Figure 1: wordcloud of item frequencies

Market basket Analysis

(Finding Frequent Items on LinkedIn Jobs & Skills Dataset)

Algorithms for Massive data project report

Samaneh Azarakhsh

Master student of University of Milan, Data Science and Economics, Matricular Number 963748, samaneh.azarakhsh@studenti.unimi.it

Abstract

Market basket analysis, a key data mining technique, explores co-occurrence patterns to reveal relationships between concurrently purchased products, also known as association analysis or frequent itemset mining. This study utilizes the FP-Growth algorithm to examine the "1.3M LinkedIn Jobs and Skills 2024" dataset, a substantial compilation of LinkedIn job listings. By treating text strings as market items, this project aims to discern prevalent item sets that reflect the most in-demand skills and job requirements, offering a unique insight into the labor market trends and skill demand dynamics.

Key words: Market Basket Analysis, Frequent Items set, LinkedIn Jobs & Skills, NLP, FPGrowth, PySpark

1. Introduction

Frequent item sets, also referred to as association rules, play a crucial role in association rule mining—an essential technique in data mining that aims to uncover connections between items within a dataset. The primary objective is to identify and understand relationships among items that frequently co-occur in the dataset.

A frequent item set consists of items that commonly appear together in a dataset. The frequency of an item set is determined by its support count, representing the number of transactions or records in the dataset that include the specific item set.

Mining frequent items, itemsets, subsequences, or other substructures is usually among the first steps to analyze a large-scale dataset, which has been an active research topic in data mining for years.

The purpose of association rule mining algorithms like FP-Growth and Apriori is to identify frequently occurring item sets and produce association rules. Iteratively creating candidate item sets and eliminating those that don't meet the minimal support requirement is how these algorithms operate. The idea of confidence, which is the ratio of the number of transactions including the item set to the number of transactions containing the antecedent (left-hand side) of the rule, can be used to build association rules once the frequent item sets have been identified.

Many applications, including cross-selling, recommendation engines, and market basket research, can benefit from the use of frequent item sets and association rules. It should be highlighted, though, that association rule mining has the potential to produce a huge number of rules, many of which might be pointless or boring. Thus, it's critical to assess the generated rules' level of interest using suitable metrics like lift and conviction.

2. About LinkedIn Jobs & Skills (2024) Dataset

The "1.3M LinkedIn Jobs and Skills 2024" dataset is an extensive collection of job listings and associated skills aggregated from LinkedIn, one of the largest professional networking platforms

globally. This dataset encompasses over 1.3 million job entries, reflecting a wide array of industries, roles, and geographical locations. The primary objective of compiling this dataset is to provide a robust foundation for analyzing labor market trends, skill demand dynamics, and the evolving landscape of job requirements across different sectors.

The dataset is structured to offer a granular view of the job market, with each entry detailing specific aspects of job listings, including but not limited to:

- Job Titles: The designation or title of the job posting, which reflects the role's nature and level within the organizational hierarchy.
- Skills: A comprehensive list of skills required for the respective roles, providing insights into the qualifications and competencies employers seek in candidates.
- Industries: The sector or industry classification of the job, enabling analysis of demand patterns across various economic sectors.
- Locations: Geographical information indicating where the job is based, facilitating regional labor market analysis.
- Company Names: The names of the organizations posting the jobs, offering a lens into hiring trends among different employers.

3. About the algorithm

3.1. Basket analysis

Basket Analysis involves examining the contents of baskets in the context of shopping, whether conducted online or offline. The focus is on analyzing transactions to gather data that records the products associated with each purchase. This method provides insights into consumer behavior by studying the patterns of products bought together, aiding retailers and businesses in understanding customer preferences and optimizing strategies for marketing, inventory management, and personalized recommendations. In essence, Basket Analysis is a valuable tool for enhancing the shopping experience and tailoring business approaches based on observed transactional patterns.

3.2. Apache Spark

Apache Spark represents an open-source distributed processing system essential for handling extensive big data tasks. It employs in-memory caching and refined query execution techniques, enabling swift analytical queries on datasets of varying sizes. This technology is designed to enhance the efficiency and speed of processing large volumes of data, making it a valuable tool for diverse applications in the field of big data analytics. It supplies programming interfaces for development in Java, Scala, Python, and R, facilitating code reuse across a spectrum of workloads. These workloads span batch processing, interactive queries, real-time analytics, machine learning, and graph processing. This versatility allows users to leverage the capabilities of Apache

Spark across various tasks, promoting seamless integration and adaptability for different data processing requirements in the realms of big data analytics.

In 2009, Apache Spark originated as a research initiative within the AMPLab at UC Berkeley. The AMPLab, a collaborative effort involving students, researchers, and faculty, concentrated on domains with data-intensive applications. The primary objective of Spark was to develop an innovative framework specifically tailored for swift iterative processing, encompassing tasks such as machine learning and interactive data analysis. Importantly, Spark aimed to uphold the scalability and fault tolerance characteristics inherent in Hadoop MapReduce, ensuring its applicability across diverse and demanding computational scenarios.

3.3. PySpark

PySpark serves as the Python Application Programming Interface (API) for Apache Spark, facilitating real-time, extensive data processing within a distributed framework using the Python programming language. Additionally, it furnishes a PySpark shell that allows interactive examination of data.

The synergy between PySpark and Python capitalizes on the latter's user-friendly nature and ease of learning, merging it with the robust capabilities of Apache Spark. This amalgamation empowers individuals proficient in Python to engage in the processing and analysis of data at any scale. The user-friendly interface and the potency of Apache Spark make PySpark a valuable tool for a broad audience seeking efficient and scalable data processing solutions. PySpark is faster than Pandas when processing large datasets. It can leverage the computing power of a cluster of machines to perform parallel processing. This can significantly reduce processing times.

3.4. FP-Growth

The FP-growth algorithm is a method for mining frequent patterns without the need for candidate generation. In this context, "FP" denotes frequent pattern. In the initial stage of FP-growth, when applied to a dataset of transactions, the primary step involves determining the frequencies of individual items and recognizing those items that occur frequently. This process entails identifying and quantifying the items that manifest with high regularity in the dataset, laying the foundation for subsequent stages in the FP-growth algorithm.

The FP-growth (Frequent Pattern growth) algorithm involves several key steps in mining frequent patterns from a dataset without the conventional approach of candidate generation.

Here are the fundamental steps of the FP-growth algorithm:

1. Counting the occurrences of individual items
2. Filter out non-frequent items using minimum support
3. Order the itemsets based on individual occurrences
4. Create the tree and add the transactions one by one

The FP-growth algorithm efficiently discovers frequent patterns by leveraging the FP-tree structure and avoiding the need for candidate generation, making it a valuable tool in association rule mining for large datasets.

In the context of association rule mining, various metrics are employed to evaluate and quantify the significance and quality of discovered rules. These metrics play a crucial role in assessing the strength and reliability of associations between items in a dataset. Here are some common metrics used in the evaluation of association rules:

Support:

Support quantifies the fraction of transactions in the dataset containing the items specified in the rule. Increased support implies broader applicability of the rule across a more extensive portion of the dataset.

Confidence:

Confidence evaluates the probability of the consequent occurring when the antecedent is present. Higher confidence values indicate a more robust association between the antecedent and consequent.

Lift:

Lift measures the ratio of the observed support for the rule to the expected support, assuming independence between the antecedent and consequent. Lift values exceeding 1 signify a positive impact, indicating a non-random association in the rule.

4. Experimental results

Upon obtaining data from Kaggle and subsequently extracting and updating the file path for the CSV, I proceed to import the data into a PySpark DataFrame. The primary focus of this project lies within the textual content contained within the " job_skills" column of the job_skills.csv file.

The process initiates with the retrieval of data from Kaggle, followed by the extraction of files and the adjustment of the file path for the CSV. Subsequently, the data is imported into a PySpark DataFrame for further analysis and manipulation.

Upon completion of these steps, the dataset is effectively loaded into the PySpark environment, ready for exploration and examination. The ensuing analysis will predominantly center on the textual information encapsulated within the designated " job_skills" column of the job_skills.csv file.

job_link	j
https://www.linkedin.com/jobs/view/housekeeper-i-pt-at-jacksonville-state-university-3802280436	B
https://www.linkedin.com/jobs/view/assistant-general-manager-huntington-4131-at-ruby-tuesday-3575032747	C
https://www.linkedin.com/jobs/view/school-based-behavior-analyst-at-ccres-educational-and-behavioral-health-services-3739544400	A
https://www.linkedin.com/jobs/view/electrical-deputy-engineering-group-supervisor-at-energy-jobline-3773709557	E
https://www.linkedin.com/jobs/view/electrical-assembly-lead-at-sanmina-3704300377	E
https://www.linkedin.com/jobs/view/senior-lead-technician-programmer-at-security-101-3785441848	A
https://www.linkedin.com/jobs/view/program-consultant-at-methodist-family-health-3588621456	C
https://www.linkedin.com/jobs/view/veterinary-receptionist-at-wellhaven-pet-health-3803807922	V
https://www.linkedin.com/jobs/view/sr-technician-receiving-inspection-at-abbott-3799867135	O
https://www.linkedin.com/jobs/view/experienced-hvac-service-technician-at-lane-valente-industries-3798208587	H

only showing top 10 rows

job_skills
Building Custodial Services, Cleaning, Janitorial Services, Materials Handling, Housekeeping, Sanitation, Waste Management, Floor Maintenance, Equipment Mainter
Customer service, Restaurant management, Food safety, Training, Supervision, Scheduling, Inventory, Cost control, Sales, Communication, Problemsolving, Leadersh
Applied Behavior Analysis (ABA), Data analysis, Behavioral assessment, Positive behavior support, Programming development, Progress monitoring, Staff training,
Electrical Engineering, Project Controls, Scheduling, Estimating, Engineering Efforts, Planning, Work Packaging, Communication Skills, Verbal Communication, Wri
Electrical Assembly, Point to point wiring, Stripping and crimping of wiring, Reading blueprints and SOPs, Leadership skills, Communication skills, Directing ar
Access Control, Video Management Systems, IPbased video systems, Intrusion Alarm Systems, SQL, Databases, Genetec, Software House, Avigilon Unity, Avigilon Alta
Consultation, Supervision, InService Training, PreService Training, Youth Intake, Individual Treatment Plans, Program Development, Data Collection, Quality Cont
Veterinary Receptionist, AAHAaccredited, Customer service, Communication skills, Organizational skills, Problemsolving skills, Time management skills, Ability t
Optical Inspection Equipment Programming, MS Excel, Microsoft Word, FDA regulations, ISO 13485, Gage R&R, NCMR dispositioning, Blueprint reading, Sampling techr
HVAC, troubleshooting, Preventative maintenance, Inspections, Repairs, Record keeping, Parts management, Time management, Travel

Figure 2:first 10 rows of DataFrame after loading in PySpark environment

Due to the massive dataset size, I decided to work with a small fraction for modeling. This helps in managing computational resources efficiently. By selecting a small representative sample, we strike a balance between accuracy and resource constraints, allowing for meaningful analysis and interpretation.

4.1. Data preprocessing

In the next step, I focused on preparing the textual data for analysis. Firstly, I converted all text in the " job_skills " column to lowercase to ensure uniformity and consistency in text representation. Following this, I removed non-alphabetic characters using regular expressions, aiming to clean the text data from unnecessary symbols and characters. Subsequently, I tokenized the text, breaking it down into individual words, which facilitates further analysis. Lastly, I removed common stop words, such as "the," "is," and "and," to eliminate noise and focus on meaningful words for analysis.

To handle a sizable dataset, I divided the DataFrame into 8 partitions with the goal of improving computational efficiency. This was followed by redistributing the data across these partitions. This redistribution strategy aims to balance the workload and enhance parallel processing capabilities, thereby refining the performance of the market basket analysis model.

Displayed below is the Processed Data Frame, prepared and primed for analysis

```
+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|transactions
+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|[construction, management, engineering, architecture, design, build, project, management, client, relations, subcontractor, relations, scheduling, budgeting, cost, management, proj
|[nursing, patient, care, medical, nursing, care, plan, infection, control, nursing, assessment, resident, care, care, plans, nursing, diagnosis, quality, resident, care, educationa
|[merchandising, inventory, control, asset, control, safety, sanitation, baking, training, customer, service, employee, relations, sales, distribution, labor, control, shrink, contr
|[group, exercise, fitness, zumba, yoga, first, aid, cpr, aed]
|[curriculum, development, instructional, strategies, learning, styles, data, analysis, feedback, positive, learning, environment, classroom, culture, family, communication, profess
|[billing, analysis, invoice, creation, invoice, review, financial, reporting, customer, contract, interpretation, carrier, report, analysis, salesforce, management, billing, issue,
|[customer, service, retail, experience, pos, system, stocking, shelves, merchandise, organization, lifting, heavy, objects, working, flexible, schedule, communication, skills, prob
|[legal, advice, compliance, management, service, agreement, negotiation, transaction, document, review, compliance, system, design, monitoring, amlctf, program, support, breach, co
|[survey, equipment, topographical, surveys, measured, building, surveys, setting, surveys, survey, data, processing, data, capture, multidisciplined, survey, projects, project, man
|[microsoft, office, suite, outlook, word, excel, powerpoint, access, database, management, report, writing, complex, document, creation, data, analysis, english, language, written,
|[substitute, teacher, curriculum, adaptation, lesson, planning, unit, planning, rubric, creation, assessment, development, datadriven, instruction, feedback, provision, individual,
|[chemistry, ms, office, rd, chemical, formulation, technical, service, quality, control, quality, management, system, perseverance, collaboration, patience, attention, detail, chem
|[communication, skills, problemsolving, skills, conflict, management, skills, analytical, skills, observation, skills, numerical, reasoning, skills, leadership, skills, teamwork, s
|[customer, service, chairside, assisting, dental, imaging, infection, control, equipment, sterilization, clerical, functions, information, management, patient, safety, employee, ed
|[calendar, management, microsoft, office, suite, data, entry, faxing, filing, monitoring, paper, chart, organization, hipaa, compliance, em, tableau, scanning, printing, copying, r
|[communication, time, management, customer, service, organization, management, friendly, personality, problemsolving]
|[ekg, telemetry, ivs, foley, catheters, urinary, catheterization, bladder, scan, telemetry, specimen, blood, glucose, testing, vital, signs, pulse, oximetry, adls, wound, care, res
|[mri, imaging, radiography, nuclear, medicine, diagnostic, medical, sonography, american, registry, radiologic, technologists, arrt, certified, nuclear, medicine, technologist, cnm
|[civil, engineering, construction, management, project, management, building, construction, costeffective, construction, methods, subcontractor, management, contract, management, s
|[medical, assistant, patient, care, medical, history, taking, treatment, procedures, explanation, lab, specimen, collection, preparation, medication, preparation, physician, assist
+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
only showing top 20 rows
```

Figure 3:The DataFrame after preprocessing

4.2. Model results

In this segment, I utilized the FPGrowth algorithm, a technique commonly employed for mining frequent itemsets in large datasets.

Firstly, I ensured that each transaction within the DataFrame, represented by the 'transactions' column, contained only distinct elements. This step removes any duplicate items within transactions, ensuring accurate analysis.

Subsequently, the FPGrowth algorithm was applied to the DataFrame. This algorithm requires the specification of two crucial parameters: minimum support and minimum confidence.

Minimum Support (min_support): This parameter determines the minimum frequency threshold that an itemset must meet to be considered frequent. Items occurring less frequently than this threshold are disregarded.

Minimum Confidence (min_confidence): This parameter establishes the minimum confidence level required for association rules generated by the algorithm. Confidence measures the proportion of transactions containing the antecedent that also contain the consequent.

By adjusting these parameters, analysts can fine-tune the performance of the FPGrowth algorithm to better suit the specific characteristics of the dataset and the objectives of the analysis. The values assigned to min_support and min_confidence can be adjusted according to the desired level of granularity and the nature of the association rules sought.

Upon fitting the FPGrowth model to the DataFrame, the algorithm identifies frequent itemsets and generates association rules based on the specified parameters. These association rules capture meaningful relationships between items in the transactions, providing valuable insights for market basket analysis and other applications.

The figure depicted below provides a detailed exposition on the mathematical principles underlying the metrics utilized in the association rules algorithm. It offers an in-depth examination of how these metrics are calculated and interpreted, shedding light on the mathematical framework that drives the analysis of association rules.

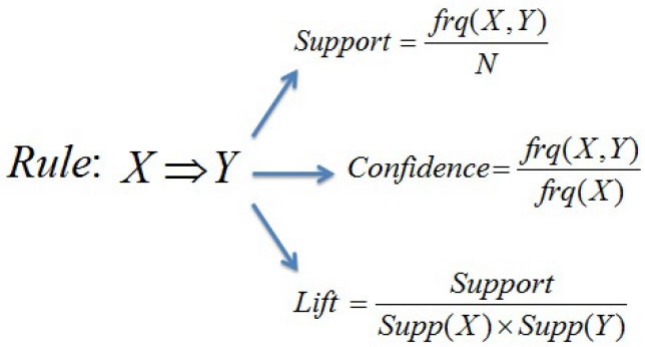

Figure 4:Metrics of Association rules algorithm [5]

This table showcases the top 20 rows, revealing the items along with their corresponding frequencies. It offers a concise yet insightful glimpse into the dataset, highlighting the prevalence of various items and their occurrences within the dataset.

Table 1:Item Frequencies

items	freq
[parental]	16
[parental, leave]	16
[reimbursement]	41
[reimbursement, insurance]	22
[reimbursement, vision]	15
[reimbursement, vision, insurance]	15
[reimbursement, employee]	16
[reimbursement, professional]	14
[reimbursement, medical]	19
[reimbursement, life]	17
[reimbursement, life, insurance]	16
[reimbursement, k]	14
[reimbursement, health]	16
[reimbursement, dental]	15
[reimbursement, dental, insurance]	15
[reimbursement, dental, vision]	15
[reimbursement, dental, vision, insurance]	15
[reimbursement, paid]	15
[reimbursement, program]	17
[reimbursement, license]	14

only showing top 20 rows

This table depicts the association rules generated by the algorithm along with their corresponding metrics, such as support and confidence.

Table 2:Association rules and their related metrics

Association Rules:

antecedent	consequent	confidence	lift	support
[school, high, license, service]	[customer]	1.0	3.586021505376344	0.015742128935532233
[school, high, license, service]	[communication]	0.7619047619047619	1.371634213739477	0.01199400299850075
[school, high, license, service]	[drivers]	0.9047619047619048	12.977982590885816	0.01424287856071964
[school, high, license, service]	[diploma]	0.9047619047619048	8.683110654333676	0.01424287856071964
[analysis, leadership, experience, management, communication]	[skills]	0.8260869565217391	2.1778656126482216	0.01424287856071964
[school, high, degree, customer, service]	[communication]	1.0	1.8002699055330635	0.01199400299850075
[microsoft, , experience, management]	[communication]	0.7272727272727273	1.3092872040240462	0.01199400299850075
[microsoft, , experience, management]	[office]	0.7727272727272727	4.643325143325144	0.012743628185907047
[microsoft, , experience, management]	[years]	0.6818181818181818	9.3767572633552	0.011244377811094454
[diploma, school, teamwork, service, management, communication]	[customer]	1.0	3.586021505376344	0.01199400299850075
[diploma, school, teamwork, service, management, communication]	[high]	1.0	7.801169590643275	0.01199400299850075
[analysis, sales, service, management]	[communication]	1.0	1.8002699055330635	0.010494752623688156
[merchandising, retail, leadership, communication]	[customer]	1.0	3.586021505376344	0.011244377811094454
[merchandising, retail, leadership, communication]	[service]	1.0	3.3857868020304567	0.011244377811094454
[diploma, school, customer, service, experience, skills]	[sales]	0.5161290322580645	3.681904433327583	0.01199400299850075
[diploma, school, customer, service, experience, skills]	[management]	0.5483870967741935	1.1100885995398697	0.012743628185907047
[diploma, school, customer, service, experience, skills]	[communication]	0.9032258064516129	1.6260502372556702	0.020989505247376312
[diploma, school, customer, service, experience, skills]	[high]	1.0	7.801169590643275	0.023238380809595203
[attention, ability, teamwork]	[detail]	1.0	9.200000000000001	0.01649175412293853
[attention, ability, teamwork]	[problemsolving]	0.6363636363636364	4.742508887760284	0.010494752623688156

only showing top 20 rows

The presented association rules reveal significant relationships between different terms or concepts in the dataset. Each rule consists of an antecedent (the condition or set of conditions) and a consequent (the outcome or result). Here's an interpretation of the metrics:

Antecedent: This is the first part of an association rule. It represents a condition or combination of items that occur together. In your dataset, these are sets of skills or attributes related to LinkedIn job postings.

Consequent: This is the second part of an association rule. It represents an item or set of items that are likely to be found in conjunction with the antecedent. In this context, it could be the outcome or skill that is associated with the antecedent skills.

Confidence: This metric measures the reliability of the inference made by the rule. In simpler terms, it is the probability of seeing the consequent in a transaction given that it also contains the antecedent. A confidence of 1.0 means that every time the antecedent occurs, the consequent occurs as well.

Lift: This measures how much more often the antecedent and consequent of the rule occur together than we would expect if they were statistically independent. A lift value greater than 1 indicates that the antecedent and consequent appear together more often than expected, which suggests a strong association between them.

Support: This represents the proportion of transactions in the dataset that contain both the antecedent and the consequent. It is a measure of how frequently the association rule occurs in the dataset.

To interpret a single rule from the table:

For the rule where the antecedent is [school, high, license, service] and the consequent is [customer], the confidence is 1.0, which indicates that in all cases where the antecedent appeared, the consequent [customer] also appeared. The lift is approximately 3.586, which suggests that [customer] is more likely to occur when [school, high, license, service] occurs than it would by chance. The support of about 0.0157 indicates that this rule is relevant for about 1.57% of the transactions in the dataset.

5. Conclusion:

This report presented a detailed examination of market basket analysis applied to a significant dataset representing the job market landscape on LinkedIn. Utilizing the FP-Growth algorithm within the PySpark framework allowed for the efficient mining of frequent itemsets, demonstrating the co-occurrence of specific job-related skills. The results highlighted crucial relationships between various skills, suggesting trends and combinations that are valued in the job market. These insights can aid in guiding workforce development, informing educational focus, and shaping recruitment strategies. The implementation of this analysis showcases the potential of large-scale data mining in extracting meaningful patterns and supports the strategic decisions within the professional and educational realms. Future work could explore optimizing the process for even larger datasets and delving deeper into the implications of these associations on job market forecasting.

6. Declaration by author

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. 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.

References

[1] Hossain, Maliha, AHM Sarowar Sattar, and Mahit Kumar Paul. "Market basket analysis using apriori and FP growth algorithm." 2019 22nd international conference on computer and information technology (ICCIT). IEEE, 2019.

[2] Mythili, M. S., and AR Mohamed Shanavas. "Performance evaluation of apriori and fp-growth algorithms." International Journal of Computer Applications 79.10 (2013).

[3] K, Joos. “The FP Growth Algorithm | Towards Data Science.” Medium, 5 Jan. 2022, towardsdatascience.com/the-fp-growth-algorithm-1ffa20e839b8.

[4] Ali, Amir. “Association Rule(Apriori and FP-Growth Algorithms) With Practical Implementation.” Medium, 7 Dec. 2021, medium.com/machine-learning-researcher/association-rule-apriori-and-eclat-algorithm-4e963fa972a4.

[5] Khanna, Sakshi. “What Are Association Rules in Data Mining?” Analytics Vidhya, 16 Feb. 2024, www.analyticsvidhya.com/blog/2023/11/what-are-association-rules-in-data-mining.

[6] Firmansyah, Firmansyah. "Market Basket Analysis for Books Sales Promotion using FP Growth Algorithm, Case Study: Gramedia Matraman Jakarta." Journal of Informatics and Telecommunication Engineering 4.2 (2021): 383-392.