filters and parameters. All graphics on this poster show the top 10 counts for all technician occupations over the 18-month analysis period.

In the dashboard, users can customize the counts and occupations for all visualizations and can also adjust the timeframe shown in the map.

Skills Results

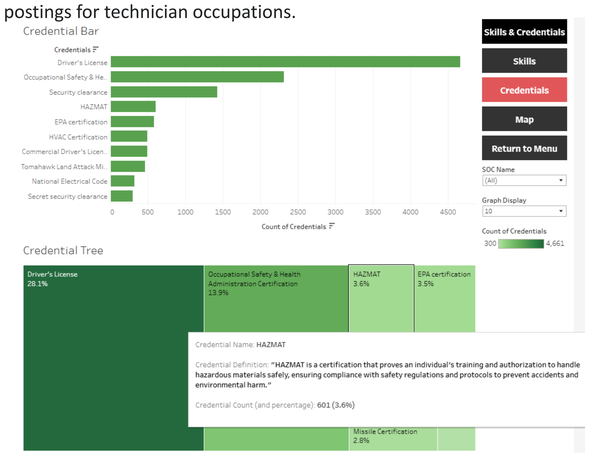
Customizable visualization of the top skills appearing in job postings for technician occupations.



Above is the skill dashboard which includes a Horizontal Bar Chart and a tree map to cater to different user's preferences for understanding visualized data. Here, the top skill of Troubleshooting is selected to demonstrate the information available to users about each skill.

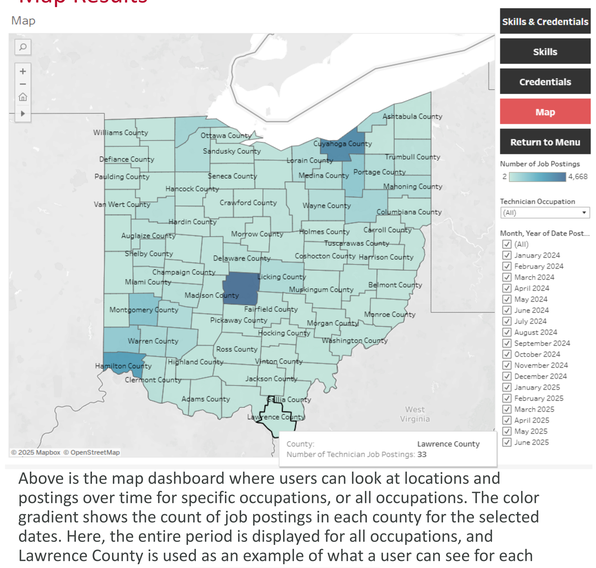
Credentials Results

Customizable visualization of the top credentials appearing in job postings for technician occupations.



Above is the credential dashboard which includes a horizontal bar chart and a tree map. Here, the fourth most common credential of HAZMAT is selected to demonstrate the information available to users about each credential.

Map Results



Above is the map dashboard where users can look at locations and postings over time for specific occupations, or all occupations. The color gradient shows the count of job postings in each county for the selected dates. Here, the entire period is displayed for all occupations, and Lawrence County is used as an example of what a user can see for each county.

The above QR code links to the interactive dashboard as published on the Ohio Labor Market Information (LMI) website, with both desktop and mobile friendly versions available.

Natural Language Processing of Raw Job Ad Data

Context

JFS researchers have done Natural Language Processing (NLP) on Raw Job Ad postings using the tidytext library in R, a library that standardizes text cleaning specifically for NLP. Due to the nature of Raw Job Ad postings, the top results were usually standard Human Resources (HR) terms that almost every posting had some variations of, rather than the skills that JFS was looking to identify. A few of the most common examples included: "*equal opportunity employer,*" "*why work with us,*", "*benefits include…,*" etc…

To solve this issue, I created a custom function that manually recreates most of what tidytext does, but in a different order that allows more customization and control to better fit the needs of this specific case.

From there, two different methodologies are tested on finding and removing HR text: calculating cosine similarities and training a supervised machine learning model. Early versions of this script used a random forest machine learning algorithm, but a switch to a logistic regression model (using the caret and glmnet libraries) resulted in enormous efficiency gains.

R Script Steps:

1. Custom builds a function for cleaning text and splitting it into sentences
2. Defines samples of HR text and creates a Term Frequency Inverse Document Frequency (TF-IDF) matrix
3. Uses sample HR text to calculate cosine similarity scores for each sentence using the TF-IDF matrix
4. Creates labels for use in a model
5. Trains a logistic regression supervised machine learning model
6. Compares results between the model and the cosine similarity calculations
7. Exports filtered and/or unfiltered results to a .csv

Roadblocks

The goal had been to create a universal script that could be plugged into any occupation. The script's manual calculations did a decent job at filtering out the HR text but still resulted in more false positives than would be desirable for a quality analysis. The machine learning model was far more likely to misidentify a sentence as a false positive for HR text. In both cases, this is due to the request for a universal model. Theoretically this issue could be solved by defining clear lists of sample skill & credential sentences and phrases for each SOC code, but due to how greatly this would vary from occupation to occupation, it is not a feasible solution for all 867 SOC codes.

Conclusions

1. Pre aggregated data on skills and credentials from private companies such as TalentNeuron **can be beneficial to job-seekers** by providing a good place to start for those exploring requirements in a career they may be unfamiliar with. However, this product has many limitations in terms of its cost and the data provided to clients such as JFS, who have no control and limited information on TalentNeuron's process of creating these skill and credential categories.
2. Doing NLP on raw job ads in-house **can be effective on a case-by-case basis** for specific in-demand occupations, but building a generalized model for all SOC codes is a far more complex task. This likely also explains the limited control and knowledge JFS has of TalentNeuron's process as mentioned in conclusion 1.
3. Future work on improving the NLP process would require a clearly defined skills list for a model to better identify what is or isn't a skill or credential. One methodology of doing this could be using O*NET's defined skill lists to start very broadly and then continue to customize based on specific needs.

Bibliography

TalentNeuron. (n.d.). *Exploring TalentNeuron Data*. TalentNeuron Support.

U.S. Bureau of Labor Statistics. (2018). *Standard Occupational Classification System*.

Acknowledgements

OERC - Dr. Josh Hawley, Ceanna Burnheimer, Mark Oleson
JFS - Chris Dixon, Richard Banks, Lacey Hall, Nick Wallace
My fellow interns – Floria Liu, Killian Hoyt, Yaqi Zhang

THE OHIO STATE UNIVERSITY

https://ohiolmi.com/technicians_dashboard