Università degli Studi di Milano

Market Basket Analysis Project

Apriori Algorithm

Algorithms for massive data

Professor Dario Malchiodi

Danial Forouzanfar, 12988A

Danial.forouzanfar@studenti.unimi.it

2023-2024

Contents

Abstract

This report performs a Market Basket Analysis on a dataset of 1.3 million job listings scraped from LinkedIn in the year 2024. The focus of this report will be to highlight those skills that are in demand within the job market. The Apriori algorithm was used to extract these job skills to find common patterns. In this project, we didn't use the whole dataset and a sample of the dataset was mined. The mining came up with 81 single frequent skills, 67 skill pairs, and 25 triples of skills. These findings highlight key skillsets in demand and gave insight into the common requirements in the job market.

1. Introduction

1.1. Background

Data science is a multidisciplinary area that includes Mathematics, Statistics, Computer science, and domain expertise to analyze and interpret complex data to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning. (1)

Machine learning[1] is an area of research in Artificial Intelligence[2] that focuses on developing systems that learn or improve performance based on data without being programmed explicitly. It is important to point that ML and AI don't have a same meaning. AI generally refers to systems or machines that mimic human intelligence. An important distinction is that although all machine learning is AI, not all AI is machine learning. (2) ML is typically divided into supervised learning, unsupervised learning, and reinforcement learning, each with different methods of learning from data.

Algorithms are the engines that power machine learning. The Apriori algorithm is a classic data mining technique used for mining association rules. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. This algorithm identifies frequently occurring itemsets in large datasets and derives association rules that describe how the occurrence of one item is related to the occurrence of another. (3)

Apache Spark is an open-source, distributed processing system that is used for big data workloads. It uses in-memory caching and optimized query execution for fast analytic queries against data of any size. It enables large-scale data to be processed across nodes in a cluster, making it ideal for big data analytics. (4)

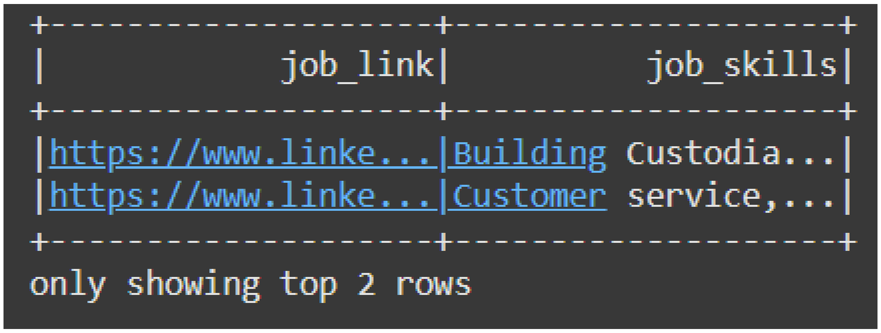
1.2. Problem statement

In today's job market, knowing which skills employers want is essential for both job seekers and companies. But sorting through job listings manually to find common skills is challenging. This project uses data science techniques, like the Apriori algorithm and Market

[1] Machine Learning - ML
[2] Artificial Intelligence - AI

Basket Analysis, to efficiently find the most frequently required skills in a large dataset of job postings. We aim to uncover skill combinations that often appear together, offering insights into the skills employers value most.

The dataset that we used in this project contain 1.3 million job listings scraped from LinkedIn in 2024. Each job listing includes a list of required skills, making it possible to analyze and use the apriori algorithm to find the patterns and common skill demand. We focused on the skills column, which lists the specific skills needed for each job.



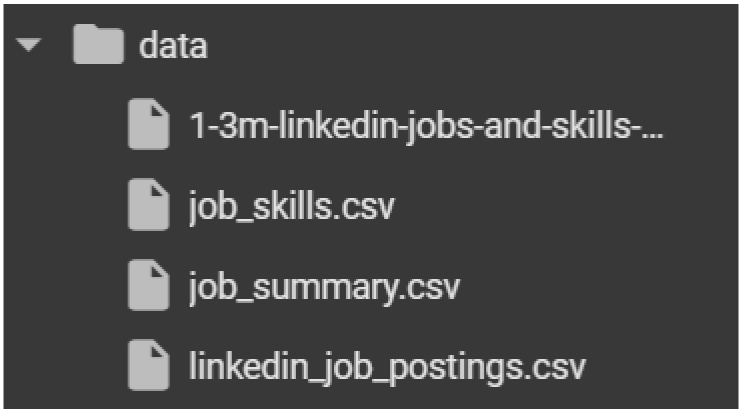
```
+--------------------+--------------------+
|            job_link|          job_skills|
+--------------------+--------------------+
|https://www.linke...|Building Custodia...|
|https://www.linke...|Customer service,...|
+--------------------+--------------------+
only showing top 2 rows
```

Figure 1. The data set

2. Methodology

2.1. Data Preparing & Preprocessing

Preparing the data was the first step to start the project with using the Apriori algorithm. We began by creating a specific path for our data in Google Colab environment. Then, by using the Kaggle API, the dataset was downloaded and stored in the prepared path. From there the dataset was loaded it into PySpark for processing.



data

1-3m-linkedin-jobs-and-skills-...

job_skills.csv

job_summary.csv

linkedin_job_postings.csv

Figure 2. The directory

Since only the skills column was relevant for our analysis, firstly we removed the column that contains the link and then we remove the rows that don’t have any value. In the next step, for each row, a list of skills was created from comma-separated strings into individual items within a list. This allowed us to structure the data into itemsets that the Apriori algorithm could process effectively.

2.2. The Apriori Algorithm

The apriori algorithm is a classic association rule mining intended to find frequent itemsets in a dataset and generate meaningful association rules. In this project, the "items" are skills listed in job postings and the goal is to uncover frequent skill sets, including their relationships, to indicate demands that are common in the job market.

The main Concepts in Apriori:

1. Support: The frequency of occurrence of the item or itemset in the dataset; it gives a base frequency for filtering out infrequent itemsets (by setting a threshold).
$$Support(I) = \frac{Frequency\ of\ I\ in\ the\ dataset}{Total\ number\ of\ transactions}$$
2. Confidence: It measures the likelihood that a particular item appears in a transaction given that another item is already present. Stronger associations between skills are defined by higher values of confidence.
$$Confidence(a \rightarrow B) = \frac{Support\ (A\ \cup B)}{Support\ (A)}$$
3. Lift: It helps identify relationships that are more meaningful than random co-occurrences.
$$Lift(a \rightarrow B) = \frac{Support\ (A\ \cup B)}{Support\ (A) *\ Support\ (B)}$$

First, the Apriori algorithm finds all the individual items that meet the support threshold. This would form the "singleton" frequent itemsets. Then, it systematically puts these together into pairs, then triples, and thus larger sets, but it retains only those that are meeting the support threshold. This tends to cut down the number of itemsets generated and narrows down to the most common and, thus, relevant combinations.

In this project, Apriori was applied on a sample of job listings. By setting a thresholds, we were able to:

- Identify the most frequent skills (singletons) that appeared across the dataset.
- Find common skill pairs (such as "teamwork" and "communication") and triples (such as "teamwork, communication, and customer service") that employers often require together.
- Derive association rules showing patterns of skill requirements, which provide insights into what combinations of skills are valued in the job market.

3. Result

By applying the algorithm on a sample of the dataset, we identified key individual skills, skill pairs, and skill triples that are frequently required by employers. These patterns help reveal the skill combinations that are most in demand across job listings.

- Frequent Singletons

The analysis found 81 frequent item, or individual skills, that met the minimum support threshold. Among these, "communication", “customer service”, "problem-solving", “teamwork”, and “leadership” emerged as the frequent mentioned skill across the dataset.

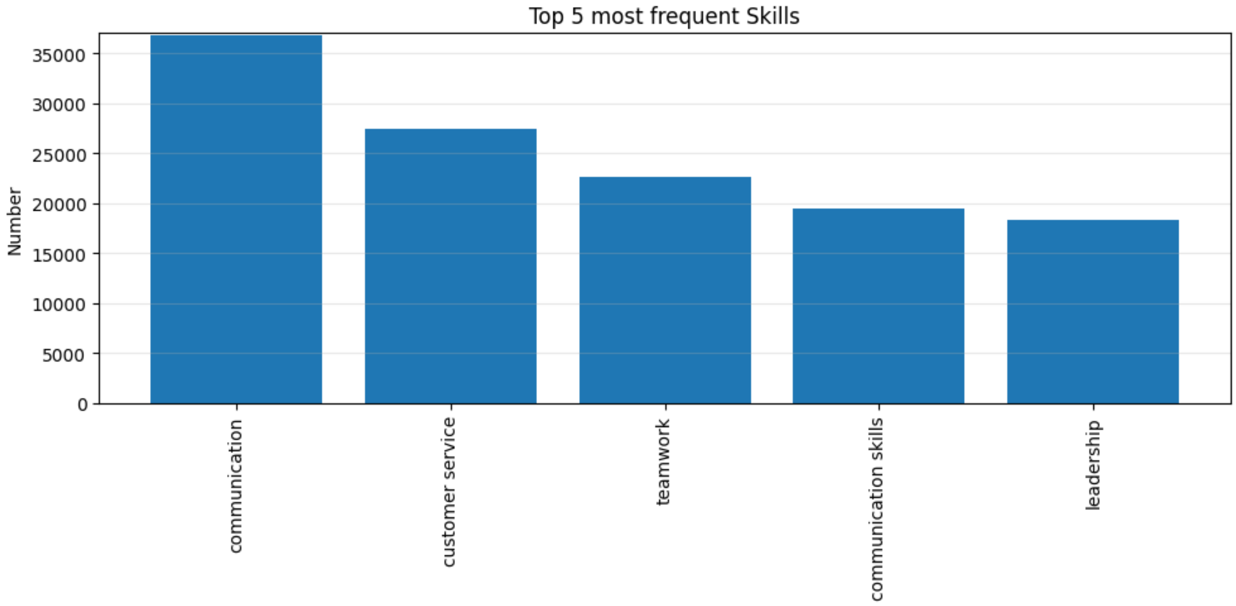


Figure 3. Most frequent singletons

- Frequent Pairs

In terms of skill pairs, the apriori algorithm identified 67 frequent pairs. The most common pair was "teamwork” and “communication" underscoring the emphasis on collaborative and interpersonal skills in many job roles.

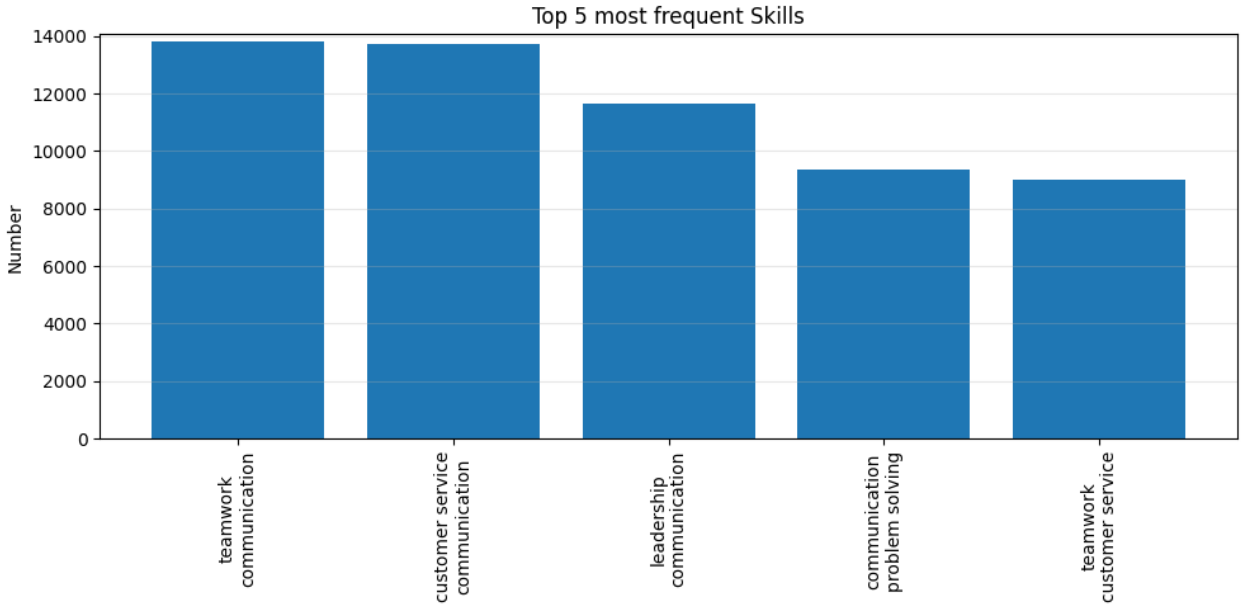


Figure 4. Most frequent pairs

- Frequent Triples

The analysis also discovered 25 frequent triples. The most common triple was "teamwork", "communication", and "customer service" which points to a pattern of employers valuing both collaboration and customer-facing abilities.

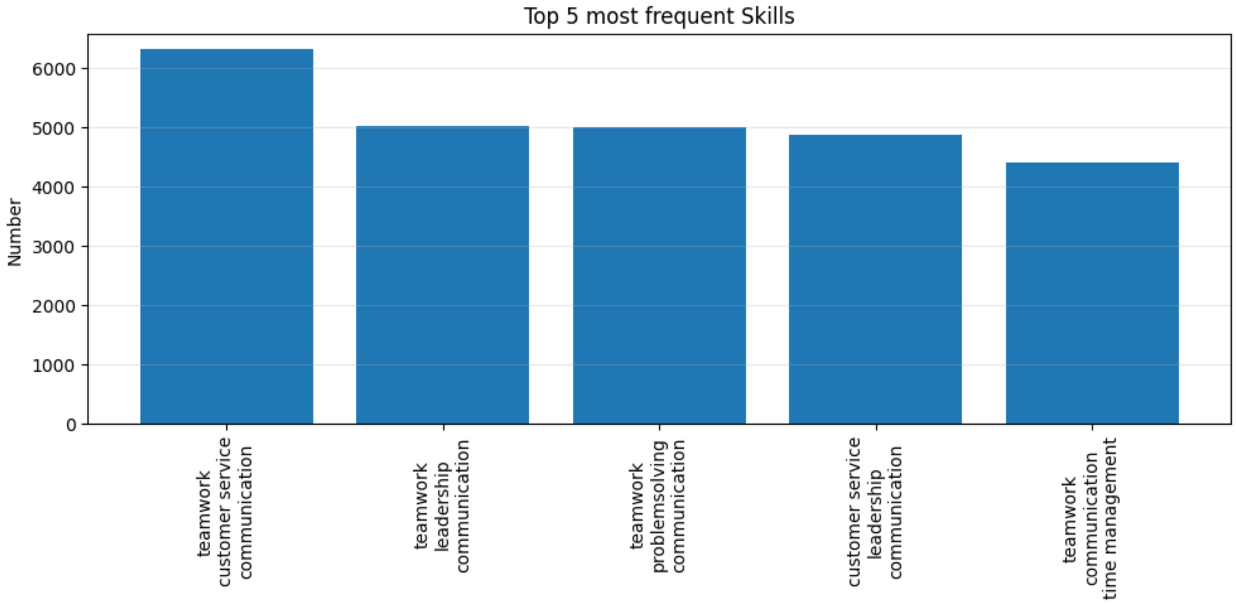


Figure 5. Most frequent triples

4. Discussion

The results show the effectiveness of the Apriori algorithm in identifying patterns of job listing and the combination of skills. The frequent itemsets and association rules provide insights into skill requirements that are not apparent when reviewing individual job postings. Organizations and individuals, based on the identified pattern, can have a better understanding of what type of competencies are highly valued in the present job market.

5. References

1 . https://www.ibm.com/topics/data-science
2. https://www.oracle.com/ca-en/artificial-intelligence/machine-learning/what-is-machine-learning/
3 . https://www.geeksforgeeks.org/apriori-algorithm/
4 . https://aws.amazon.com/what-is/apache-spark/

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work, and including any code produced using generative AI systems. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.