Algorithms for Massive Data project:

Market Basket Analysis

Kseniya Yerameichyk
kseniya.yerameichyk@studenti.unimi.it

04/07/2024

1. Introduction

This report presents the implementation and results of a Market Basket Analysis project performed on the LinkedIn Jobs & Skills dataset from Kaggle. The goal was to identify frequent itemsets within job posts, where each job skill entry is a basket, and each skill is an item. The Apriori algorithm was chosen for this task as the most basic and common algorithm for MBA. Due to the large size of the dataset, a sample was used to ensure manageable computation time within thirty minutes.

2. Data Analysis

2.1 Data description

The dataset used for the project is the LinkedIn Jobs & Skills dataset, published on the Kaggle platform. For the project purposes was used only *job_skills.csv* file. This file contains two columns: *job_link* and *job_skills*. For the purposes of this project, only the *job_skills* column (which includes strings of skills associated with various job posts) was considered. The initial dataset contained 1,296,381 rows.

2.2 Data preprocessing

The first step in data preprocessing was to clean the data by dropping the *job_link* column, as it was useless for the analysis. Subsequently, rows containing any missing values were removed. This reduced the dataset to 1,294,374 rows, ensuring that only complete records were used in the analysis.

Given the large size of the dataset, a sample was taken to make sure the computations finish within a reasonable timeframe. A sample size of 0.2% was chosen, resulting in 2,521 rows. This sample size was selected to limit the computation time to approximately 30 minutes.

The sampled DataFrame was transformed into an RDD (Resilient Distributed Dataset) for further processing. The RDD was cached to store all the RDD in-memory and use it efficiency across parallel operations. The number of partitions was set to twice the number of available cores as recommended in the Apache Spark documentation. Each row in the *job_skills* column from a string of skills separated by commas was transformed into a list of skills, creating individual baskets. An analysis of the basket sizes revealed that the largest basket contained 144 skills, the smallest basket contained one skill, and the average basket size is 21.

The support value for the Apriori algorithm was calculated based on a common book rule of using 1% of the dataset. This threshold was adjusted for the sample size, resulting in a support value of 26. This means that for an itemset to be considered frequent, it must appear in at least 26 baskets in the sampled dataset.

For more effective memory use during the Apriori algorithm, skills were mapped from strings to integers using a hash-table. This transformation allowed the algorithm to process the data more efficiently. The baskets were then converted to contain these hashed values of the skills instead of the original strings.

3. Apriori Algorithm

3.1 Implementation details

The Apriori algorithm is a popular technique used to find frequent itemsets and generate association rules. It is based on the principle that any subset of a frequent itemset must also be frequent. The algorithm works iteratively to discover larger itemsets, starting from single items and expanding to pairs, triplets, and so on, while filtering non-frequent itemsets at each step.

In the provided implementation, the function *apriori_algorithm* takes three parameters: *baskets* (an RDD containing the lists of skills), *support* (the minimum support threshold), and *hash_dict* (a hash mapping dictionary to retrieve skills string representation instead of integer one). The function begins by identifying frequent single items (singletons). It maps each item in the baskets to a count, aggregates these counts, and filters out items that do not meet the support threshold. If no frequent singletons are found, the function suggests lowering the support value and terminates.

In the case where frequent singletons are found, the algorithm iteratively increases the size of the itemsets being considered. For each itemset size, it generates combinations of items, checks that all subsets of the current itemsets are frequent, aggregates counts, and filters based on the support threshold. This process continues until no more frequent itemsets can be found.

3.2 Algorithm results

In this chapter are represented the results of applying the Apriori algorithm to identify frequent itemsets of skills from a given dataset sample. The goal of this analysis is to uncover common combinations of skills that frequently appear together, which can inform decisions in areas such as job market analysis and training programs. The analysis was conducted in two phases: first analyzing the entire sample as it is, which predominantly featured soft skills, and then iteratively removing the identified frequent soft skills to focus on hard skills.

The results of the first run of the algorithm are summarized as follows:

1. Frequent Singletons: 164
   Most Frequent Singleton Skill: *Communication*
   Support Value: 714
2. Frequent Pairs: 172
   Most Frequent Itemset: *Teamwork, Communication*
   Support Value: 252
3. Frequent Triplets: 70
   Most Frequent Itemset: *Teamwork, Communication, Leadership*
   Support Value: 96
4. Frequent Itemsets of Size 4: 9
   Most Frequent Itemset: *Teamwork, Problem Solving, Communication, Customer Service*
   Support Value: 48

The predominance of soft skills in the frequent itemsets shows their general importance across various job roles. However, this information is not particularly insightful for understanding the specific technical or hard skills that are crucial for particular professions.

Given the initial findings, the Apriori algorithm was rerun after removing the identified frequent soft skills to uncover underlying patterns among hard skills. This iterative process revealed the following results:

1st iteration

1. Frequent Singletons: 159
   Most Frequent Singleton Skill: *Customer Service*
   Support Value: 223
2. Frequent Pairs: 48
   Most Frequent Itemset: *Customer Service, Problem Solving*
   Support Value: 76
3. Frequent Triplets: 1
   Most Frequent Itemset: *Customer Service, Problem Solving, Time Management*
   Support Value: 32

2nd iteration

1. Frequent Singletons: 156
   Most Frequent Singleton Skill: *Communication skills*
   Support Value: 214
2. Frequent Pairs: 30
   Most Frequent Itemset: *Attention to detail, Time management*
   Support Value: 50

3rd iteration

1. Frequent Singletons: 153
   Most Frequent Singleton Skill: *Sales*
   Support Value: 189
2. Frequent Pairs: 22
   Most Frequent Itemset: *Patient Care, Nursing*
   Support Value: 49

4th iteration (Lowering Support Threshold)

1. Frequent Singletons: 234
   Most Frequent Singleton Skill: *Sales*
   Support Value: 189
2. Frequent Pairs: 73
   Most Frequent Itemset: *Patient Care, Nursing*
   Support Value: 49

3. Frequent Triplets: 6
   Most Frequent Itemset: *Walking, Lifting, Standing*
   Support Value: 26

The initial results contained exclusively soft skills such as Communication, Teamwork, and Leadership. These skills are important and valued across nearly all job roles and industries, leading to their high frequency in the dataset. However, they are too generic and render the results less informative for targeting specific technical competencies or hard skills.

By iteratively removing frequent soft skills, the algorithm was able to identify frequent itemsets containing hard skills. The frequent occurrence of skills like Patient Care and Nursing, or technical activities like Walking, Lifting, and Standing, highlights the specific demands of roles in healthcare and related fields.

After the third iteration, the support threshold was lowered from 1% to 0.7% to obtain more frequent itemsets. This adjustment enabled to capture additional patterns that were previously filtered out due to the higher threshold, finding less frequent but still significant skill combinations.

4. Conclusion

This Market Basket Analysis project successfully utilized the Apriori algorithm to identify frequent itemsets of job skills from the LinkedIn Jobs & Skills dataset. The initial phase of the analysis revealed a high frequency of soft skills such as Communication, Teamwork, and Leadership, indicating their universal importance across various job roles. These findings highlight the critical nature of soft skills in the modern job market.

However, to uncover more specific technical competencies, an iterative approach was adopted where frequent soft skills were removed progressively. This strategy allowed to identify the important hard skills, providing a clearer picture of the specific demands of certain professions, particularly in healthcare. For instance, skills such as Patient Care and Nursing, along with physical tasks like Walking, Lifting, and Standing, emerged as significant in the dataset sample. Lowering the support threshold further enabled the discovery of additional significant itemsets that were not found at a higher threshold, illustrating the importance of adjusting parameters to uncover more useful information.

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.