**Data Analysis Report: Microsoft Employee Review**

Samantha Paige Minchew

Professor Satvika Marrapu

*Data Analysis and Knowledge Discovery, INFO 4670, Section 001*

Assignment 4

6 November 2023

The thesis of this report is to understand the process of constructing a data analysis model using Rapid Miner. Using textual associations to construct data, Rapid Miner analyzes words that appear together to discover hidden or important relationships.

**DATA METHODS/PRE-PROCESSING:**

Choosing to use the *Microsoft Employee Review* data set from Kaggle, you must first download it to your computer before uploading it to Rapid Miner. After downloading, I opened Rapid Miner, went to extensions, and then downloaded Text Processing. Adding this extension allows the data to be categorized into different text formats. Using this extension, we will be able to filter for tokenization and stemming for this data report. It is important to download this extension for this report since we will need to prepare and analyze texts.

**OBJECTIVES:**

To understand the process of gathering and organizing data through Rapid Miner, and to be able to create visualizations of Microsoft employee reviews. Using Rapid Miner, we will be able to data-mine the information and find connections between attributes in the data set. After transcribing data, we should be able to find the confidence and support percentages. These two main factors dictate whether frequency patterns get translated into association rules. Confidence percent is calculated by attributes flagged as true when the associated attribute is also flagged as true. Support percentage is the number of times the rule occurred divided by the number of observations in the data set.

**SELECT ATTRIBUTES:**

Even though you are supposed to be able to right-click on the data file under the imported data repository, I was not able to do so. I read on the online Rapid Miner Studio documentation that this was a common issue with MacBook Air owners. Unfortunately, that meant that the right-click feature was bugged and there was no easy way to fix that. Working around the right-click feature, I added the Data Editor display panel and then transferred the imported data file to the panel. From here I was hoping the right-click feature would work so I could adjust the polynomial attributes to text for this assignment, but again, it was not working. To work around this, I added an operator that would change nominals to texts, so I could continue gathering the data I needed.

*Figure 1.a*

A screenshot of a computer

Description automatically generated

As seen in Figure 1.a, I added a ‘type’ operator that would change selected nominal attributes and convert them to text. As you can see, I transferred advice to management, company, employee type, job title, link, location, pros, and summary to text. You could choose whichever attribute you wanted, but for the sake of this assignment, I chose all the topics that I wanted to focus on and the ones that had missing values. I checked to see if the attributes were changed from polynomial values to text by running the data and checking the result type under statistics.

**PROCESS DOCUMENTS FROM DATA:**

Processing documents from data generates word vectors from string attributes. This parameter is a powerful tool that is used for text processing and feature extraction from raw text data. It is specifically designed to transform unstructured text data into a structured format for machine learning and data analysis. Figure 2.a shows the tokenized and filtered attributes from the processing documents from the data operator after the complete process.

*Figure 2.a*

A screenshot of a computer

Description automatically generated

Figure 2.a shows the tokenized and filtered attributes from the processing documents from the data operator after the complete process. By double-clicking *Documents from Data* attribute, you are transferred to another page of that attribute’s sub-category. In the Parameters panel, I selected a single filter and chose “cons” for the attribute. Under the Parameters panel, I selected a checkbox for creating a word vector, then chose *Binary Term Occurrences*. Under the “prune method” box, I selected “absolute value,” I kept the prune below absolute at 2, and the prune above absolute at 9999. You’re welcome to change the absolute value to whatever you want to be included in your data set.

**TOKENIZE:**

Tokenization is the process of breaking a stream of text into phrases, words, or other meaningful elements known as tokens. The goal of tokenization is to organize and filter the exploration of words in a sentence.

*Figure2.b*

A screenshot of a computer

Description automatically generated

Inside the operator, I added several textual processing operators. Under the Operators panel, search for “Tokenize,” then drag to the processing canvas. From there you will be able to choose the mode, but I kept it at “non-letters.”

**TRANSFORM CASE:**

*Figure 2.c*

A screenshot of a computer

Description automatically generated

Transform Cases is a Rapid Miner text process attribute that transforms the cases of characters in a document to either upper case or lower case. To add this, search for “Transform Cases” in the Operator panel, then add the operator to the canvas. Transform the cases to lowercase in the Parameters panel.

**FILTER STOP WORDS:**

This operator filters English stop words from a document by removing every token that equals a stop word from the built-in stop word list. Since I selected the English stop word operator, every token should represent a single English word only.

*Figure 2.d*

A screenshot of a computer

Description automatically generated

Under the Parameter panel, search for “Filter Stop Words.” Add in the “Filter Stopwords (English)” operator to the attributes. This attribute does not have parameters to change, so there is nothing else needed to be done in this step.

**FILTER TOKENS:**

Filter Tokens is an operator that filters tokens based on their content. A token is kept in the document, if the token either equals, contains, or does not contain the given value.

*Figure 2.e*

A screenshot of a computer

Description automatically generated

then add in “Filter Tokens (by Length)” to the canvas. Here, I set the minimum characters to 4 and the maximum to 99 characters. This operator filters tokens based on the number of characters they contain, which can be adjusted and filtered in the Parameter panel. Now, we can go back to the original canvas by selecting the “Process” button.

**FP GROWTH:**

The FP-Growth, or “frequent pattern growth” operator is applied to generate frequent item sets based on numeric patterns. This algorithm method is common for pattern mining in data sets. Constructing a frequent pattern tree from the input data gives a representation of the data set that captures frequency and association information of the items in the gathered data.

*Figure 3.a*

A screenshot of a computer

Description automatically generated

Now that we are back on the original canvas page, we can begin to finish our data preparation. In the Operators panel, I typed in “FP” and selected “FP-Growth” to add to the process canvas. In the Parameters panel, I selected 0.05 for the minimum support attribute and kept 0 for maximum items (meaning no limit was set).

**CREATE ASSOCIATION RULES:**

Association rules are created by analyzing data for frequent patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database and confidence of a class value that states how certain the model is that a document belongs to that class.

*Figure 4.a*

A screenshot of a computer

Description automatically generated

In the Operators panel, search for “Create Association” then drag the Create Association Rules operator to the process canvas. From here, connect the fre port from the FP-Growth operator to the ite port on the Create Association Rules operator. Then connect the rul port to the res port on the process canvas. In the Create Association Rules operator Parameters panel, I selected and changed the minimum confidence attribute from 0.3 to 0.2. I was finally able to run my program and check the results.

**RESULTS:**

*Figure 5.a*

A screenshot of a computer

Description automatically generated

Under the Results page, I clicked the FP-Growth sets to check the results on this page. Figure 4.a shows my frequency growth output results from size, support, and items listed. Now click on the “Association Rules” to see the outcome for the processed data (Figure 5.b).

*Figure 5.b*

A screenshot of a computer

Description automatically generated

Figure 5.b is the association rules outcome. Here we can see the premises, conclusion, support, confidence, and lift. The Support bar measures the proportion of documents that contain the item set. From the results, you can see that nearly 10% of all documents contain an association of the terms Information and Users. Whereas the Confidence bar shows the proportion of documents that include the conclusion term given the premise term. The Lift bar measures how many times Confidence is larger than the expected (baseline) Confidence. A lift value that is greater than 1 is desirable.

*Figure 5.c*

A screenshot of a computer

Description automatically generated

To see Association Rules, click the Description tab. Here you can view the Conclusion tab’s associations with confidence percentage. To view the graph’s interactive scatter block chart, click on the Graph tab as seen in figure 5.c. This is interpreted that in the selected 'cons' column the words life and balance are a premise to the word work with a confidence of 99.2%. In the selected column the words life and balance were strongly associated with the word work. With this, we can conclude that employees were not happy with their work-life balance while working at Microsoft.

*Figure 5.d*

A screenshot of a computer

Description automatically generated

Here is a visual scatter-block chart of the results for Microsoft Employee Review. Under the filter tab, you can search for specific conclusions or pair them using lift criteria as seen in the following figure.

*Figure 5.e*

A screenshot of a computer

Description automatically generated

As seen in this figure, only the results for balance, life, and work are shown. You can examine the graph results by sliding up the minimum criterion value, you can see the confidence levels for support and lift also change. This represents a visualization of the association rules.

**CONCLUSION:**

Overall, by examining the data set of Microsoft Employee Review, we can gather that the employees were not happy with the work-life balance. We gathered this information by understanding Rapid Miner to analyze and data mine the given data set. From transcribing nominal values to text values, I was able to sort out data by using text processing tools. Learning how to process documents from data is a valuable tool that allows us to visualize and understand data. Analyzing the Microsoft Employee Review data set allows us to understand the confidence values by constructing a data analysis model. Using textual associations to construct data, Rapid Miner analyzes words that appear together to discover hidden or important relationships, so that we can manipulate the data to form visualized and interactive graphs.