Rashmi Jakkani

INFO 5810.001

Summer 2022

Text Mining

Dataset: Microsoft Employees Reviews (It contains 9000 records)

About the data and source: I used the Microsoft Employee Reviews data which has 9000 records in it.

This dataset is collected from Kaggle. I choose this data to know about the reviews of the Microsoft employees using some data mining techniques. This data has different columns like company name, location, dates the reviews were wrote, job title, summary, pros, cons, etc.

I choose this to perform the Correlation, Association Analysis and Clustering analysis. I have cleaned the dataset while performing the analysis using remove duplicates operator.

The goal of the analysis:

The goal of my analysis is to determine the Correlation Analysis, Association Analysis, Cluster Analysis using rapid miner.

The correlation produces the vector weights between the attributes and gives a correlation matrix. It is a statistical method between variables and correlation analysis is used to compute the linear relation with 2 variables and merge their association.

Association analysis is to determine the relations between the categorical variables and some special values for large datasets. It is used in many applications like in medical field, business field etc.

Clustering analysis in rapid miner is used to combine the objects that are same to each other and dissimilar to the objects that are clusters.

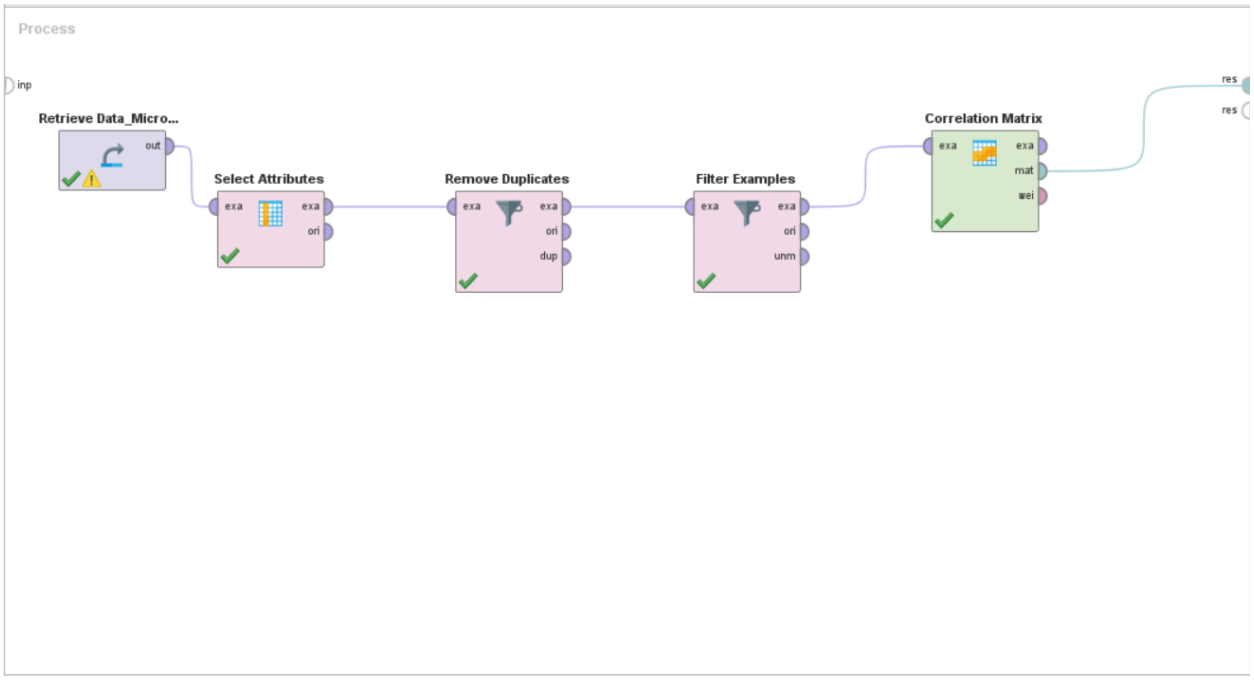
Data Mining Techniques:

Correlation:

Correlation establishes the relationships between each attribute and generates a visual representation of those relationships. Using this correlation method, it is clear that a few attributes are connected. Actually, the correlation between two qualities is implied by the relationship between -1 and 1.

Positive and negative associations can occur in two different ways. Huge values from one set will be associated with high values from another set if the correlation is positive, while few values from one set would be associated with little values from another set. The high values of one set correlate with the low values of another set when a correlation returns a negative result.

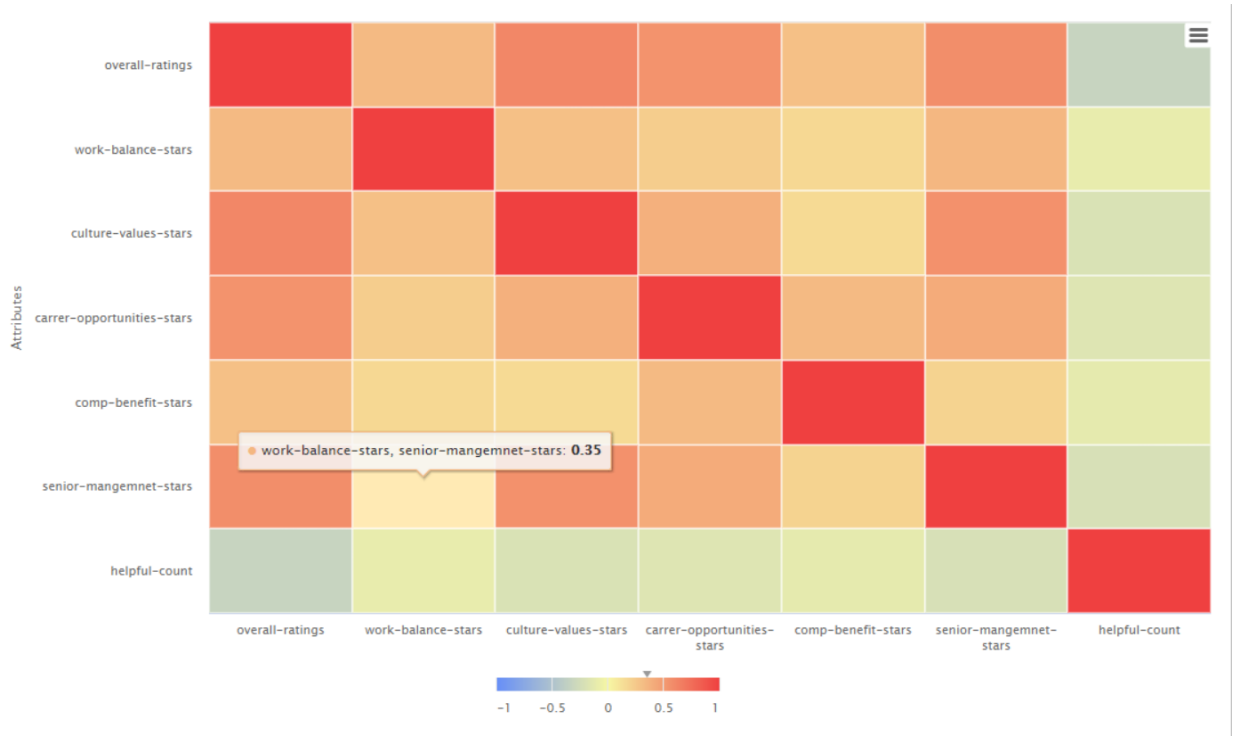
Here is the Correlation Analysis design:



Results:

Attribut...	overall-...	work-b...	culture-...	carrer-...	comp-b...	senior-...	helpful-...
overall-r...	1	0.333	0.610	0.549	0.298	0.576	-0.334
work-bal...	0.333	1	0.302	0.229	0.168	0.350	-0.102
culture-v...	0.610	0.302	1	0.390	0.157	0.557	-0.206
carrer-op...	0.549	0.229	0.390	1	0.331	0.417	-0.177
comp-be...	0.298	0.168	0.157	0.331	1	0.199	-0.124
senior-m...	0.576	0.350	0.557	0.417	0.199	1	-0.223
helpful-c...	-0.334	-0.102	-0.206	-0.177	-0.124	-0.223	1

Visualization:

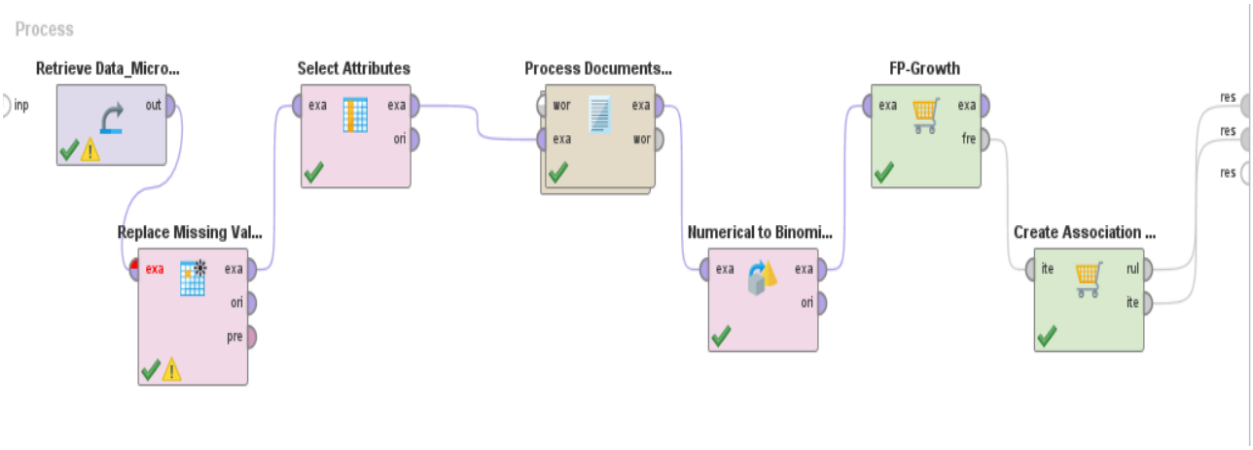


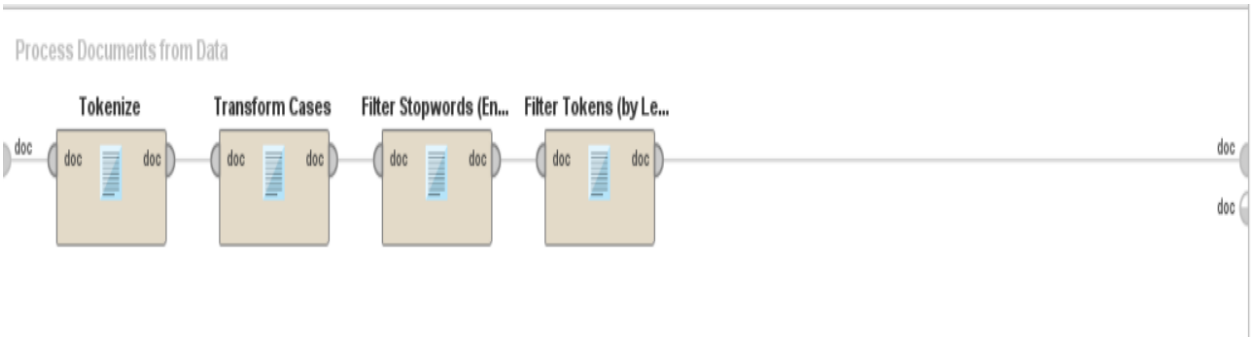
Association Analysis:

A collection of association rules are produced by this operator. They are shown using if or else lines, which show how the things appear in the database, while the other shows the actual objects. The majority of their uses are in bioinformatics and other diverse fields.

The FP-Growth is attached to the create association rules after being dropped on the design phase adjacent to it. When the operator was selected, its parameters were presented on the right panel with the criterion set to confidence and the minimum confidence set to 0.8. Its output rule port was also connected to the result port.

Association Analysis Design:





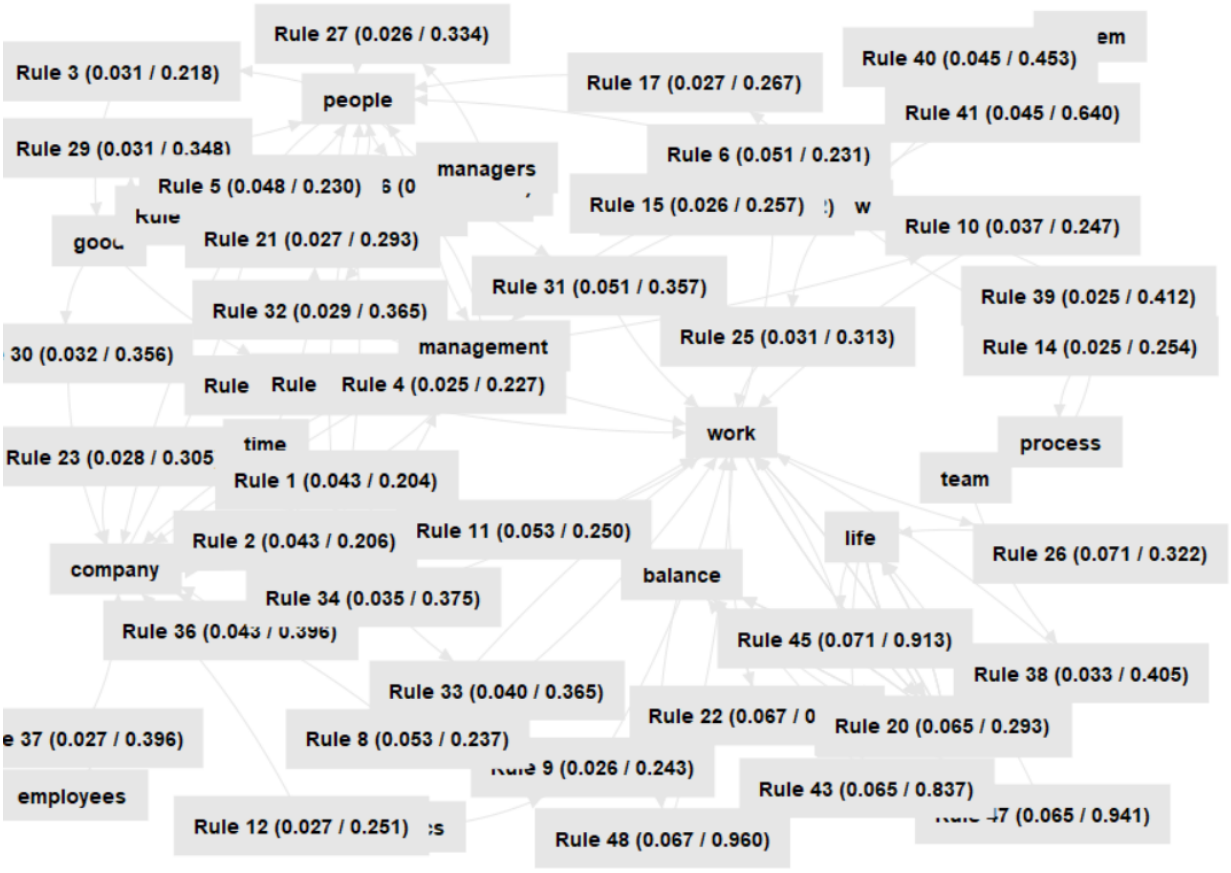
Results:

AssociationRules

```
Association Rules
[company] --> [management] (confidence: 0.204)
[company] --> [microsoft] (confidence: 0.206)
[people] --> [good] (confidence: 0.218)
[microsoft] --> [management] (confidence: 0.227)
[company] --> [people] (confidence: 0.230)
[work] --> [people] (confidence: 0.231)
[people] --> [microsoft] (confidence: 0.235)
[work] --> [company] (confidence: 0.237)
[politics] --> [work] (confidence: 0.243)
[management] --> [work] (confidence: 0.247)
[company] --> [work] (confidence: 0.250)
[politics] --> [company] (confidence: 0.251)
[management] --> [people] (confidence: 0.252)
[review] --> [process] (confidence: 0.254)
[review] --> [management] (confidence: 0.257)
[people] --> [management] (confidence: 0.263)
[review] --> [people] (confidence: 0.267)
[review] --> [company] (confidence: 0.272)
[management] --> [company] (confidence: 0.288)
[work] --> [life, balance] (confidence: 0.293)
[time] --> [people] (confidence: 0.293)
[work] --> [balance] (confidence: 0.301)
[time] --> [company] (confidence: 0.305)
[microsoft] --> [people] (confidence: 0.306)
[review] --> [work] (confidence: 0.313)
[work] --> [life] (confidence: 0.322)
[managers] --> [people] (confidence: 0.334)
[people] --> [company] (confidence: 0.338)
[good] --> [people] (confidence: 0.348)
[good] --> [company] (confidence: 0.356)
[people] --> [work] (confidence: 0.357)
[managers] --> [work] (confidence: 0.365)
```

```
[managers] --> [work] (confidence: 0.365)
[microsoft] --> [work] (confidence: 0.365)
[time] --> [work] (confidence: 0.375)
[good] --> [work] (confidence: 0.383)
[microsoft] --> [company] (confidence: 0.396)
[employees] --> [company] (confidence: 0.396)
[team] --> [work] (confidence: 0.405)
[process] --> [review] (confidence: 0.412)
[review] --> [system] (confidence: 0.453)
[system] --> [review] (confidence: 0.640)
[life] --> [work, balance] (confidence: 0.830)
[life] --> [balance] (confidence: 0.837)
[work, life] --> [balance] (confidence: 0.908)
[life] --> [work] (confidence: 0.913)
[balance] --> [work, life] (confidence: 0.933)
[balance] --> [life] (confidence: 0.941)
[balance] --> [work] (confidence: 0.960)
[work, balance] --> [life] (confidence: 0.972)
[life, balance] --> [work] (confidence: 0.992)
```

Graph:

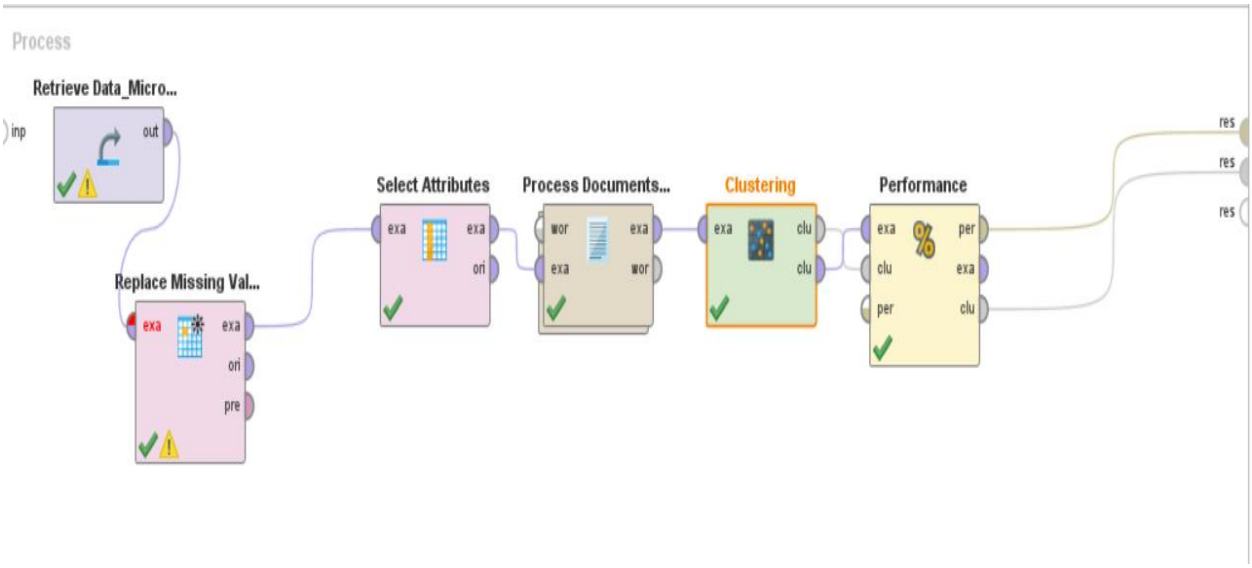


K-Means Clustering:

Unsupervised clustering algorithm K-means divides unlabeled data into a predetermined number of distinctive groupings. K-means also identifies observations that have key similarities and groups them into clusters. The notification makes reference to numerous distinct K-means articles.

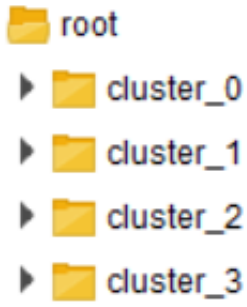
Without any parameters, an unsupervised algorithm will discover the smallest number of clusters. Comparisons are done using unsupervised K-means and other techniques. By looking at the results, the suggested k-means clustering algorithm's advantages become clear.

K-means Clustering Design:

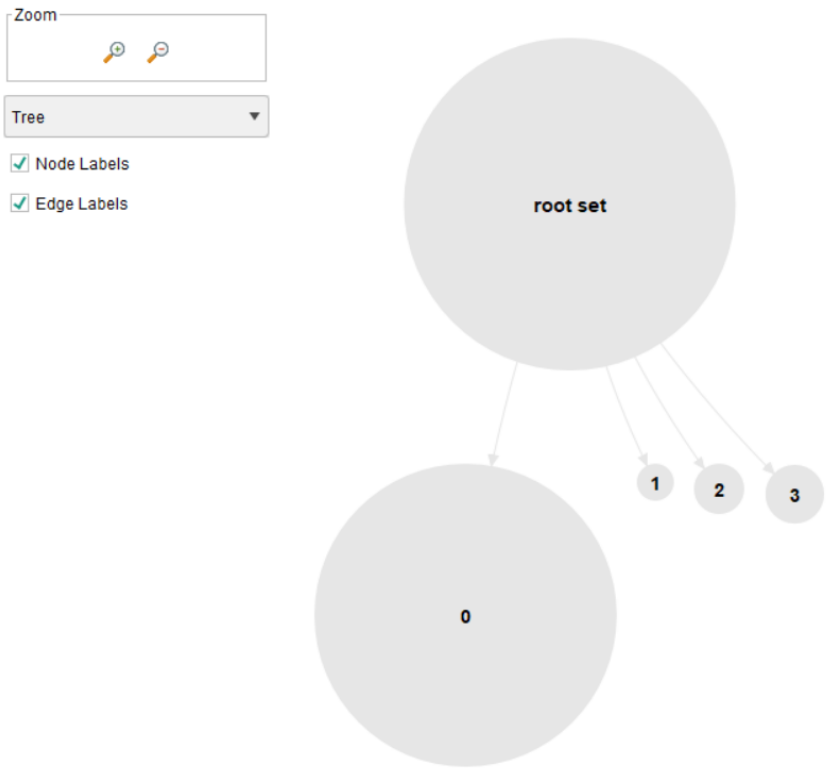


Results:

Folder View:



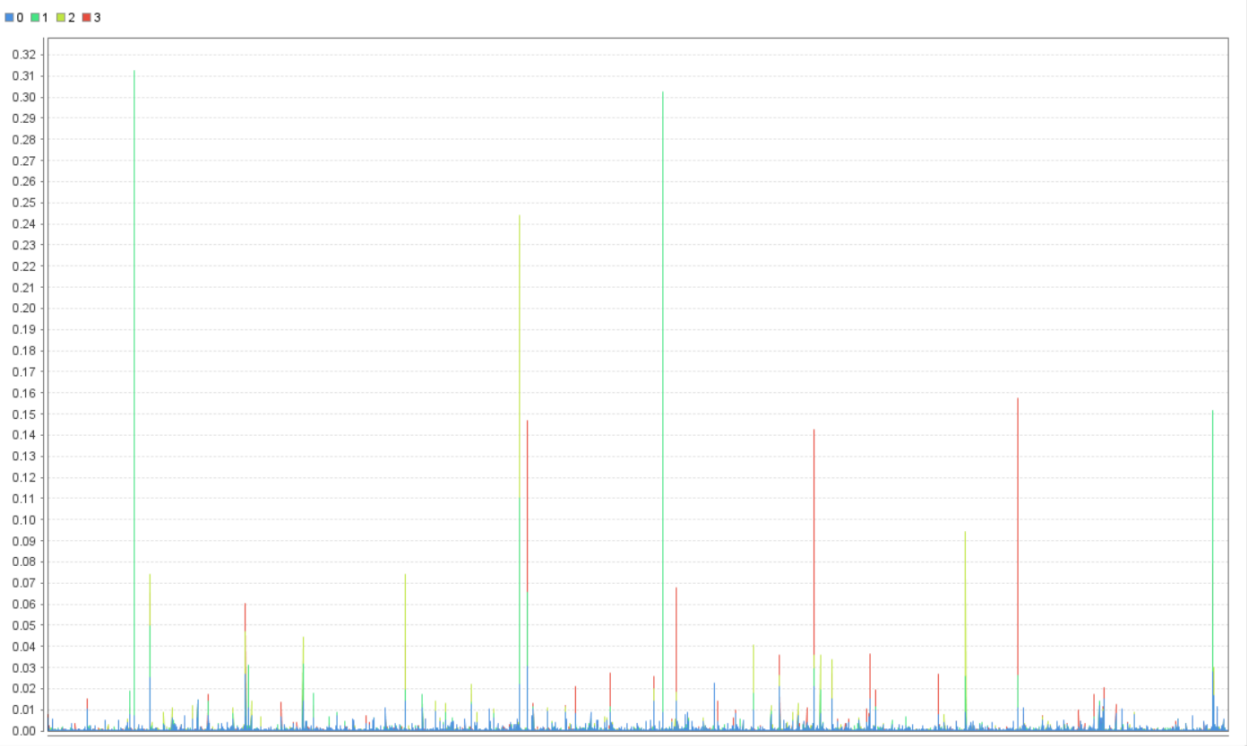
Graph:



Centroid Table:

Attribute	cluster_0 ↓	cluster_1	cluster_2	cluster_3
great	0.030	0.065	0.031	0.147
work	0.028	0.152	0.052	0.079
company	0.027	0.019	0.047	0.060
benefits	0.025	0.050	0.074	0.065
microsoft	0.023	0.002	0.007	0.010
good	0.022	0.110	0.244	0.030
people	0.021	0.030	0.036	0.142
opportunities	0.021	0.020	0.026	0.036
working	0.017	0.012	0.030	0.024
technology	0.015	0.007	0.013	0.020
place	0.015	0.011	0.034	0.027
products	0.015	0.011	0.012	0.036
career	0.015	0.014	0.014	0.012
lots	0.014	0.011	0.018	0.068
culture	0.014	0.032	0.044	0.023
environment	0.014	0.019	0.074	0.041
learn	0.014	0.006	0.020	0.026
opportunity	0.013	0.007	0.010	0.022

Plot:



PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.951
Avg. within centroid distance_cluster_0: -0.983
Avg. within centroid distance_cluster_1: -0.757
Avg. within centroid distance_cluster_2: -0.898
Avg. within centroid distance_cluster_3: -0.897
Davies Bouldin: -6.658
```

Davies Bouldin

```
Davies Bouldin: -6.658
```

Conclusion and Experience:

Found the correlation, clustering and association analysis using rapid miner. Because they all have positive ideals behind them, there is a positive relationship between overall ratings, work-life balance, cultural values, and career prospects. The rules are represented graphically here; it is wise to filter out the layout and compact versions here, though there are more layouts available. This basically states that since these terms frequently appear together in various documents, we should go through and number them. In order to execute this analysis, I first received the data from the data source, replaced any missing values with zeroes using the replace missing values operator and then chose favorable attributes from select attributes in order to perform k-means clustering.

References:

Just a moment. . . (2022). Retrieved from https://www.sciencedirect.com/topics/computer-science/association-analysis

GmbH, R. (2022a). Correlation Matrix - RapidMiner Documentation. Retrieved from https://docs.rapidminer.com/latest/studio/operators/modeling/correlations/correlation_matrix.html

GmbH, R. (2022b). Create Association Rules - RapidMiner Documentation. Retrieved from https://docs.rapidminer.com/latest/studio/operators/modeling/associations/create_association_rules.html