**IDS 706, Fall 2024**

# The US Labor Market through Linkedin Job Posting: A Natural Language Processing Approach Using Word2vec

**by Gedrago, Meron; Liu, Yirang and Rodriguez, Diego, Gazder, Adil**

## Abstract

The U.S. labor market is one of the most dynamic in the world. As data science students, we are curious about the future job market and eager to apply something learned in class. In the following study, we used natural language processing (NLP) to analyze job postings from LinkedIn, focusing on identifying the key skills employers seek. By processing over 60,000 job descriptions, we applied a Latent Dirichlet Allocation (LDA) model, and Word2Vec after data cleaning and tokenization. Through the models, we aimed to detect patterns in skill demand and understand relationships between different job qualifications and their regional demand. The results from the Word2Vec model provide valuable insights into the skills driving the U.S. labor market today, offering insights to jobseekers the areas where their skills are demanded across the country. Furthermore, when evaluating our model's performance on expected correlated skills, measured by cosine distance, our customized model beat the commonly used Google News Word2Vec model.

## Introduction

The U.S. labor market is changing quickly because of new technologies and industry needs. Many companies now post job openings online, and LinkedIn has become a popular place for this. These job postings are full of information about what skills employers are looking for. Analyzing this data can help job seekers learn what skills they need, help employers understand what to look for in candidates, and help policymakers create better plans for workforce training.

In this study, we used natural language processing (NLP) to study job postings from LinkedIn. We wanted to find out which skills are most in demand, how similar skills can be grouped together, and how they spread across the country. To do this, we cleaned and processed the job descriptions and used three methods: tokenization, topic modeling, and Word2Vec.

We tested two models: Latent Dirichlet Allocation, and Word2Vec—to see which one worked best for finding useful patterns in the job postings. Based on our need for explainability, Word2Vec was our selected choice since the model performance was better than other pre-trained models for this task. This paper explains how we did the analysis, what we found when analyzing skills, how the skills demand behaves across the country, and how the results can be improved in the future.

## Data

Our study utilized three datasets in total. This included two datasets from Kaggle, each providing unique and complementary information for analyzing the U.S. labor market through job postings and skill requirements. Additionally, we used the geographic data from the Census Bureau to enhance the visualization of our findings, so we plot the skills demanded across the country. Below is a brief description of each dataset and its relevance to the study:

Job Description (Koneru, 2024)
This dataset contains a large collection of job postings from LinkedIn, including job titles, descriptions, and other metadata. It served as the primary dataset for analyzing the skills demanded by employers across various industries. By processing the job descriptions, we extracted essential patterns and relationships to understand skill requirements in the labor market.

Job Skills (Johnson, 2017)
Johnson used raw data from Google Careers and extracted keywords. The output provided a collection of information commonly used in data science roles. It was valuable for identifying domain-specific skills and technical requirements and helped us with narrowing down the skill keywords. This data complemented the broader analysis of LinkedIn job postings by offering a deeper understanding of specialized skills.

Census Bureau's (TIGER, 2024)
The Census Bureau provides detailed shapefiles and geospatial information through programs like the Topologically Integrated Geographic Encoding and Referencing (TIGER) system, enabling precise mapping and analysis of regions such as states, counties, and census tracts. This data helps us to create map visualization of skills demands across the USA.

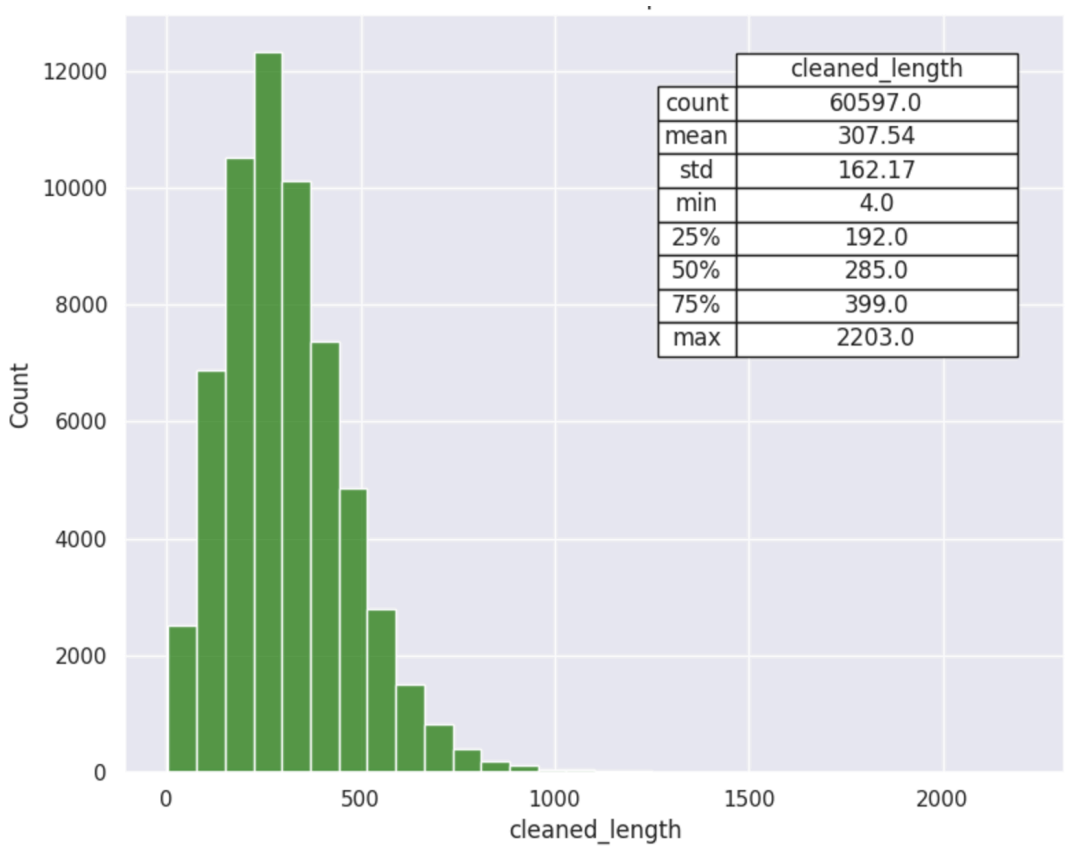
# Methodology

### Job Posting Data Processing

The first step is to tokenize the words in the job description. This is required to deal with all kinds of writing as job postings could contain irrelevant information from terms about the company, skills required, ways to connect to the company and other details of the hiring process[1]. Additionally from the regular process of tokenization, the removal of email addresses, emoticons, hyperlinks, urls, and the removal of common stop words[2] had to be implemented to get tokens ready for analysis. Figure 1 shows the distribution of tokenized job posting descriptions in the dataset. The average length of the description is 307 tokens and over 60,000 observations.

Figure 1: Amount of tokens in job posting descriptions

[1] For a sample look into the text we tokenized for this analysis check Appendix 1.
[2] A comprehensive list can be found in Appendix 2.




| | cleaned_length |
|---|---|
| count | 60597.0 |
| mean | 307.54 |
| std | 162.17 |
| min | 4.0 |
| 25% | 192.0 |
| 50% | 285.0 |
| 75% | 399.0 |
| max | 2203.0 |

After the description cleaning, we checked the most common applicable words. For this purpose, by looking into the word cloud to check the most common words, we find a mix of skills, position-related information, such as location, regulation or requirement, and industry fields. Figure 2, shows the 10 most common relevant words, while Figure 3 shows the word-based word cloud library to have a representation of the most common words within the job post description dataset.

Figure 2: 10 most common words in the LinkedIn job post descriptions

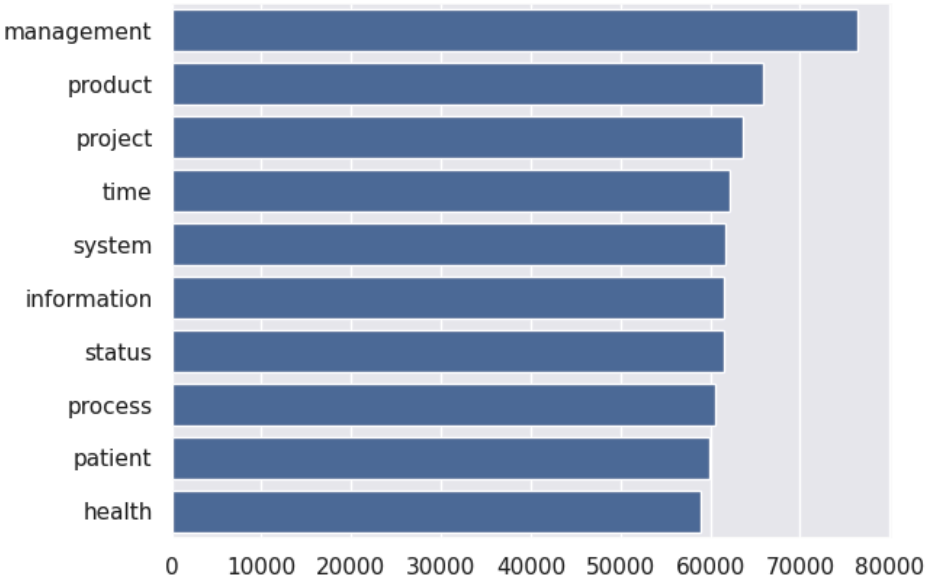
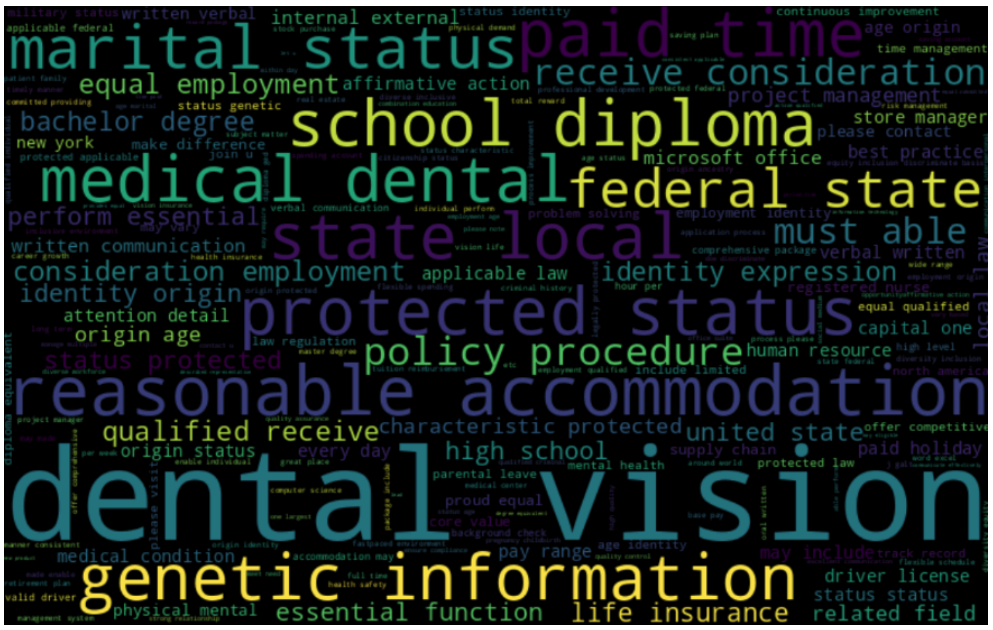



Figure 3: Wordcloud over LinkedIn job post descriptions




The first stage to find the skills within each job description is to use topic modeling. The institution behind this approach is to leverage Gibbs sampling to pick unique words and then identify a particular topic. For this purpose we used LDA to find the underlying topics, within our dataset, we iterated with many options of topic number during the process. Table 1 shows a sample of a LDA with 50 topics. However, as Table 1 shows, most of the topics found within this iteration were a mix of job-related information, job skills, and job fields.

Table 1: Sample of the top 5 words by topic (group) using LDA

| Topic 2 | |
|---|---|
| **Word** | **Prob** |
| store | 0.043 |
| customer | 0.028 |
| sale | 0.017 |
| retail | 0.013 |
| product | 0.012 |

| Topic 36 | |
|---|---|
| **Word** | **Prob** |
| business | 0.012 |
| customer | 0..010 |
| product | 0.008 |
| role | 0.008 |
| people | 0.008 |

| Topic 41 | |
|---|---|
| **Word** | **Prob** |
| health | 0.027 |
| program | 0.012 |
| care | 0.012 |
| community | 0.010 |
| family | 0.009 |

| Topic 49.000 | |
|---|---|
| **Word** | **Prob** |
| status | 0.035 |
| employment | 0.022 |
| state | 0.020 |
| protected | 0.018 |
| law | 0.016 |

| Topic 12.000 | |
|---|---|
| **Word** | **Prob** |
| member | 0.013 |
| paid | 0.010 |
| dental | 0.009 |
| care | 0.009 |
| program | 0.009 |

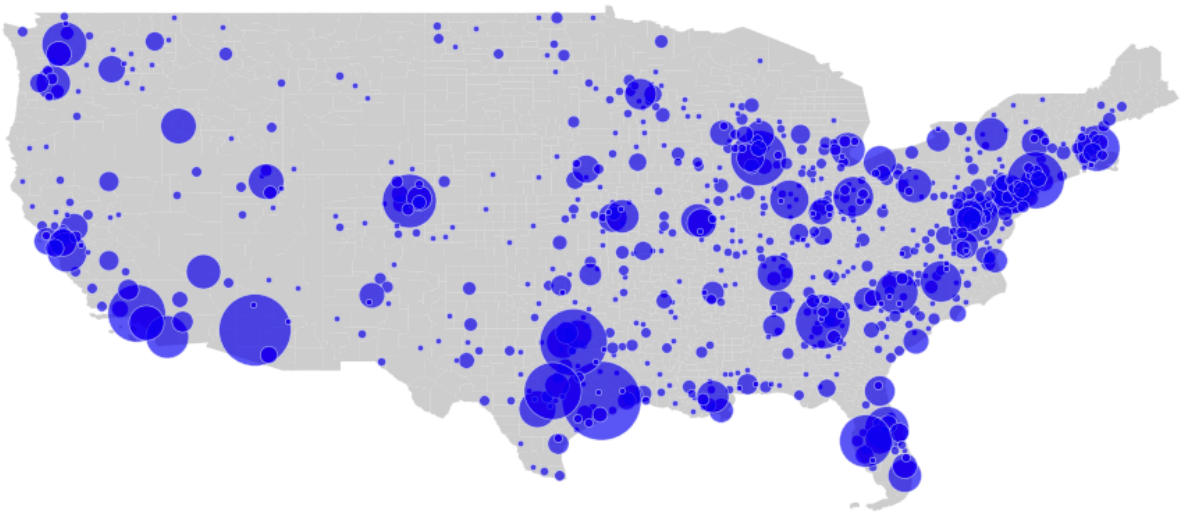
| Topic 8.000 | |
|---|---|
| **Word** | **Prob** |
| event | 0.008 |
| employment | 0.007 |
| resume | 0.007 |
| may | 0.007 |
| pay | 0.007 |

**Job Skills Dataset - Data Processing**

Then we turned to the skill dataset to use it as a dictionary to extract the skill-related tokens from the job description dataset. The tokenization process for the skills dataset was the

same as the one highlighted in the job posting section. Once tokenized, we used the tokens as a set of words to scrap the tokens associated with skills in job descriptions.
At this point, we can already answer simple, but interesting questions related to the skills in demand in the US labor market, such as, how many positions have demanded a particular type of skills. For example, the following graph shows how many positions that demand construction are open across the US.

Figure 4: Job Demand for construction skills in the USA



However, given the current information, we are not able yet to see how different skills relate to each other. To complete this analysis we used Word2Vec, which was a computational framework designed to generate dense, distributed representations (embeddings) of words in a vector space. It leverages shallow neural networks to capture semantic and syntactic relationships among words based on their co-occurrence patterns in a corpus. This was achieved by training a weight matrix that serves as the embedding layer, mapping words into a high-dimensional vector space where similar words are positioned closer together.

**Word2Vec Methodology**
Word2Vec is particularly effective for checking the proximity between words in a semantic space. By representing words as dense vectors in a continuous vector space, the proximity between words can be quantified using similarity measures, such as cosine similarity, Euclidean distance, or dot product. Words with similar meanings or contextual usage in the training corpus tend to have closer vector representations, enabling comparisons based on their semantic closeness. This makes it an ideal tool compared to other options such as one-hot encoding or TF-IDF, which treat words as independent entities.

For example, given the embeddings of words like *cat*, *dog*, and *car*, the cosine similarity between *cat* and *dog* would be higher than that between *cat* and *car*, reflecting their contextual and semantic proximity. This property is critical for tasks like clustering similar terms, retrieving related words, and building recommendation systems. Word2Vec's ability to encode nuanced relationships in this manner provides a robust tool for analyzing semantic proximity in natural language, making it ideal to understand the skills required in the US labor market. Those interactions are explored in the following section.

## Results

As outlined in the introduction and methodology, we tested two models to extract insights from job descriptions and identify the skills required for various job types. Both models varied in output and interpretability, however, we found that the **Word2Vec** provided better interpretable outputs, aligning well with our project objectives of finding the skills required within the labor market

- **LDA:** We trained the LDA model with two different numbers of topics: 5 and 20 are presented. In the 5-topic model, the limited number of topics combined with the diversity of skills resulted in top-related skills that did not make much sense. However, increasing the number of topics to 20 provided greater nuance, allowing us to infer specific themes. For instance, topics 2 and 19 appear to be related to computer/software skills, topic 17 to health-related skills, and topic 6 to technician/electrical skills, as shown in Table 3 below.

  Table 2: Top 10 skills for LDA model 1 (with 5 topics)

| Topic number | Top 10 related skills for each topic |
|---|---|
| 1 | project, degree, management, engineering, technical, support, design, system, related, process |
| 2 | equipment, safety, customer, perform, ability, product, lift, time, skill, must |
| 3 | care, health, service, medical, time, program, life, state, day, provide |
| 4 | degree, office, skill, communication, management, related, information, process, support, ability |
| 5 | business, product, skill, new, service, development, build, across, industry, need |

  Table 3: Top 10 skills for LDA model 2 (with 20 topics)

| Topic number | Top 10 related skills for each topic |
|---|---|
| 1 | identity, employment, status, age, information, qualified, genetic, please, expression, action |
| 2 | system, security, technical, support, network, technology, issue, software, computer, configuration |
| 3 | store, customer, retail, sale, service, associate, ability, time, policy, operation |
| 4 | engineering, science, development, technical, system, degree, design, technology, engineer, analysis |

| | |
|---|---|
| 5 | people, career, make, join, looking, every, world, ha, value, grow |
| 6 | equipment, repair, electrical, maintenance, mechanical, technician, system, tool, safety, part |
| 7 | sale, relationship, business, marketing, customer, new, product, market, account, strategy |
| 8 | project, management, degree, ensure, process, manager, communication, related, engineering, quality |
| 9 | communication, monthly, skill, system, process, payment, service, written, account, record |
| 10 | employment, identity, status, law, age, consideration, receive, qualified, ability, program |
| 11 | safety, school, time, equipment, customer, diploma, service, order, area, product |
| 12 | perform, lift, essential, ability, function, stand, may, must, use, hand |
| 13 | employment, identity, age, status, please, law, genetic, information, qualified, state |
| 14 | range, factor, based, pay, life, total, world, identity, may, diverse |
| 15 | client, financial, degree, office, skill, firm, accounting, communication, business, strong |
| 16 | staff, appropriate, education, policy, procedure, information, andor, may, knowledge, related |
| 17 | care, patient, health, healthcare, medical, life, service, clinical, nursing, nurse |
| 18 | america, europe, world, application, technology, asia, take, fortune, service, new |
| 19 | engineer, development, technology, code, engineering, service, software, java, cloud, join |
| 20 | data, project, business, across, stakeholder, technology, tool, skill, design, management |

- **Word2Vec:** After training the Word2Vec model, we tested specific skills to identify which other skills were related according to the model. As we can see in the figures below, the majority of the related skills made intuitive sense, as we would traditionally associate those skills. Compared to the LDA, this model provided outputs that were more interpretable and aligned with our objectives, as demonstrated in the figures below.

Figure 5: Top skills similar to “model”

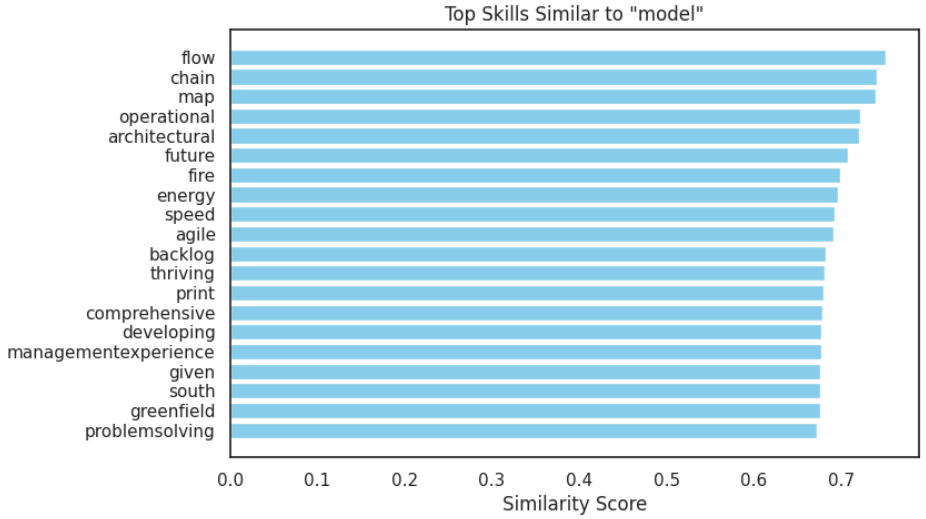



Figure 6: Top skills similar to “aws”

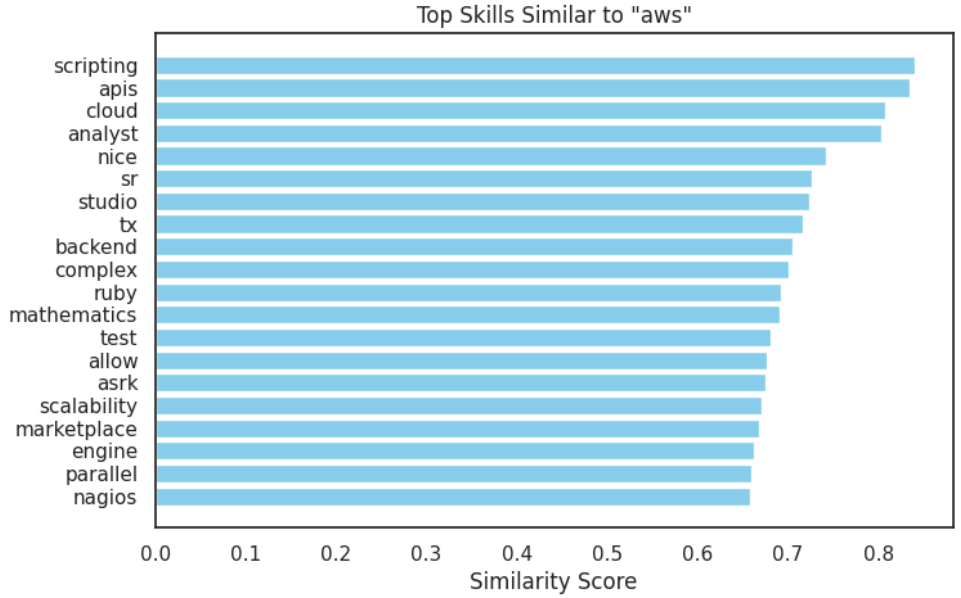


Figure 7: Top skills similar to “ml”

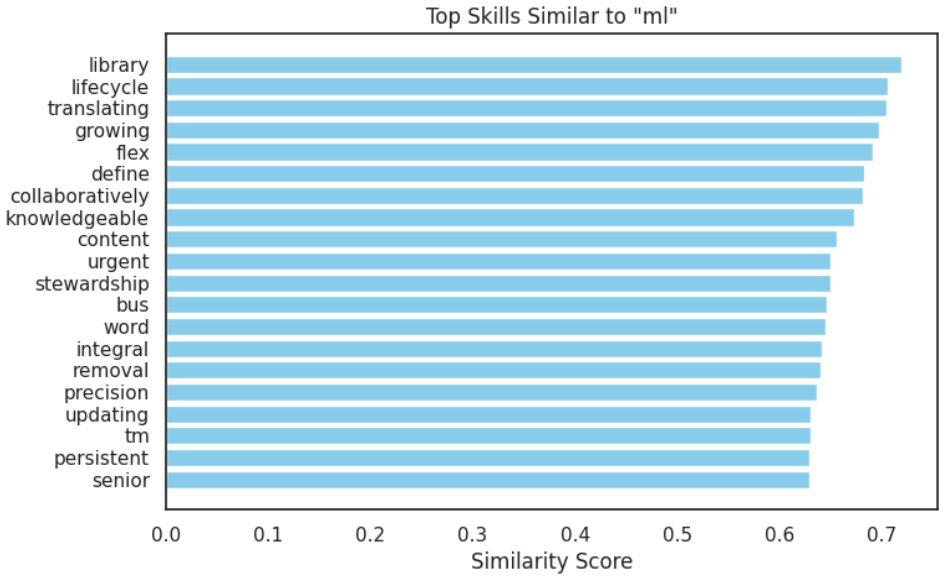
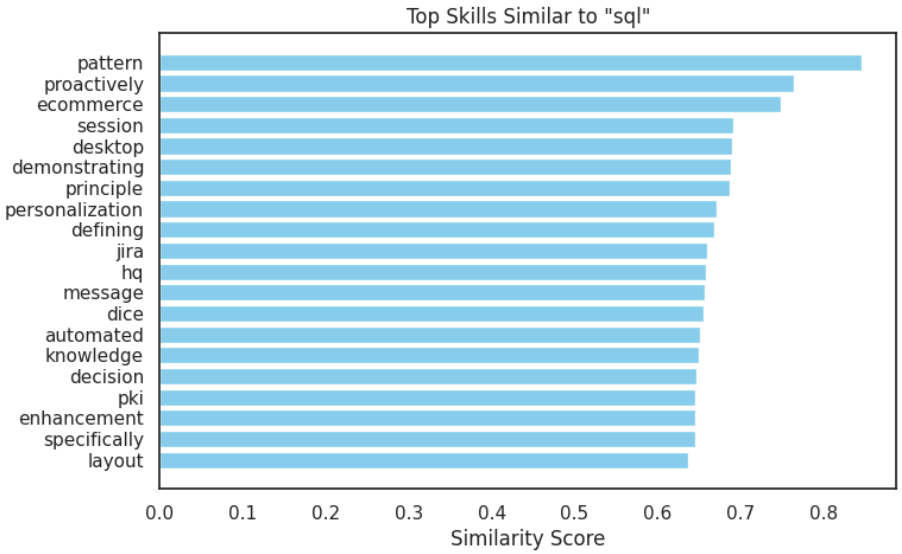



Figure 8: Top skills similar to “sql”




As shown in the figures above, many of the related skills identified by the model are logical and align with expectations, suggesting that this approach shows as the more likely skills related to one another within a specific job post description.

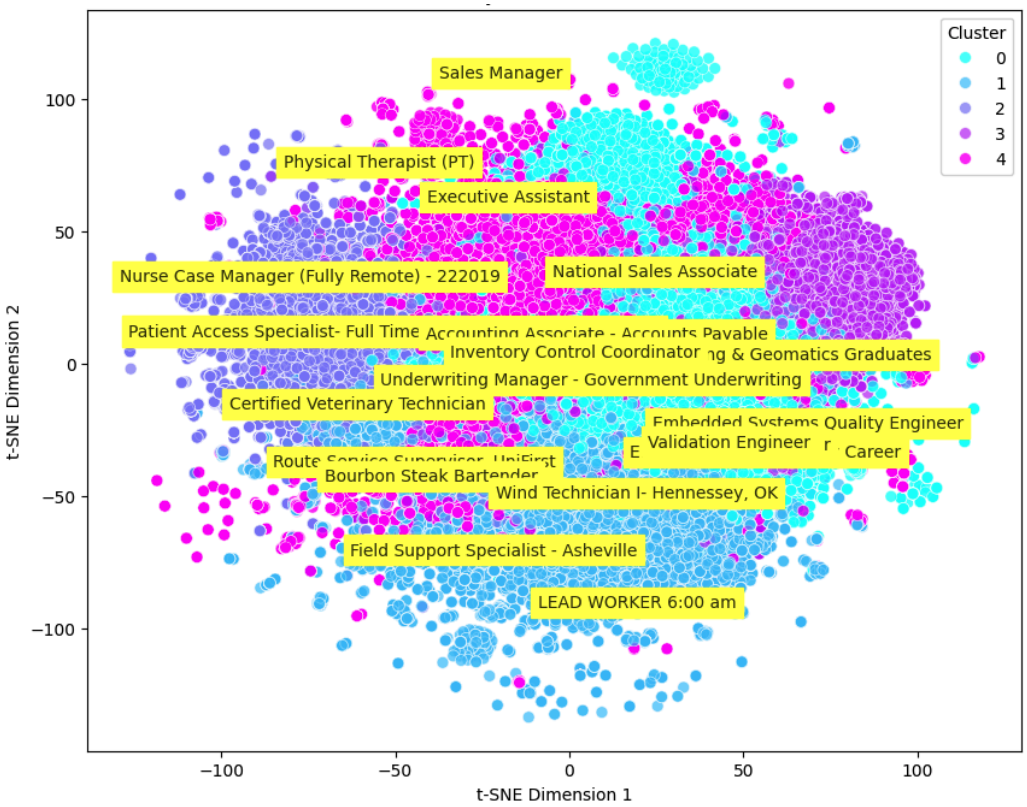
When comparing our custom-trained model to a pre-trained model (such as the Google News Word2Vec model), we find that our model performs better at identifying unique and relevant skills related to the input skills. Table 4 summarizes these findings, in most cases, our model suggests skills that are closely connected or make sense in a professional context. In contrast, the Google model often generates suggestions that are either not unique or unrelated to a professional setting, as shown in the 'java' and 'aws' comparison below. This is expected, as we are specifically modeling a list of skills, while the pre-trained model is trained using Google News and focusses less on job-related/labor market insights.

Table 4: Top similar skills and their cosine similarity for our model and Google News model

| Top skills similar to 'java' | | Top skills similar to 'design' | | Top skills similar to 'aws' | |
|---|---|---|---|---|---|
| Our model (NLP Din) | Google News model | Our model (NLP Din) | Google News model | Our model (NLP Din) | Google News model |
| postgresql: 0.8 | coffee: 0.65 | flow: 0.71 | designs: 0.77 | scripting: 0.84 | limi: 0.55 |
| cloudbased: 0.75 | o_joe: 0.63 | server: 0.71 | designing: 0.72 | apis: 0.83 | efficace: 0.53 |
| domain: 0.74 | chai_latte: 0.58 | scheduling: 0.69 | Design: 0.7 | cloud: 0.81 | orion: 0.52 |
| mongodb: 0.73 | joe: 0.58 | aptitude: 0.69 | architectural: 0.64 | analyst: 0.8 | je_ne: 0.52 |
| bi: 0.72 | espresso: 0.57 | map: 0.68 | designers: 0.63 | nice: 0.74 | stato: 0.51 |
| sdks: 0.71 | Stumptown_coffee: 0.56 | developing: 0.67 | Engenio_logo: 0.63 | sr: 0.73 | pratique: 0.51 |
| tableau: 0.67 | latte: 0.55 | motivated: 0.67 | architecture: 0.63 | studio: 0.72 | mmj: 0.51 |
| toad: 0.67 | frappuccino: 0.55 | architectural: 0.66 | Manufacturability_DFM: 0.59 | tx: 0.72 | Qu'est_ce: 0.5 |
| statistic: 0.67 | mocha: 0.55 | operational: 0.66 | redesign: 0.59 | backend: 0.71 | dpa_jbp: 0.5 |
| concurrent: 0.66 | lattés: 0.54 | speed: 0.66 | stylized_Activant_logo: 0.58 | complex: 0.7 | certi: 0.5 |

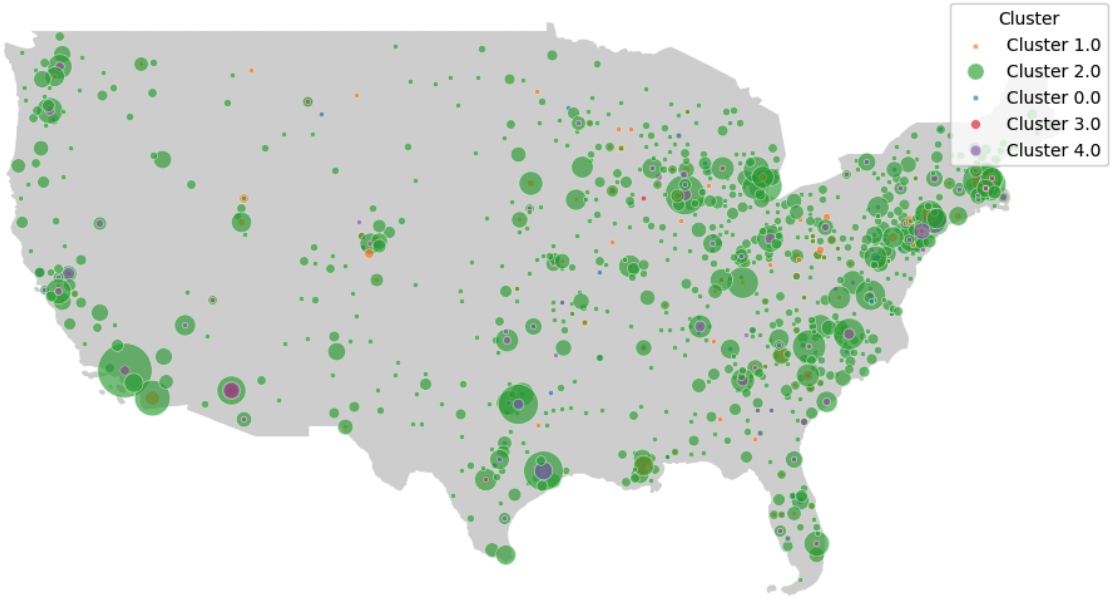
We now have access to a reliable word embedding representation of skills, enabling deeper analysis. One approach is to apply clustering to explore whether specific skills are grouped by different areas of interest. Figure 9 visualizes the results using t-SNE for dimensionality reduction and k-means to cluster job postings into five groups based on skills. Labels for twenty randomly selected job titles are overlaid on the graph to illustrate the relationship between job titles and their embedded skill representations.

Figure 9: Skill embeddings visualized with t-SNE dimensionality reduction and k-means clustering




As shown in the figure above, certain job positions are distinctly segmented within the sphere. Engineering roles tend to cluster toward the lower-right, while health-related positions are more concentrated on the right side of the cloud. Leveraging this segmentation, we can specifically analyze health-focused positions within the second cluster. Figure 10 highlights the distribution of job postings requiring nursing skills by cluster, demonstrating that the second cluster has a stronger association with healthcare roles compared to the others.

Figure 10: Demand for Nursing Skills Across the USA by K-Means Cluster



Finally, Word2Vec embeddings enabled us to assess regional demand for specific skills. By extracting vector representations for these skills and combining them into a composite vector, we establish a benchmark for comparison. Figure 11 illustrates regional skill demand by averaging the embeddings of these skills from job postings within each county. Cosine similarity is then applied to quantify how closely each region's demand aligns with the selected skill set, with normalized similarity scores improving interpretability. The difference among Figures 11.A and 11.B highlight that some skills are highly concentrated in some areas, while others are more demanded across the country. This approach highlights the effectiveness of Word2Vec in capturing semantic relationships between skills and providing actionable insights into regional demand patterns among different types of skills.

Figure 11.A: Demand for python, sql, analysis, machine,learning based on average similarity across the USA

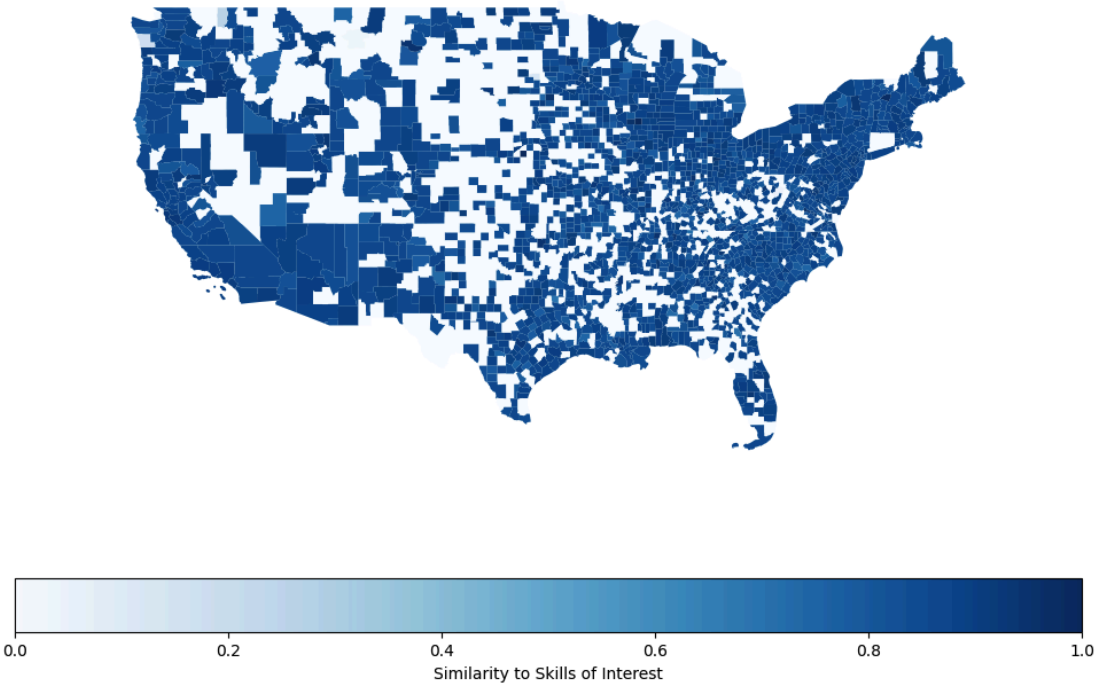
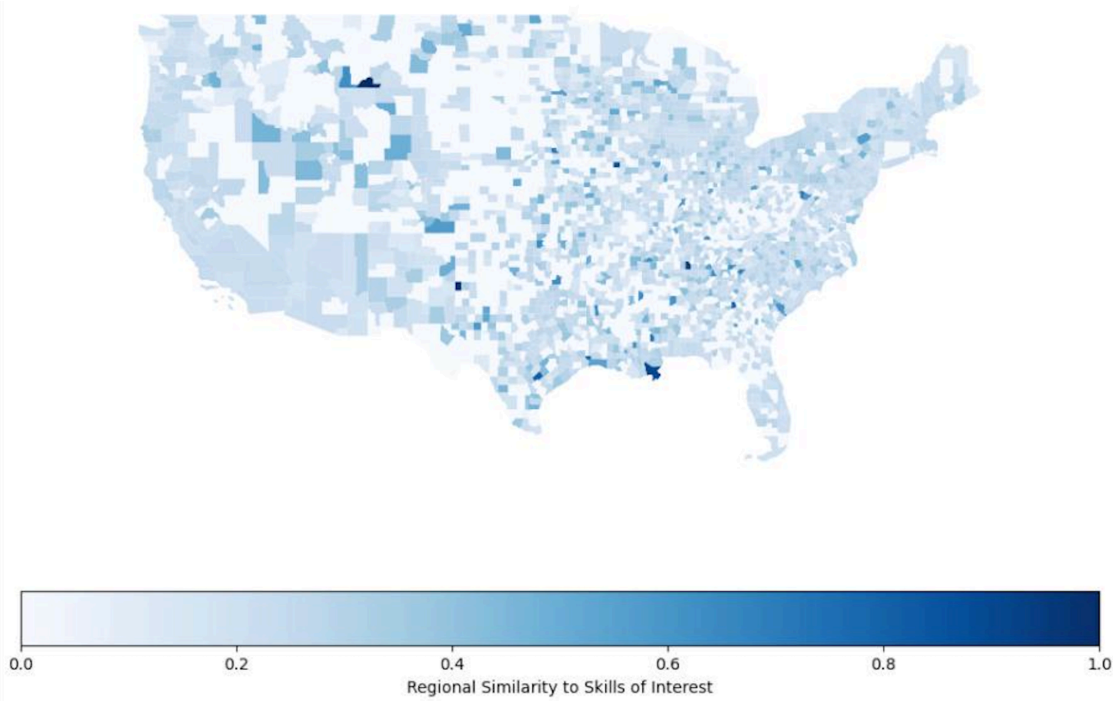



Figure 11.B: Demand for Metal, Engine, Manufacture, based on average similarity across the USA




## Conclusions

The study underscores the dynamic and variety of skills required in the U.S. labor market. By leveraging natural language processing (NLP) techniques, we analyzed over 60,000 LinkedIn job postings to identify the skills demanded to understand their relationships to geographic patterns. Among the methods tested, Word2Vec demonstrated its strength in capturing semantic relationships between skills, allowing us to uncover actionable insights about skill clustering and regional demand trends. Our findings have practical implications to focus efforts on skills development within the labor market. Specifically, job seekers will use these results to help align their skill sets with market demands, employers can refine their hiring strategies, and policymakers are enabled to design targeted workforce development programs.

## Future Work

When conducting the above analysis, the study also highlighted challenges, including data quality issues and the need for more robust preprocessing techniques. Addressing these limitations will be essential for future analyses to improve accuracy and applicability. Some example areas to pursue in future revisions include:

> Skill Data quality**:** Our primary limitation was the quality of the skills dataset. We consider this skill dataset at face value, meaning if new skills emerge or some of them are outdated, we currently don't have a way to manage this situation. A possible solution is to monitor future job posting scrap specifically job requirements or job duties sections to see the specifics of those sectors. Again, this wouldn't be trivial work as each company has unique ways to post their job description process

and will require trial and error. If we trained our model with a more concise and cleaner skill dataset, it would improve the skills identified in the job description and lead to more insightful results from both of the models we discuss above.

Tokenization/pre-processing: To train the models, we employed standardized tokenization techniques, such as lowercasing text, removing certain words from the job descriptions, and tokenizing by word. While necessary, these preprocessing steps introduced limitations as each word is tokenized independently. For example, important elements like two-words, abbreviations, and proper nouns (e.g. “Microsoft Excel”) were not used for training, potentially impacting the quality of the model. Future approaches will allow us to account for grouped words.

Expand Job Description Dataset: While the current dataset contains over 60,000 job descriptions from the period 2023-2024, we know that yearly job openings in the US are averaging nine million a month on the Bureau of Labor Statistics. Increasing the dataset size would help have more accurate information about the state of the skill demanded across the USA. Also, having access to an extended period of time would help to understand shifts on demand of some skills across time, like seeing the explosion of AI demand after large language models become more common in late 2022.

# Appendix

1. **Sample Job Posting - Techolution - Senior Robotic Engineer**

Title: Senior Commercial Grade Robotics Engineer (Inventory Management)
Experience: 2+ years (in the relevant role) Location: Ridgewood NJ - 07450
Techolution is looking for a smart and dynamic Senior Commercial Grade Robotics Engineer to lead the design, development, and implementation of cutting-edge autonomous robotic systems. (...)
Roles and Responsibilities: Oversee the end-to-end design, development, and deployment of commercial-grade robotic systems, from concept to production, ensuring adherence to quality standards and project requirements.Design and develop advanced linear motion

track systems specifically tailored for cobot applications, ensuring seamless integration and optimized performance in various industrial settings(...)Develop advanced control algorithms for robotic systems, leveraging sensors, actuators, and feedback mechanisms to achieve precise control and intelligent behavior. Ensure hardware compliance by staying abreast of industry standards and regulations, incorporating them into the design and development processes. Utilize end-to-end prototyping methods, including 3D printing, CNC machining, laser cutting, and optionally PCB fabrication and assembly .Demonstrate proficiency in mechanical design principles, encompassing joint mechanisms, Design for Assembly (DFA), and Design for Manufacturing (DFM).

Preferable Skills: Bachelor's/Master's degree in Mechanical Engineering, Electrical Engineering, Mechatronics Engineering, or related field.Proven experience in designing, developing, and deploying commercial-grade robotic systems, with a minimum of 2 years of relevant experience in a leadership capacity.Exceptional attention to detail, proactive thinking, and strong problem-solving skills, with a demonstrated ability to lead and inspire a team to achieve project objectives. Collaborate closely with cross-functional teams, including R&D, product management, and quality assurance, to align the development process with organizational goals and market needs.Stay abreast of technological advancements and industry trends to continually enhance system capabilities and introduce innovative solutions.

Foster a culture of safety and continuous improvement, implementing best practices in engineering and manufacturing to achieve excellence in all project deliverables.
Experience with commercial-grade robotic components, sensors, actuators, and motion systems, as well as familiarity with industry standards and regulations.Exposure in pre-sales activities, showcasing technical expertise and presenting solutions to potential clients.Excellent communication skills, with the ability to articulate technical concepts clearly and concisely to both technical and non-technical audiences.

About Techolution: Techolution is a Product Development firm on track to become one of the most admired brands in the world for "innovation done right". Our purpose is to harness our expertise in novel technologies to deliver more profits for our enterprise clients while helping them deliver a better human experience for the communities they serve. (...)

Let's give you more insights!
One of our amazing products with Artificial Intelligence:1. https://faceopen.com/ : Our proprietary and powerful AI Powered user identification system which is built on artificial intelligence technologies such as image recognition, deep neural networks, and robotic process automation. (No more touching keys, badges or fingerprint scanners ever again!)

Some videos you wanna watch!Computer Vision demo at The AI Summit New York 2023 Life at TecholutionGoogleNext 2023Ai4 - Artificial Intelligence Conferences 2023WaWa - Solving Food Wastage Saving lives - Brooklyn HospitalInnovation Done Right on Google CloudTecholution featured on Worldwide Business with KathyIrelandTecholution presented by ION World's Greatest

Visit us @www.techolution.com : To know more about our revolutionary core practices and getting to know in detail about how we enrich the human experience with technology.

2. **Stopwords list in description**

"job", "role", "position", "responsibility", "responsibilities", "duties", "duty", "requirement", "requirements", "qualification" "qualifications", "description", "descriptions", "candidate", "candidates", "applicant", "applicants", "opportunity", "opportunities", "team", "teams", "work", "working", "employee", "employees", "employer", "employers", "company", "companies", "location", "locations", "department", "department", "report", "reports", "reporting", "benefit", "benefits", "compensation", "salary", "experience", "experienced", "year", "years", "gender", "race", "color", "sex", "orientation", "sexual", "religion", "national", "identify", "veteran", "nation", "including", "required", "disability", "regard", "without".

## Reference

Arsh Koneru. (2024). LinkedIn Job Postings (2023 - 2024) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/9200871

Johnson, R. (2017). *Job skills* [Kaggle notebook]. Kaggle. Retrieved October 26, 2024, from https://www.kaggle.com/code/rayjohnsoncomedy/job-skills/input?select=job_skills.csv