Market-Basket Analysis for MeDAL Dataset

Fatma Er, 967585

March 05, 2023

Abstract

The aim of this project is to develop a system for identifying frequent itemsets and conducting market basket analysis on the Medal dataset. Specifically, the analysis focuses on treating text as baskets and words as items, with the goal of identifying patterns and relationships within the dataset. To achieve this, we applied the Frequent Pattern Growth Algorithm, which allowed us to identify frequent item sets and association rules. By conducting this analysis, we gained valuable insights into the structure and content of the dataset, which can inform future research and decision-making.

Key words: Market-Basket Analysis and FP Growth.

1 Introduction

Association algorithms are widely used in retail analysis of transactions, recommendation engines, and online click stream analysis across web pages, etc. One of the popular applications of this technique is called market basket analysis, which finds co-occurrences of one retail item with another item within the same retail purchase transaction (Kotu and Deshpande, 2019).

Market Basket Analysis (MBA) is an accidental transaction pattern that purchasing some products will affect the purchasing of other products. MBA is used to predict what products that customer interested in. MBA has three parameters which are support, confidence, and lift. Support is a proportion of event B because of event A. Confidence is a probability event B happened because of event A dependently. Lift is a probability of event B happened because of event A independently (Halim et al., 2019).

There are different stages that are attached with the market basket analyses. For this study, firstly pre-processing of the text preparing (Tokenization, Normaliszation, Lemmatization, Clean Stopwords), after this process using Apriori ,FP Growth,SETM Algorithm and AIS algorithms for undertandin support,confidence,lift and compare results. In this project, we prefereed Apriori and FP Growth algoritms.

The input to our models are related Medal Dataset.We analysed the performance of FP Growth algorithms.

The paper is organized as follows: Section II describes the related work and background acknowledge about Market Basket Basket Analysis and Section III introduces the dataset and features we are using in this project. The methods, models and text processing section IV, followed by presenting the results and discussion details about each model and draw some perspectives about the future work section V.

2 RELATED WORKS

In this section, studies published in the literature on market basket analysis are reviewed.

(Mustakim et al., 2018) proposed a model for Market Basket Analysis using the FP-Growth algorithm to plan and optimize product availability. The study aimed to identify consumer spending patterns at Berkah Mart in Pekanbaru, by analyzing 8,307 items sold in December 2017, with an average of 400 daily transactions, and between 1 and 20 items per transaction. The FP-Growth algorithm proved to be useful in generating informative association rules, which were instrumental in determining the layout and planning of goods availability. The experimental results showed that both the FP-Growth and Apriori algorithms can increase the overall profit of Berkah Mart by analyzing consumer spending patterns. However, the FP-Growth algorithm is recommended as it has a faster processing speed and generates rules with superior support and confidence values. Therefore, it is an effective tool for optimizing product availability and increasing the profitability of the retail store.

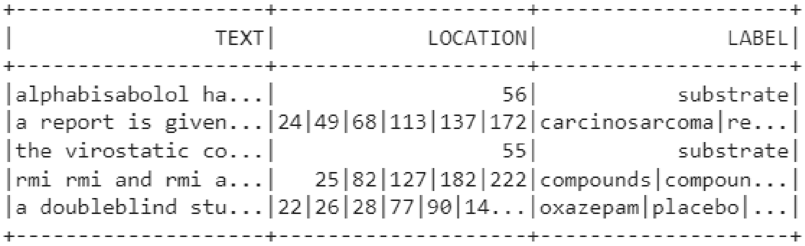
(Patil and Khot, 2022) proposed a discussion on Association rules and use of Apriori principle used for Market basket Analysis. Data mining provides the way to use précised information from the large dataset. Association rules find the relationship between items by analyzing the data and provide the accurate solution to the retailer to make better business decisions.

(Patil and Khot, 2022) proposed a method to identify the relationship between a set of selling attributes using a market basket analysis approach. Their objective was to determine the patterns of relationships that exist in transaction data for outdoor goods. The Apriori and frequent pattern growth (FP-growth) algorithms were employed for analysis, and association rules were generated. The study produced ten rules using the Apriori algorithm and four rules using the FP-growth algorithm. One of the association rules generated was "if a consumer purchases a portable stove, they are likely to purchase portable gas as well." The strength of this rule was measured with a minimum support of 0.296 and a confidence level of 0.774 using the Apriori algorithm and 0.296 and 0.750 using the FP-growth algori

3 DATASET AND FEATURES

3.1 Medal Dataset

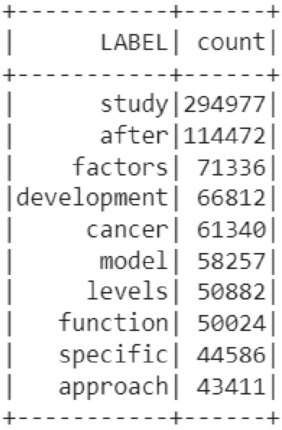
This project analyzed the Medal Dataset, which contains 14.393.619 lines of text data. The dataset has three features, namely "text", "location", and "label". The "text" feature contains the full text of the data files, while the "location" feature indicates where the data was collected. The "label" feature provides information on the classification of the data, which was used for our analysis.



```
+--------------------+--------------------+--------------------+
|                TEXT|            LOCATION|               LABEL|
+--------------------+--------------------+--------------------+
|alphabisabolol ha...|                  56|           substrate|
|a report is given...|24|49|68|113|137|172|carcinosarcoma|re...|
|the virostatic co...|                  55|           substrate|
|rmi rmi and rmi a...|    25|82|127|182|222|compounds|compoun...|
|a doubleblind stu...|22|26|28|77|90|14...|oxazepam|placebo|...|
+--------------------+--------------------+--------------------+
```

Figure 1: Data Structure

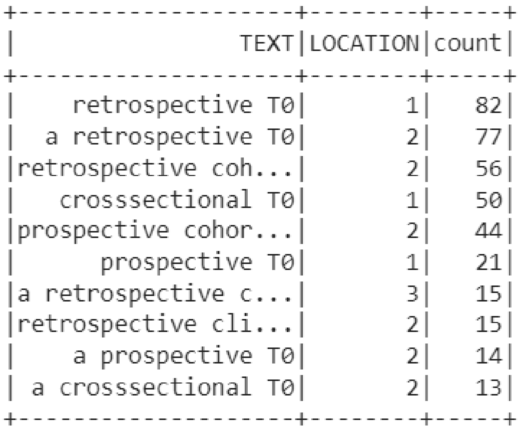
The dataset does not contain any blank rows and includes various labels.



```
+-----------+------+
|      LABEL| count|
+-----------+------+
|      study|294977|
|      after|114472|
|    factors| 71336|
|development| 66812|
|     cancer| 61340|
|      model| 58257|
|     levels| 50882|
|   function| 50024|
|   specific| 44586|
|   approach| 43411|
+-----------+------+
```

Figure 2: Label Distribution (Top)

As shown in Figure 2, the dataset consists of different labels. For our analysis, we focused on the "study" label, which contained a total of 294,977 rows of text.



```
+--------------------+--------+-----+
|                TEXT|LOCATION|count|
+--------------------+--------+-----+
|    retrospective T0|       1|   82|
|  a retrospective T0|       2|   77|
|retrospective coh...|       2|   56|
|   crosssectional T0|       1|   50|
|prospective cohor...|       2|   44|
|      prospective T0|       1|   21|
|a retrospective c...|       3|   15|
|retrospective cli...|       2|   15|
|    a prospective T0|       2|   14|
| a crosssectional T0|       2|   13|
+--------------------+--------+-----+
```

Figure 3: Duplicate text (Top)

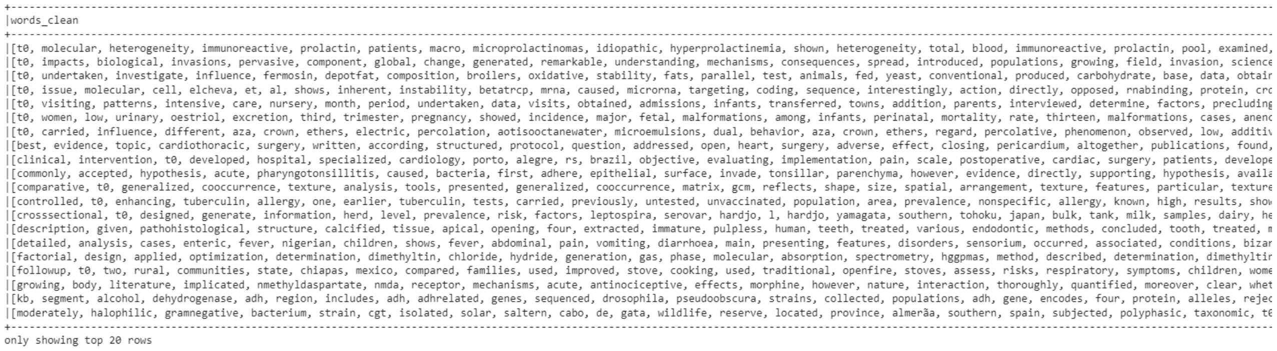
To work with the "text" feature, we also need to clean the data from any duplicates. After removing the duplicate texts, our dataset contains 294,266 rows of data.

3.2 Data Processing Cleaning Text

In order to prepare the text data for analysis, we need to apply natural language processing (NLP) techniques. We used Spark libraries for NLP. The NLP steps we applied to the data are Tokenization, Normalization, Lemmatization, and Stopword Removal.

Tokenization refers to the process of splitting the text into individual words or tokens. Normalization is used to convert all the text to lowercase to ensure consistency. Lemmatization is the process of reducing words to their base form or lemma. Finally, stopword removal involves removing common words such as "the," "and," and "in" as they do not provide much meaning to the text.

By applying these NLP steps to the data, we can preprocess the text and transform it into a format that is suitable for further analysis.



```
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|words_clean
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|[t0, molecular, heterogeneity, immunoreactive, prolactin, patients, macro, microprolactinomas, idiopathic, hyperprolactinemia, shown, heterogeneity, total, blood, immunoreactive, prolactin, pool, examined,
|[t0, impacts, biological, invasions, pervasive, component, global, change, generated, remarkable, understanding, mechanisms, consequences, spread, introduced, populations, growing, field, invasion, science
|[t0, undertaken, investigate, influence, fermosin, depotfat, composition, broilers, oxidative, stability, fats, parallel, test, animals, fed, yeast, conventional, produced, carbohydrate, base, data, obtair
|[t0, issue, molecular, cell, elcheva, et, al, shows, inherent, instability, betatrcp, mrna, caused, microrna, targeting, coding, sequence, interestingly, action, directly, opposed, rnabinding, protein, crc
|[t0, visiting, patterns, intensive, care, nursery, month, period, undertaken, data, visits, obtained, admissions, infants, transferred, towns, addition, parents, interviewed, determine, factors, precluding
|[t0, women, low, urinary, oestriol, excretion, third, trimester, pregnancy, showed, incidence, major, fetal, malformations, among, infants, perinatal, mortality, rate, thirteen, malformations, cases, anenc
|[t0, carried, influence, different, aza, crown, ethers, electric, percolation, aotisooctanewater, microemulsions, dual, behavior, aza, crown, ethers, regard, percolative, phenomenon, observed, low, additiv
|[best, evidence, topic, cardiothoracic, surgery, written, according, structured, protocol, question, addressed, open, heart, surgery, adverse, effect, closing, pericardium, altogether, publications, found,
|[clinical, intervention, t0, developed, hospital, specialized, cardiology, porto, alegre, rs, brazil, objective, evaluating, implementation, pain, scale, postoperative, cardiac, surgery, patients, develope
|[commonly, accepted, hypothesis, acute, pharyngotonsillitis, caused, bacteria, first, adhere, epithelial, surface, invade, tonsillar, parenchyma, however, evidence, directly, supporting, hypothesis, availa
|[comparative, t0, generalized, cooccurrence, texture, analysis, tools, presented, generalized, cooccurrence, matrix, gcm, reflects, shape, size, spatial, arrangement, texture, features, particular, texture
|[controlled, t0, enhancing, tuberculin, allergy, one, earlier, tuberculin, tests, carried, previously, untested, unvaccinated, population, area, prevalence, nonspecific, allergy, known, high, results, show
|[crosssectional, t0, designed, generate, information, herd, level, prevalence, risk, factors, leptospira, serovar, hardjo, l, hardjo, yamagata, southern, tohoku, japan, bulk, tank, milk, samples, dairy, he
|[description, given, pathohistological, structure, calcified, tissue, apical, opening, four, extracted, immature, pulpless, human, teeth, treated, various, endodontic, methods, concluded, tooth, treated, m
|[detailed, analysis, cases, enteric, fever, nigerian, children, shows, fever, abdominal, pain, vomiting, diarrhoea, main, presenting, features, disorders, sensorium, occurred, associated, conditions, bizar
|[factorial, design, applied, optimization, determination, dimethyltin, chloride, hydride, generation, gas, phase, molecular, absorption, spectrometry, hggpmas, method, described, determination, dimethyltir
|[followup, t0, two, rural, communities, state, chiapas, mexico, compared, families, used, improved, stove, cooking, used, traditional, openfire, stoves, assess, risks, respiratory, symptoms, children, wome
|[growing, body, literature, implicated, nmethyldaspartate, nmda, receptor, mechanisms, acute, antinociceptive, effects, morphine, however, nature, interaction, thoroughly, quantified, moreover, clear, whet
|[kb, segment, alcohol, dehydrogenase, adh, region, includes, adh, adhrelated, genes, sequenced, drosophila, pseudoobscura, strains, collected, populations, adh, gene, encodes, four, protein, alleles, rejec
|[moderately, halophilic, gramnegative, bacterium, strain, cgt, isolated, solar, saltern, cabo, de, gata, wildlife, reserve, located, province, almerãa, southern, spain, subjected, polyphasic, taxonomic, t0
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
only showing top 20 rows
```

Figure 4: Tokenization text

Figure 4 displays the tokens obtained after performing NLP steps such as Tokenization, Normalization, Lemmatization, and cleaning stopwords. The resulting tokens are used for market basket analysis.

4 METHODS

In this section, we discussedd Market Basket Analysis.

4.1 Market Basket Analysis and Algorithms

Frequent itemset mining is a technique that allows for the identification of relationships and correlations between items within large transactional or relational datasets. As vast amounts of data are constantly being collected and stored, many industries are recognizing the importance of extracting these patterns from their databases. The disclosure of correlation relationships across multiple transaction records can be extremely valuable in various decision-making processes, such as catalog design, cross-marketing, and customer exchange analysis(Kadlaskar).

Market Basket Analysis (MBA) is considered to be a popular data mining tool to improve many business decisions(Jirapatsil and Phumchusri, 2022).

Types of Market Basket Analysis in Data Mining. There are three types of market basket analysis in data mining:

Descriptive Market Basket Analysis This approach of the most popular Market Basket Analysis in Data Mining draws its conclusions from previous data. The analysis does not make any predictions, but instead uses statistical approaches to rate the relationship between items. It uses the unsupervised data mining model.

Predictive Market Basket Analysis In Data Mining this Market Basket analysis uses supervised data mining models such as classification and regression. Its primary purpose is to imitate the market to find out what's causing things to happen. Essentially, it takes into account products purchased in a particular order to calculate cross-selling.

Differential Market Basket Analysis This type of market basket analysis in data mining helps in the analysis of competitors. To uncover fascinating patterns in consumer behavior, stores compare purchase histories between seasons, periods, days of the week, and other variables(Ganiyu).

There are Multiple Techniques and Algorithms Used in Market Basket Analysis.

1)Apriori Algorithm

2)AIS Algorithm

3)SETM Algorithm

4)FP Growth Algorithm

1. Apriori Algorithm

Market Basket Analysis can be implemented by using the Apriori algorithm . This algorithm determines the frequent data or item set from the transaction database. It determines the frequent item sets from the database using candidate item set generation(Rao and Kiran, 2021).

The Apriori Algorithm was proposed by Agrawal and Srikant in 1994. Apriori is designed to work in databases containing transactions (for example, collections of items purchased by customers, or details of a website visit or IP addresses) (Jirapatsil and Phumchusri, 2022).

The Apriori algorithm is a level-based, breadth-first algorithm that counts operations containing prior knowledge of frequent item set properties. Apriori uses an iterative approach known as level-level search, in which n itemsets are used to discover (n+1) itemsets. The Apriori property is used here to increase the efficiency of frequent itemsets by level. The apriori property insists that all non-empty subsets of a frequent itemset must also be frequent. This is due to the anti-monotonic nature of the support measure. Support for a set of items never exceeds Support for its subsets. A two-stage joining and pruning process is done iteratively (Annie and Kumar, 2012).

Limitations of the Apriori Algorithm

Although the Apriori algorithm is clear and simple, it has some weaknesses. The main limitation is very frequent item sets, low minimum support, or costly wasting of time to hold large item sets and large number of candidate sets. For example, if there are 104 of the frequent 1 item set, it should generate more than 107 candidates with length 2 to be sequentially tested and accumulated. Also, detecting frequent patterns of size 100 (eg) v1, v2…v100 would require generating 2100 candidate item-sets, which is costly and wastes candidate generation time. Therefore, it will check many candidate item sets and also iteratively scan the database to find candidate item sets. Apriori will be very low and inefficient when memory capacity is limited to a large number of processes(Al-Maolegi and Arkok, 2014).

2) AIS Algorithm

The AIS algorithm is the first algorithm developed and published by Agrawal, Imielinski and Swami in 1993 to generate all common product sets in the database. The algorithm is focused on increasing the functionality of databases to make decision support queries. It carries the restriction of ordering the product names in the database from A to Z(Dunham et al., 2001).

The AIS algorithm creates multiple passes over the entire database or transactional data. Scans all processes on each pass. On the first pass, it counts the support of individual items and then determines which ones are frequent in the database. Huge sets of items in each pass are scaled up to create candidate itemsets. After each scan of a transaction, it identifies common itemsets between those itemsets of the previous pass, and then items from that transaction(Kadlaskar).

Advantage: It is convenient to use the AIS algorithm to find out if there is a relationship between the items.

Disadvantage: The main disadvantage of the AIS algorithm is that it generates too many candidate sets after it turns out to be small.

3)SETM Algorithm

SETMs were proposed by Houtsmal in 1995 for aggregation of frequent clusters(Dunham et al., 2001).

This Algorithm is quite similar to the AIS algorithm. The SETM algorithm creates batch migrations on the database. On the first pass, it counts the support of individual items and then determines which ones are frequent in the database. Then it also creates candidate itemsets by extending the large itemsets of the previous pass. In addition, the SETM algorithm recalls TIDs (transaction IDs) of transactions created with candidate itemsets(Kadlaskar).

Advantage: When generating candidate itemsets, the SETM algorithm arranges the candidate itemsets sequentially along with their TID(transaction ID).

Disadvantage: There is a relationship to the Tid for each item set, so it requires more space to store a large number of TIDs.

SETM algoritması da AIS algoritmasında olduğu gibi bir çok kez tarama yapar. Bu iki algoritmada gereksiz aday oluşturduğu için çok tercih edilen algoritmalar değildir(Kumbhare and Chobe, 2014).

4)FP Growth Algorithm

FP Growth is known as Frequent Pattern Growth Algorithm. The FP growth algorithm is a concept that represents data in the form of an FP tree or Frequent Pattern. So FP Growth is a method of Mining Frequent Item Sets. This algorithm is an improvement over the Apriori Algorithm. No candidate generation is required to create a frequent pattern. This frequent pattern tree structure preserves the relationship between sets of items(Kadlaskar).

The FP-growth algorithm is used to overcome the two disadvantages of the Apriori algorithm. FP growth requires the creation of the FP tree. This requires two passes. FPgrowth uses the divide and conquer strategy. Requires the database to be scanned twice. First, it calculates a list of frequently used items in descending order (F-List) and sorted by frequency during the initial database scan. In the second scan, the database is compressed into an FP-tree. This algorithm performs recursive mining on the FP-tree. There is a problem with finding frequent sets of items that are iteratively converted into searching and tree building. Favorite itemsets are created with only two passes from the database and no lead generation process. The pattern generation process has two sub-processes: generating the FP-tree and generating the pattern from the FP-tree(Kumbhare and Chobe, 2014).

4.2 Understanding Association Rule Terminology

Association rule mining uses special terminology to refer to the items on either side of the rule(Clarke). Association analysis measures the strength of co-occurrence between one item and another. The objective of this class of data science algorithms is not to predict an occurrence of an item, like classification or regression algorithms do, but to find usable patterns in the co-occurrences of the items. Association rules learning is a branch of unsupervised learning processes that discover hidden patterns in data, in the form of easily recognizable rules(Kotu and Deshpande, 2019).

The model outcome of an association analysis can be represented as a set of rules, like the one below:

{Item A} - {Item B}

This rule indicates that based on the history of all the transactions, when Item A is found in a transaction or a basket, there is a strong propensity of the occurrence of Item B within the same transaction. Item A is the antecedent or premise of the rule and

Item B is the consequent or conclusion of the rule. The antecedent and consequent of the rule can contain more than one item,like Item A and Item C. To mine these kinds of rules from the data, previous customer purchase transactions would need to be analyzed. In a retail business, there would be millions of transactions made in a day with thousands of stock keeping units, which are unique for an item(Kotu and Deshpande, 2019).

Support

The support metric indicates how frequently the itemset occurs within the dataset(Clarke).The support of an item is simply the relative frequency of occurrence of an itemset in the transaction set(Kotu and Deshpande, 2019).

Confidence

The confidence value tells you how often the rule proves to be true(Clarke).The confidence of a rule measures the likelihood of the occurrence of the consequent of the rule out of all the transactions that contain the antecedent of the rule.Confidence provides the reliability measure of the rule(Kotu and Deshpande, 2019). Confidence of the rule (X-Y) is calculated by

Confidence (X-Y) = Support (X U Y) / Support (X)

Lift

Lift is essentially the support (or the probability of all the items in a rule occurring together) divided by the product of the probabilities of the items on either side appearing if there was no association. The higher the lift, the stronger the association between the antecedent and the consequent(Clarke). Though confidence of the rule is widely used, the frequency of occurrence of a rule consequent (conclusion) is largely ignored. In some transaction itemsets, this can provide spurious scrupulous rule sets because of the presence of infrequent items in the rule consequent. To solve this, the support of a consequent can be put in the denominator of a confidence calculation. This measure is called the lift of the rule(Kotu and Deshpande, 2019).

Life (X-Y)= Support (X U Y) / Support (X) * Support (Y)

Rule Generation

1. Finding all frequent itemsets. For an association analysis of n items it is possible to find $2^n - 1$ itemsets excluding the null itemset. As the number of items increase, there is an exponential increase in the number of itemsets. Hence it is critical to set a minimal support threshold to discard less frequently occurring itemsets in the transaction universe.

2. Extracting rules from frequent itemsets. For the dataset with n items it is

possible to find $3^n + 2^n + 1 + 1$rules This step extracts all the rules with a confidence higher than a minimum confidence threshold.

Hyperparameters

minSupport - The minimum support of an item to be considered in a frequent itemset.

minConfidence - The minimum confidence for generating an association rule from an itemset.

This two-step process generates hundreds of rules even for a small dataset with dozens of items. Hence, it is important to set a reasonable support and confidence threshold to filter out less frequent and less relevant rules in the search space. The generated rules can also be evaluated with support, confidence, lift, and conviction measures. In terms of computational requirements, finding all the frequent itemsets above a support threshold is more expensive than extracting the rules. Fortunately, there are some algorithmic approaches to efficiently find the frequent itemsets. The Apriori and Frequent Pattern (FP)-Growth algorithms are two of the most popular association analysis algorithms (Kotu and Deshpande, 2019).

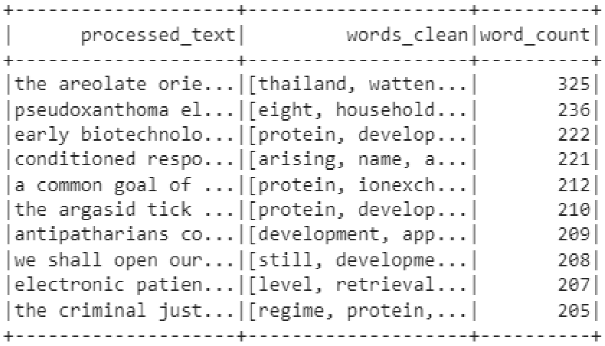
4.3 Models, Market Basket Analyses

This study employed the FP Growth Algorithm, a widely used method in market basket analysis, to conduct the analysis. The analysis was performed using Python and Pyspark libraries, specifically FPGrowth. The results were reported with an emphasis on interpreting key metrics such as consequent, confidence, lift, and support.

4.4 Market Basket Analysis with FP Growth Algorithm

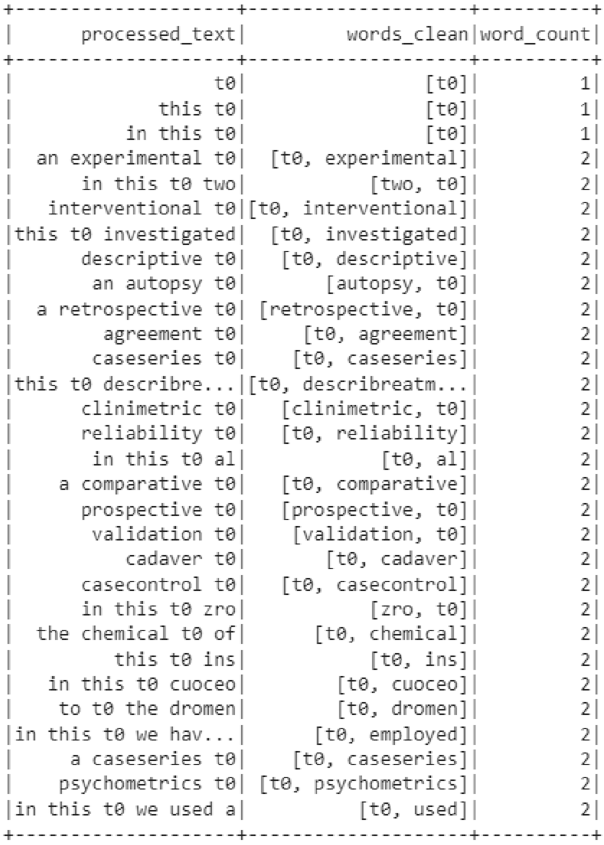
The market basket analysis was conducted using the Medal Dataset, which consists of 294,266 lines of text and "study" label. For analyses our basket was "words clean".

When attempting to use the full dataset, we encountered issues with performance. As a result, we calculated the "word count" values to understand the number of words included in each basket. This approach proved to be more effective for our analysis.



```
+--------------------+--------------------+----------+
|      processed_text|         words_clean|word_count|
+--------------------+--------------------+----------+
|the areolate orie...|[thailand, watten...|       325|
|pseudoxanthoma el...|[eight, household...|       236|
|early biotechnolo...|[protein, develop...|       222|
|conditioned respo...|[arising, name, a...|       221|
|a common goal of ...|[protein, ionexch...|       212|
|the argasid tick ...|[protein, develop...|       210|
|antipatharians co...|[development, app...|       209|
|we shall open our...|[still, developme...|       208|
|electronic patien...|[level, retrieval...|       207|
|the criminal just...|[regime, protein,...|       205|
+--------------------+--------------------+----------+
```

Figure 5: Word Count (desc)



```
+--------------------+--------------------+----------+
|      processed_text|         words_clean|word_count|
+--------------------+--------------------+----------+
|                  t0|                [t0]|         1|
|             this t0|                [t0]|         1|
|          in this t0|                [t0]|         1|
|  an experimental t0|  [t0, experimental]|         2|
|      in this t0 two|           [two, t0]|         2|
|   interventional t0|[t0, interventional]|         2|
|this t0 investigated|   [t0, investigated]|         2|
|      descriptive t0|    [t0, descriptive]|         2|
|       an autopsy t0|       [autopsy, t0]|         2|
|  a retrospective t0| [retrospective, t0]|         2|
|        agreement t0|      [t0, agreement]|         2|
|       caseseries t0|     [t0, caseseries]|         2|
|this t0 describre...|[t0, describreatm...|         2|
|       clinimetric t0|    [clinimetric, t0]|         2|
|       reliability t0|    [t0, reliability]|         2|
|       in this t0 al|             [t0, al]|         2|
|    a comparative t0|    [t0, comparative]|         2|
|       prospective t0|    [prospective, t0]|         2|
|        validation t0|     [validation, t0]|         2|
|           cadaver t0|        [t0, cadaver]|         2|
|       casecontrol t0|    [t0, casecontrol]|         2|
|      in this t0 zro|            [zro, t0]|         2|
|  the chemical t0 of|       [t0, chemical]|         2|
|         this t0 ins|            [t0, ins]|         2|
|   in this t0 cuoceo|         [t0, cuoceo]|         2|
|    to t0 the dromen|         [t0, dromen]|         2|
|in this t0 we hav...|       [t0, employed]|         2|
|     a caseseries t0|     [t0, caseseries]|         2|
|    psychometrics t0| [t0, psychometrics]|         2|
|in this t0 we used a|           [t0, used]|         2|
+--------------------+--------------------+----------+
```

Figure 6: Word Count (asc)

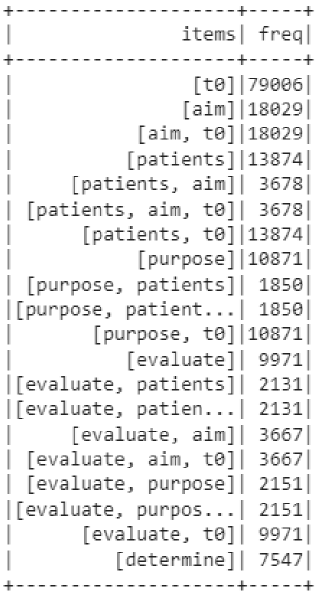
Figures 5 and 6 illustrate the distribution of word counts within our dataset, which was calculated as part of our data filtering process. These figures provide insight into the number of words included in each text, and allow us to better understand the structure and content of our dataset. By visualizing this distribution, we can identify any potential outliers or trends within the data, which can inform our analysis and improve the accuracy of our results.

To ensure the quality of our analysis, we filtered the data based on word count. Specifically, we excluded baskets with a word count less than 5 or greater than 50 % of the distribution. This approach allowed us to remove outliers and extremely lengthy texts from our dataset, which could have negatively impacted our results. By focusing on baskets within this range, we were able to improve the accuracy and reliability of our analysis.

Following the data filtering process, our analysis continued with a dataset consisting of 79,006 lines of text. This allowed us to focus on a more manageable and refined subset of the original data, which could be analyzed with greater accuracy and precision. We continued to use the FP Growth Algorithm and Python/Pyspark libraries for the analysis, with an emphasis on key metrics such as consequent, confidence, lift, and support.

The FPGrowth algorithm was used for the analysis with hyperparameters set to a minimum support of 0.02 and minimum confidence of 0.02. Using this approach, we calculated key Market Basket Measures such as Support, Lift, and Confidence, which allowed us to gain insight into the relationships and patterns within our dataset. As part of the data filtering process, we identified the most frequently published text words, which included t0, aim, patient, purpose, and evaluate. By focusing on these words, we were able to further refine our analysis and gain deeper insights into the underlying trends and patterns within the dataset.

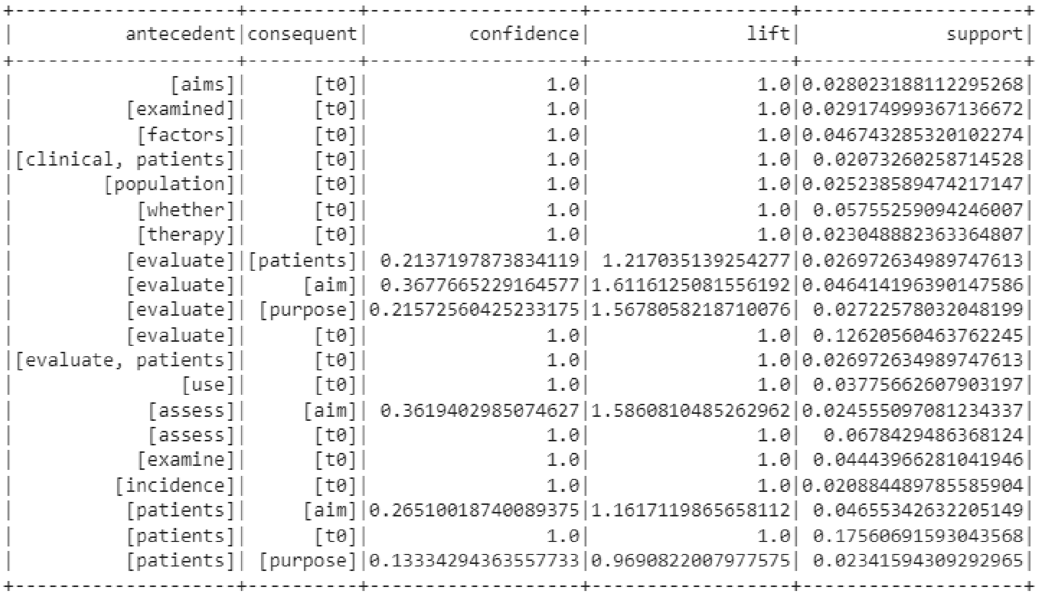
The output from the FP Growth Algorithm has been presented in figure 7 and 8. These results show the frequent item sets and their corresponding frequencies, which are calculated using the minimum support threshold of 0.02. These frequent item sets can be used to generate association rules, which can provide insights into the relationships between different items in the dataset. By analyzing the support, confidence, and lift metrics for each association rule, we can gain a deeper understanding of the underlying patterns and trends in the data.



```
+--------------------+-----+
|               items| freq|
+--------------------+-----+
|                [t0]|79006|
|               [aim]|18029|
|           [aim, t0]|18029|
|          [patients]|13874|
|     [patients, aim]| 3678|
| [patients, aim, t0]| 3678|
|      [patients, t0]|13874|
|           [purpose]|10871|
| [purpose, patients]| 1850|
|[purpose, patient...| 1850|
|       [purpose, t0]|10871|
|          [evaluate]| 9971|
|[evaluate, patients]| 2131|
|[evaluate, patien...| 2131|
|     [evaluate, aim]| 3667|
| [evaluate, aim, t0]| 3667|
| [evaluate, purpose]| 2151|
|[evaluate, purpos...| 2151|
|      [evaluate, t0]| 9971|
|         [determine]| 7547|
+--------------------+-----+
```

Figure 7: Frequent Items

The results obtained from the FP Growth Algorithm show the frequent itemsets and their corresponding frequencies. For example, the item "t0" appears in all 79,006 baskets. The itemset "aim, t0" appears in all 18,029 baskets where "aim" also appears. The support metric indicates how frequently an itemset appears in the dataset. Additionally, association rules can be derived from the frequent itemsets by setting a minimum threshold for the confidence and lift metrics. These association rules can provide insights into the relationships between items in the dataset.



```
+--------------------+----------+-------------------+------------------+--------------------+
|          antecedent|consequent|         confidence|              lift|             support|
+--------------------+----------+-------------------+------------------+--------------------+
|              [aims]|      [t0]|                1.0|               1.0|0.028023188112295268|
|          [examined]|      [t0]|                1.0|               1.0|0.029174999367136672|
|           [factors]|      [t0]|                1.0|               1.0|0.046743285320102274|
|[clinical, patients]|      [t0]|                1.0|               1.0| 0.02073260258714528|
|        [population]|      [t0]|                1.0|               1.0|0.025238589474217147|
|           [whether]|      [t0]|                1.0|               1.0| 0.05755259094246007|
|           [therapy]|      [t0]|                1.0|               1.0|0.023048882363364807|
|          [evaluate]|[patients]| 0.2137197873834119| 1.217035139254277|0.026972634989747613|
|          [evaluate]|     [aim]| 0.3677665229164577|1.6116125081556192|0.046414196390147586|
|          [evaluate]| [purpose]|0.21572560425233175|1.5678058218710076| 0.02722578032048199|
|          [evaluate]|      [t0]|                1.0|               1.0| 0.12620560463762245|
|[evaluate, patients]|      [t0]|                1.0|               1.0|0.026972634989747613|
|               [use]|      [t0]|                1.0|               1.0| 0.03775662607903197|
|            [assess]|     [aim]| 0.3619402985074627|1.5860810485262962|0.024555097081234337|
|            [assess]|      [t0]|                1.0|               1.0|  0.0678429486368124|
|           [examine]|      [t0]|                1.0|               1.0| 0.04443966281041946|
|         [incidence]|      [t0]|                1.0|               1.0|0.020884489785585904|
|          [patients]|     [aim]|0.26510018740089375|1.1617119865658112| 0.04655342632205149|
|          [patients]|      [t0]|                1.0|               1.0| 0.17560691593043568|
|          [patients]| [purpose]|0.13334294363557733|0.9690822007977575| 0.02341594309292965|
+--------------------+----------+-------------------+------------------+--------------------+
```

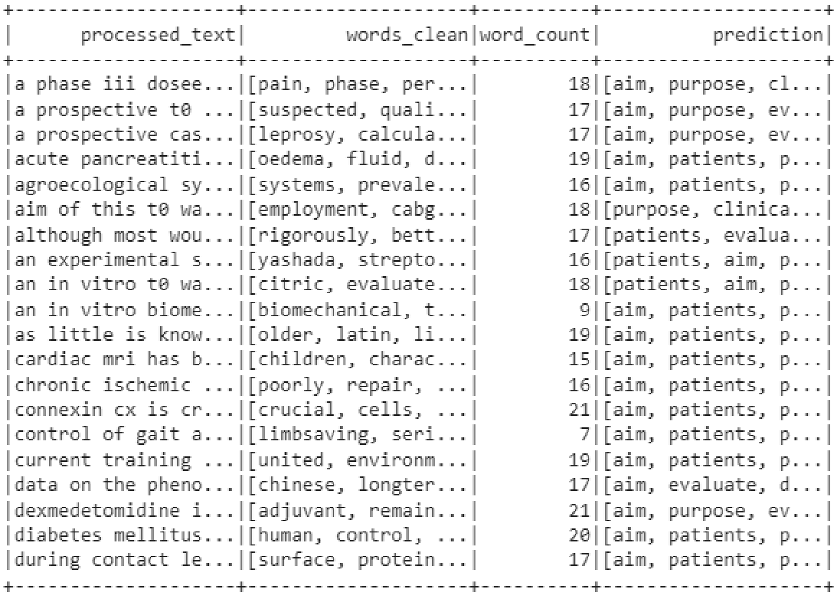
Figure 8: Association Rules

These results show the association rules generated by the FP Growth algorithm, including the antecedent (items that occur before) and the consequent (items that occur after) each rule, along with their corresponding measures of confidence, lift, and support.

Confidence is the proportion of baskets that contain the antecedent item set and also contain the consequent item set, while lift measures the strength of the association between the antecedent and consequent item sets compared to what would be expected by chance. Support measures the frequency of occurrence of the item sets in the dataset.

For example, the rule [evaluate] -> [aim] has a confidence of 0.368, indicating that 36.8% of the baskets containing [evaluate] also contain [aim]. The lift for this rule is 1.612, which means that the association between [evaluate] and [aim] is 1.612 times stronger than what would be expected by chance. The support for this rule is 0.046, indicating that this rule occurs in 4.6

Similarly, other rules can be analyzed based on their support, confidence, and lift metrics to identify meaningful patterns and associations in the dataset.



```
+--------------------+--------------------+----------+--------------------+
|      processed_text|         words_clean|word_count|          prediction|
+--------------------+--------------------+----------+--------------------+
|a phase iii dosee...|[pain, phase, per...|        18|[aim, purpose, cl...|
|a prospective t0 ...|[suspected, quali...|        17|[aim, purpose, ev...|
|a prospective cas...|[leprosy, calcula...|        17|[aim, purpose, ev...|
|acute pancreatiti...|[oedema, fluid, d...|        19|[aim, patients, p...|
|agroecological sy...|[systems, prevale...|        16|[aim, patients, p...|
|aim of this t0 wa...|[employment, cabg...|        18|[purpose, clinica...|
|although most wou...|[rigorously, bett...|        17|[patients, evalua...|
|an experimental s...|[yashada, strepto...|        16|[patients, aim, p...|
|an in vitro t0 wa...|[citric, evaluate...|        18|[patients, aim, p...|
|an in vitro biome...|[biomechanical, t...|         9|[aim, patients, p...|
|as little is know...|[older, latin, li...|        19|[aim, patients, p...|
|cardiac mri has b...|[children, charac...|        15|[aim, patients, p...|
|chronic ischemic ...|[poorly, repair, ...|        16|[aim, patients, p...|
|connexin cx is cr...|[crucial, cells, ...|        21|[aim, patients, p...|
|control of gait a...|[limbsaving, seri...|         7|[aim, patients, p...|
|current training ...|[united, environm...|        19|[aim, patients, p...|
|data on the pheno...|[chinese, longter...|        17|[aim, evaluate, d...|
|dexmedetomidine i...|[adjuvant, remain...|        21|[aim, purpose, ev...|
|diabetes mellitus...|[human, control, ...|        20|[aim, patients, p...|
|during contact le...|[surface, protein...|        17|[aim, patients, p...|
+--------------------+--------------------+----------+--------------------+
```

Figure 9: Prediction

The data in figure 9 appears to be a collection of medical research papers or studies. The "processed text" column contains the text of each paper, while the "words clean" column contains the cleaned words from each paper. The "word count" column indicates the number of words in the cleaned text. The "prediction" column contains predictions made by a machine learning model about the aim or purpose of each study based on the text.

5 CONCLUSION

In conclusion, our market basket analysis using the FP Growth algorithm on the Medal Dataset has provided valuable insights into the relationships and patterns within the dataset. By filtering the data based on word count and focusing on baskets within a certain range, we were able to improve the accuracy and reliability of our analysis. Our results show the frequent item sets and their corresponding frequencies, which can be used to generate association rules and provide insights into the relationships between different items in the dataset.

Through the analysis of key market basket measures such as support, confidence, and lift, we were able to identify meaningful patterns and associations in the data. We also identified the most frequently published text words, which helped to refine our analysis and gain deeper insights into the underlying trends and patterns within the dataset.

Overall, our analysis has demonstrated the usefulness of market basket analysis and the FP Growth algorithm in understanding complex datasets. The insights gained from this analysis can inform future research and decision-making processes, particularly in the medical research field where data analysis plays a critical role in identifying trends and patterns

6 Declaration

I declare that this material, which we 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 our 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

M. Al-Maolegi and B. Arkok. An improved apiori algorithm for association rules. *International Journal on Natural Language Computing (IJNLC)*, 3(1):1–9, 2014.

L. C. Annie and A. Kumar. Market basket analysis for a supermarket based on frequent itemset mining. *IJCSI International Journal of Computer Science Issues,*, 9(5):257–264, 2012.

M. Clarke. How to use the apriori algorithm for market basket analysis. URL `https://practicaldatascience.co.uk/data-science/how-to-use-the-apriori-algorithm-for-market-basket-analysis`.

M. Dunham, Y. Xiao, L. Gruenwald, and Z. Hossain. A survey of association rules. *Researchgate*, pages 1–65, 2001.

I. S. Ganiyu. Market basket analysis in data mining simplified 101. URL `https://hevodata.com/learn/market-basket-analysis-in-data-mining/`.

S. Halim, T. Octavia, and C. Alianto. Designing facility layout of an amusement arcade using market basket analysis. *ScienceDirect*, (161):623–629, 2019.

P. Jirapatsil and N. Phumchusri. Market basket analysis for fresh products location improvement: A case study of e-commerce business warehouse. *MSIE 2022, April 28–30, 2022, Chiang Mai, Thailand*, pages 1–28, 2022.

A. Kadlaskar. A comprehensive guide on market basket analysis. URL `https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/`.

V. Kotu and B. Deshpande. Association analysis. *Science Direct,*, pages 1–22, 2019.

T. Kumbhare and P. S. Chobe. An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies,*, 5(1):927–930, 2014.

Mustakim, D. Herianda, and A. Ilham. Market basket analysis using apriori and fp-growth for analysis consumer expenditure patterns at berkah mart in pekanbaru riau. *Journal of Physics*, pages 1–10, 2018.

B. Patil and L. Khot. A study on market basket analysis using apriori algorithm. *Gogte Institute of Technology Belagavi*, pages 1–5, 2022.

A. Rao and J. Kiran. Market basket analysis for fresh products location improvement: A case study of e-commerce business warehouse. *Int J Syst Assur Eng Manag*, pages 1–6, 2021.