CUKUROVA UNIVERSITY / FACULTY OF ENGINEERING

COMPUTER ENGINEERING

CEN445 – Introduction to Data Visualization

Finding Frequent Patterns Using FP-Growth Algorithm

Asst. Prof. Dr. Mehmet SARIGÜL

Tufan Akbaş - 2019555002

Ömer Dinçer – 2020555402

Introduction

As technology continues to develop and become more widespread, the amount of data stored in digital environments has grown significantly. However, data in its raw form is often useless. Nevertheless, with appropriate data processing methods, such as data analysis, data mining, and data visualization, it can become a valuable asset in every aspect of life.

Data mining is an analytical method that aims to extract meaningful information from large datasets. The process involves collecting and preparing data, which includes techniques such as data cleaning, organizing missing data, and removing unnecessary information. Then, through data exploration, models suitable for the dataset are found. These models represent patterns in the dataset and can include machine learning algorithms. During the evaluation phase, the models' performance is tested. If the evaluation is successful, the models are put into use.

The FP-Growth algorithm is a method for mining frequent item sets in a dataset. To efficiently generate frequent item sets, the algorithm uses a divide-and-conquer approach. There are several algorithms for finding frequent item sets, such as Apriori, but FP-Growth is a more efficient and faster method. The algorithm scans the database only once and creates an FP-tree.

In our dataset, there are approximately 42.000 news with 13 different categories. All the news are in text(.txt) format. Contents of the news include a title and a paragraph in Turkish.

Purpose and Importance of the Study

Analyzing a dataset of 42,000 news articles using the FP-Growth algorithm is significant for extracting valuable insights and identifying frequent patterns. The primary objective of this study is to uncover meaningful relationships and frequent patterns among the news articles. The FP-Growth algorithm, through association rule analysis, can identify words or topics that frequently co-occur in the news, offering potential benefits such as discovering hidden patterns and improving decision-making processes.

This project can help to identify trends and patterns in news content, making it a valuable resource for shaping marketing strategies and optimizing content production. Furthermore, identifying frequent patterns in customer feedback or areas of interest can contribute to the development of customer-centric strategies for businesses. FP-Growth analysis can reveal hidden information within large datasets, which can strengthen business strategies and enhance decision-making processes with additional insights.

Method

On that project, the FP-Growth algorithm is used to find frequent patterns in the 42.000 news dataset. The working principle is realized by creating the FP-Tree, a tree-based structure. In the first step, frequently occurring items in the dataset are identified. Then, the FP-Tree containing these items is created. The FP-Tree represents the frequent patterns in the dataset. When building the FP-Tree, the algorithm compresses the data in such a way that the frequent items are included, and thus efficiently analyzes the data set. Due to this simple structure and efficiency, FP-Growth is used for finding association rules in large data sets[1].

FP-Growth Algorithm

The FP-Growth algorithm is employed in the analysis of a vast dataset comprising 42,000 news articles distributed across 13 categories. This algorithm proves particularly valuable for revealing frequent patterns within the dataset, assisting in the extraction of meaningful insights. The process begins by establishing a frequency table, where the occurrences of individual words or topics within the news articles are computed, adhering to a predetermined support threshold that ensures a minimum frequency requirement.

Following the creation of the frequency table, the algorithm proceeds to prioritize and sort the words or topics based on their frequency. This meticulous sorting ensures that only those words or topics surpassing the predefined support threshold are considered for further analysis.

The FP-Growth algorithm's core functionality lies in the construction of an FP-Tree tailored to the characteristics of the news articles. This tree structure, formed by encoding each article in a manner that exclusively includes the identified frequent words or topics, serves as a representation of the underlying patterns within the dataset. The iterative construction of the FP-Tree involves sequentially adding each news article to the tree.

As the algorithm advances, it systematically sorts the remaining articles based on the frequency of words or topics and updates the FP-Tree accordingly. This iterative process is applied to the entire dataset, resulting in the creation, and merging of new FP-Trees in each iteration.

The culmination of this process involves merging the generated FP-Trees, yielding a consolidated FP-Tree that encapsulates the collective frequent patterns identified across the diverse categories of news articles. Subsequently, association rules are extracted from this FP-Tree, with additional filtering based on support and confidence values derived from the frequency table[2].

In essence, the integrated FP-Growth algorithm, operating within the context of 42,000 news articles categorized into 13 distinct categories, offers a powerful and efficient solution for mining frequent patterns and extracting valuable insights from the extensive dataset.

Experimental Study

First, all 42,000 news items were in 13 categories and in text format (Figure 1).

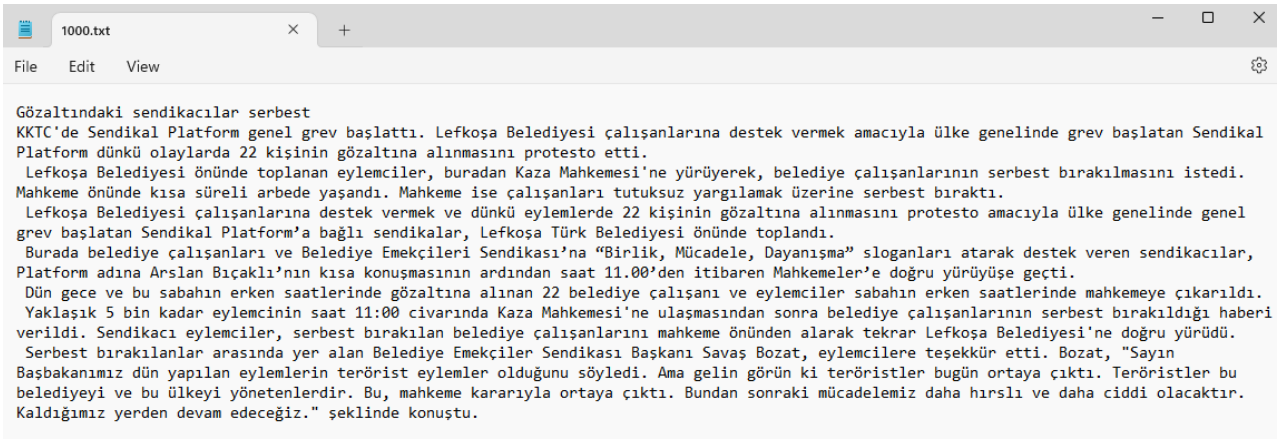

Figure 1

The first thing we did was to convert all the news we had into csv format based on categories (Figure 2).

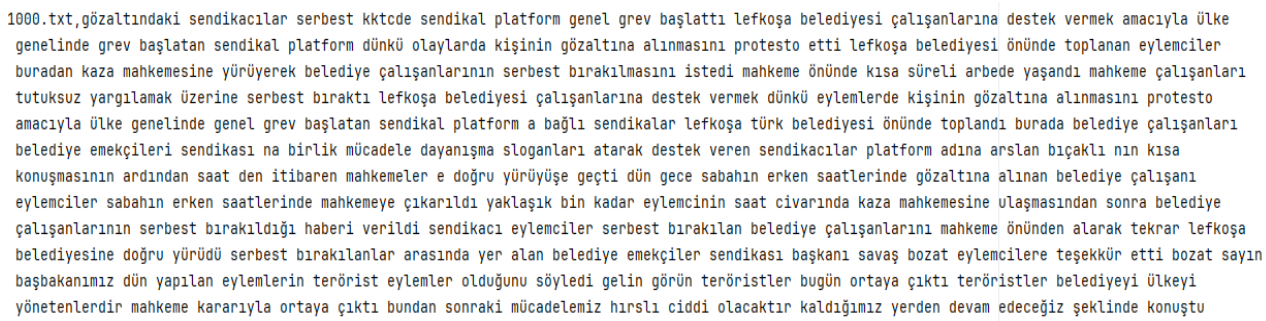

Figure 2

To clean the content of our data and get a better output, we removed single and double letter words, conjunctions, words that have no meaning on their own and numbers. We separated each word with a comma and cleaned up the formatted data (Figure 3).

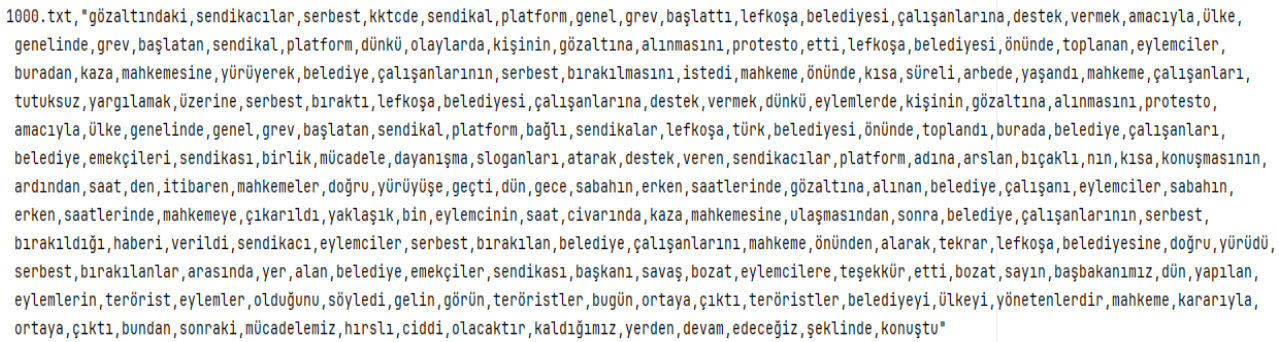

Figure 3

The output after running our algorithm for the whole news without categories. (Figure 4).

```
support,itemsets
0.203182,yer
0.208206,tarafından
0.208349,devam
0.208921,oldu
0.209659,yapılan
0.210778,yeni
0.212683,eden
0.21485,önce
0.221852,ardından
0.228401,ilgili
0.23926,olduğu
0.243689,büyük
0.245332,söyledi
0.252405,etti
0.256644,son
0.257311,ilk
0.267741,olduğunu
0.294127,sonra
0.306225,dedi
0.363641,olan
0.400957,olarak
```

Figure 4

We visualized our output with seaborn and matplotlib libraries (Figure 5).

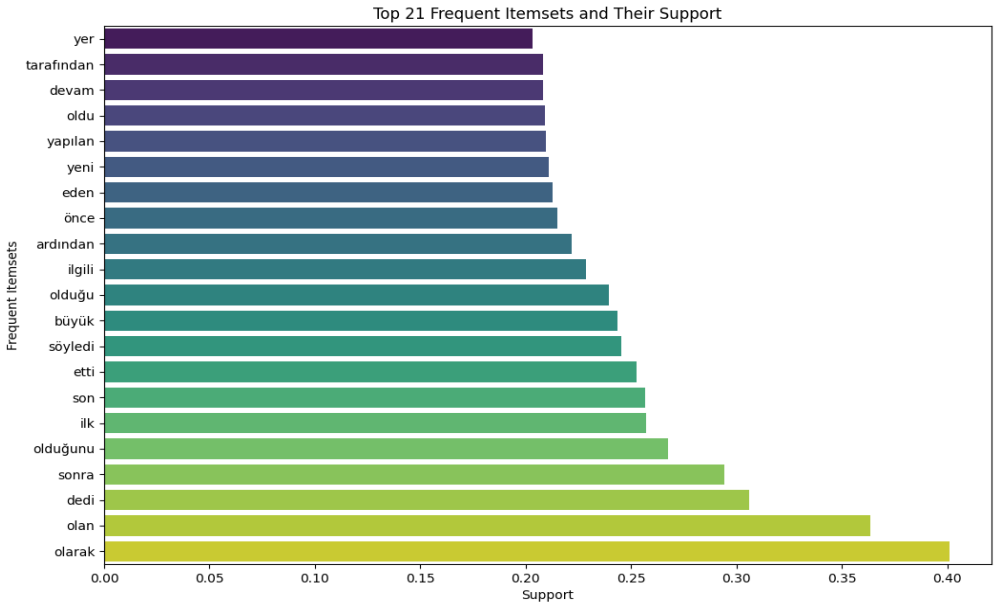

Figure 5

"Health" was the category where the algorithm did the best job of finding frequent patterns (Figure 6).

```
0.2313810556760665, olarak, eden
0.2284887924801157, olduğunu, eden
0.2031814895155459, eden, olan
0.2031814895155459, söyledi, eden
0.21475054229934923, ifade, eden
0.25090383224873464, olduğunu, ifade
0.22270426608821403, olarak, ifade
0.21691973969631237, söyledi, ifade
0.24945770065075923, olarak, fazla
0.23355025307302965, fazla, olan
0.22631959508315258, olduğunu, fazla
0.22270426608821403, olarak, sağlık
0.24150397686189443, olarak, tedavi
0.23427331887201736, tedavi, olan
0.30874909616775126, önemli, olarak
0.28054953000723065, önemli, olan
0.20824295010845986, önemli, olduğu
0.27042660882140274, olduğunu, önemli
0.2198120028922632, önemli, olarak, olan
0.20607375271149675, olduğunu, önemli, olarak
0.2299349240780911, olarak, özellikle
0.2284887924801157, özellikle, olan
0.21258134490238612, ortaya, olarak
0.21402747650036152, son, olarak
```

Figure 6

The table comparing the number of frequent patterns found by the algorithm based on the outputs for each category (Figure 7).

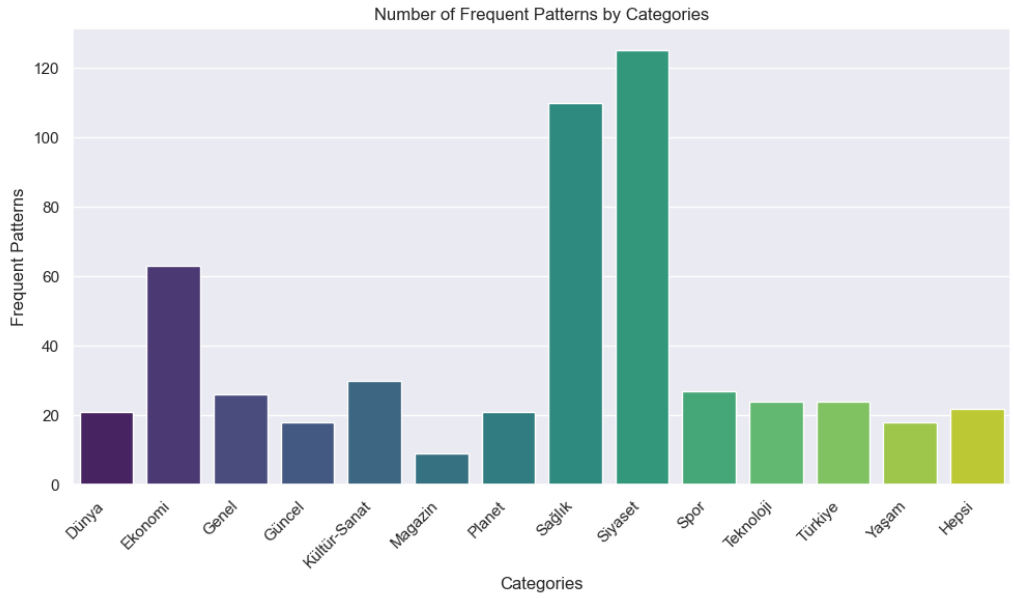

Figure 7

Finally, a table is presented below showing the top 38 most frequently used words in 42,000 news items, along with the number of times each word appears in the news items (Figure 8).

```
Word,Frequency
olarak,34054
olan,25757
sonra,21133
dedi,20959
olduğunu,19728
ilk,19076
son,18061
büyük,17969
bin,17388
var,17209
yüzde,17017
yeni,16841
türkiye,16225
ilgili,15613
söyledi,14956
yıl,14776
olduğu,14763
etti,14678
başkanı,14634
oldu,13834
yer,13537
türk,13446
önce,13165
devam,13132
tarafından,12959
ardından,12890
iyi,12623
milyon,12283
yapılan,12244
önemli,12209
eden,11926
arasında,11163
ifade,10646
konuştu,10486
genel,10407
yaptığı,10378
değil,10104
```

Figure 8

Conclusion

In this project, we analyzed 42,000 news articles in 13 categories using the FP-growth algorithm. By analyzing the words and their frequencies in these news articles, we determined the most frequently used words and the frequency of their co-occurrence for each category.

As a result of this analysis, it was observed that each category has its own unique vocabulary and that these words are used in different combinations. For example, the most frequently used words in the " culture-art" category were "kültür", "film", "ünlü", and "sanat", while the most frequently used words in the "politica" category were "türkiye", "başkan", "parti", and "başbakan".

This analysis can show that the words and their frequencies in news articles play an important role in determining the content and category of the news article. This information can be used in applications such as automatic classification and analysis of news articles.

References

1. https://medium.com/@anilcogalan/fp-growth-algorithm-how-to-analyze-user-behavior-and-outrank-your-competitors-c39af08879db

2. https://www.geeksforgeeks.org/frequent-pattern-growth-algorithm/