Trend Analysis using Resume Data

Srinidhi Shukla
Instructor- Sahar Behpour
Data Analysis and Knowledge Discovery, INFO 5810, Section number (203)
Term Project
Date: 12/07/2020

1. Introduction

The primary objective of this project is to perform data mining methods on a text corpus to extract knowledge from the dataset. The objective is to extract a meaningful context from the corpus through the implementation of Association analysis and k-means Cluster analysis on the dataset. To accomplish this task, the text given in the fields related to work experience, title and skills are used.

1.1. Data

- About the dataset

  The dataset used for this project is an Excel file, 'resume.xlsx'. It is a structured collection of resumes of job seekers collected and uploaded to Kaggle by user, Avani Siddhapura on 2020-04-11. It consists details likes Resume_title, City, State, Description and Skills. Each row in the dataset denotes an individual's resume data. This original dataset consists of 14800 resumes data.
- Attributes

  The following is the list of attributes used in this dataset
  1. Resume_title- Qualitative (Unique)
  2. City- Qualitative (Specified location)
  3. State- Qualitative (Specified location)
  4. Description- Qualitative (Unique)
  5. Skills- Qualitative (Unique)
- Reason to choose this dataset

  This dataset is a collection of resumes of people who are applying for the jobs in the field of Computer Science. It provides perfect information about the trending industry skills such as software, programming tools, technology or programming languages. I found it interesting as we can determine what this generation's individuals are focusing on to get their dream job.

1.2. Background Research

This dataset has been used by different users for different purposes. Based on the research conducted in these papers- Proposed System for Resume Analytics (Amala Deshpande et al., 2016) and Frequent Item Based Text Clustering: Big Data Analytics, I found that the dataset is quite relevant for the anlysis.

1.3. Objectives

To extract meaningful context from the dataset using

i) Textual Data Analysis using K-means Clustering

ii) Association analysis

2. Methods

2.1. Tools

The tools used in this assignment for the analysis are Microsoft Excel and RapidMiner.

2.2. Data pre-processing

- The original dataset consisted of the following columns- Resume_title, City, State, Description, work_experiences, Educations, Skills, Links, Certificates, Additional Information. Out of these the less relevant columns which are- work_experiences, Educations, Links, Certificates, Additional Information, were dropped using Microsoft Excel.
- Filtered the data by removing the missing values using 'Filter' option in Microsoft Excel.
- Removed any other missing values using the 'Replace Missing Values' operator in RapidMiner.
- Used the Text Pre Processing to change the datatype of the text columns to 'Text'.

2.3. Data Analysis

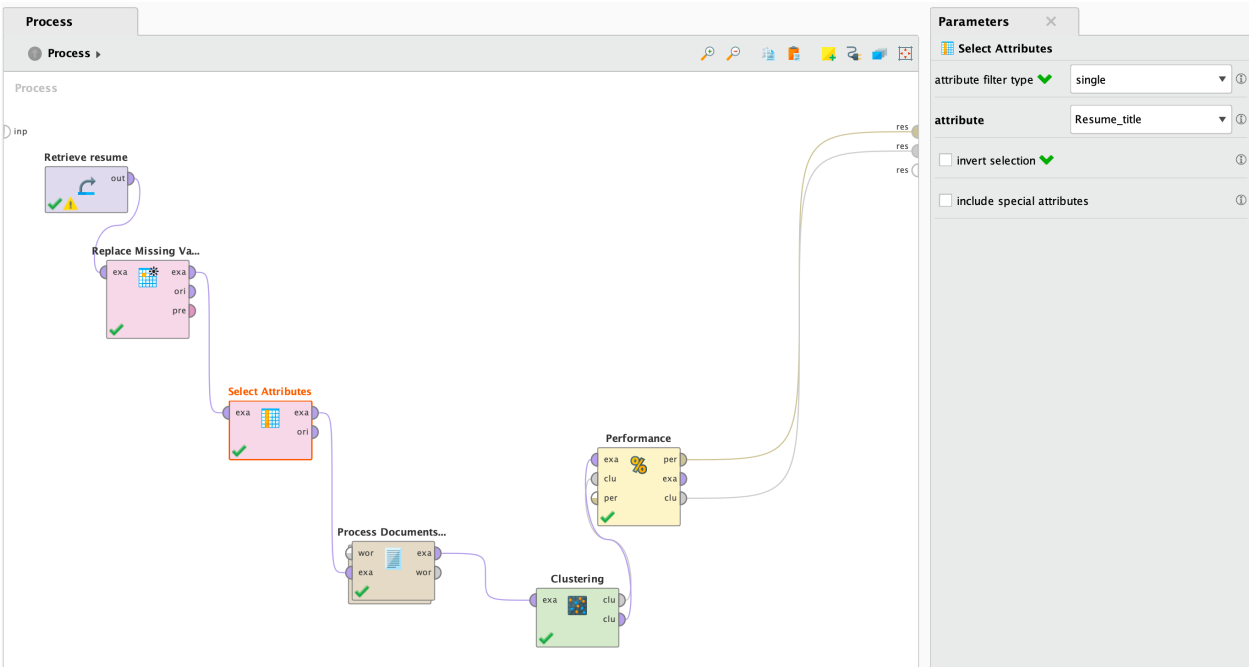
The knowledge extraction in this project has been performed using the following methods and operators

i) Method- K-means Clustering

Operators-

- *Select Attributes:*

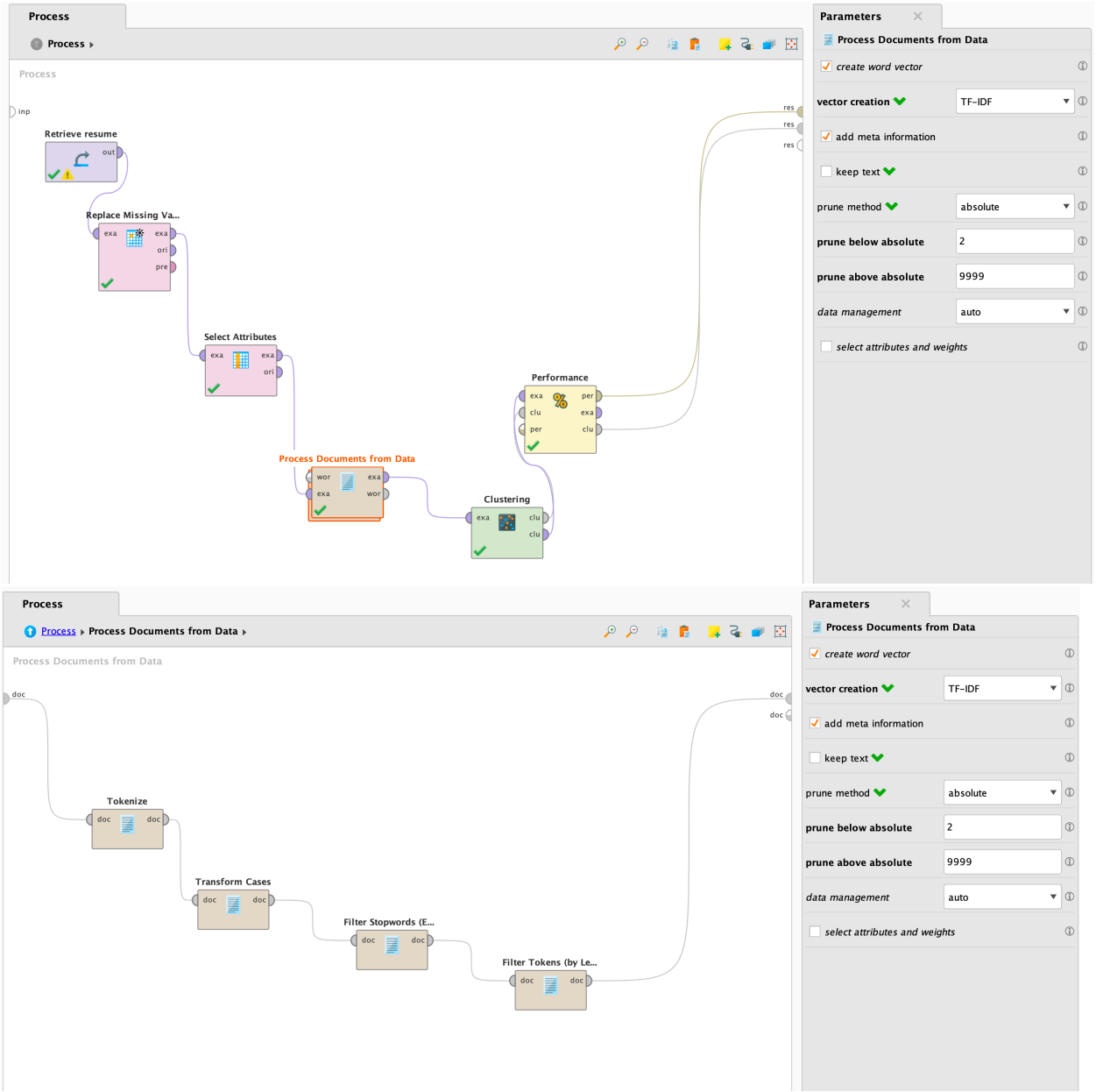
This Operator selects a subset of dataset attributes and deletes all other attributes. The subset is defined by attribute parameters. We filter the data in the column 'Resume_title' for our project.



- *Process Documents from Data:*

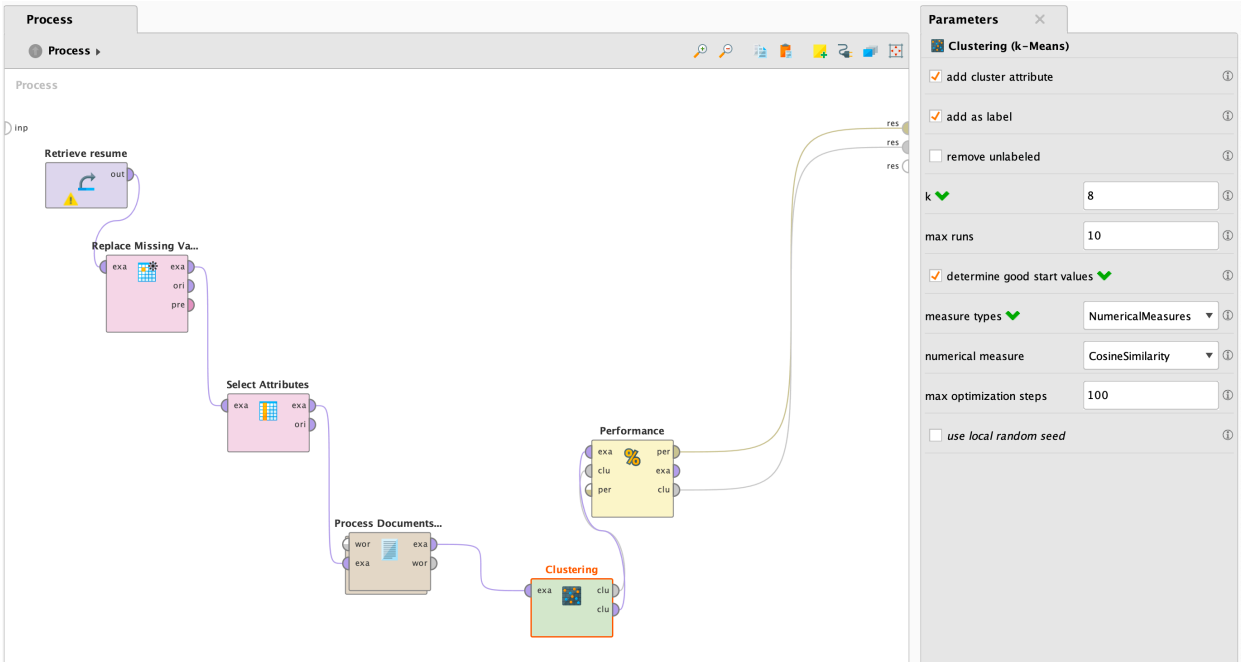
This operator concentrates more on the words to be analyzed in the text. In word vector generation, Data Operator Process Documents is used. The check box is selected to construct a word vector, and

the Binary Term occurrences is selected. 'Keep text' is checked and absolute is selected for the prune method with values '2' and '9999' below and above.



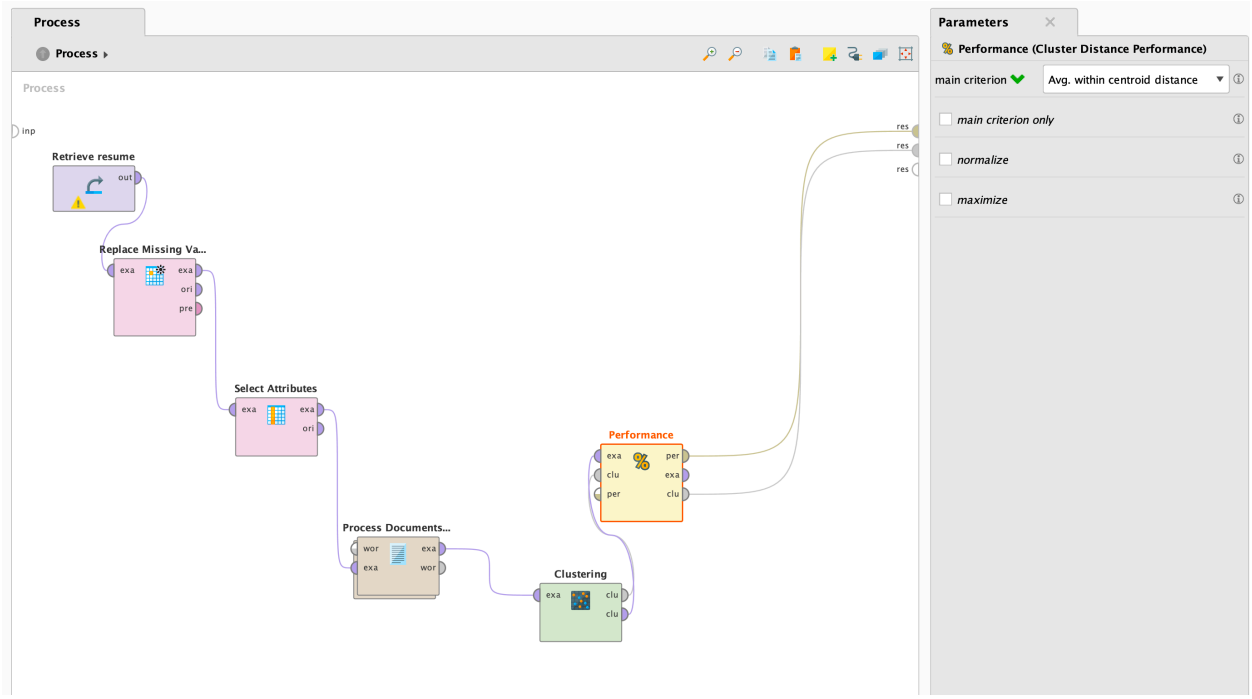
- *K-Means Clustering:*

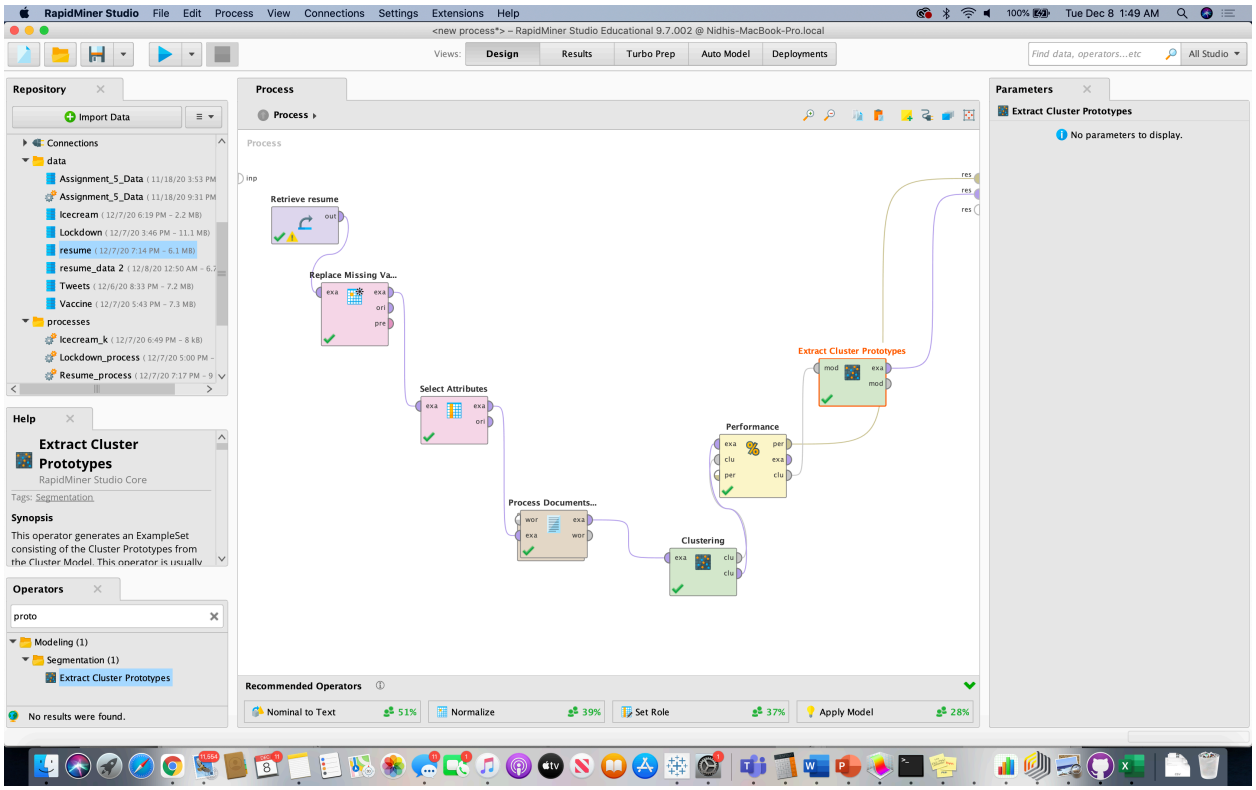
The position of the center in the n-dimensional space of the n-attributes of the dataset is determined by the cluster in the k-means algorithm. This position is known as the centroid. This may be but does not have to be the position of the dataset example. The specifics are broken down into categories. The number of clusters is known as K and indicated by the customer. In this project, we give the K- value as 8, indicating that examples are split into 8 groups. The cluster model that stores the center of clusters and cluster sets is the outputs provided by this operator.



- *Cluster Distance Performance:*

This operator is used for performance evaluation of centroid-based clustering techniques in this project. This operator offers a list of values based on cluster centroids of performance criteria and is provided with the cluster model and clustered sets as inputs in order to test its output on the basis of cluster centroids. Core knowledge is used in every cluster Form cluster centroid. The average cluster distance within and the Davies-Bouldin index help efficiency metrics in the cluster size operator. The efficiency and cluster configuration are the operator's output ports that are attached to the respective ports.



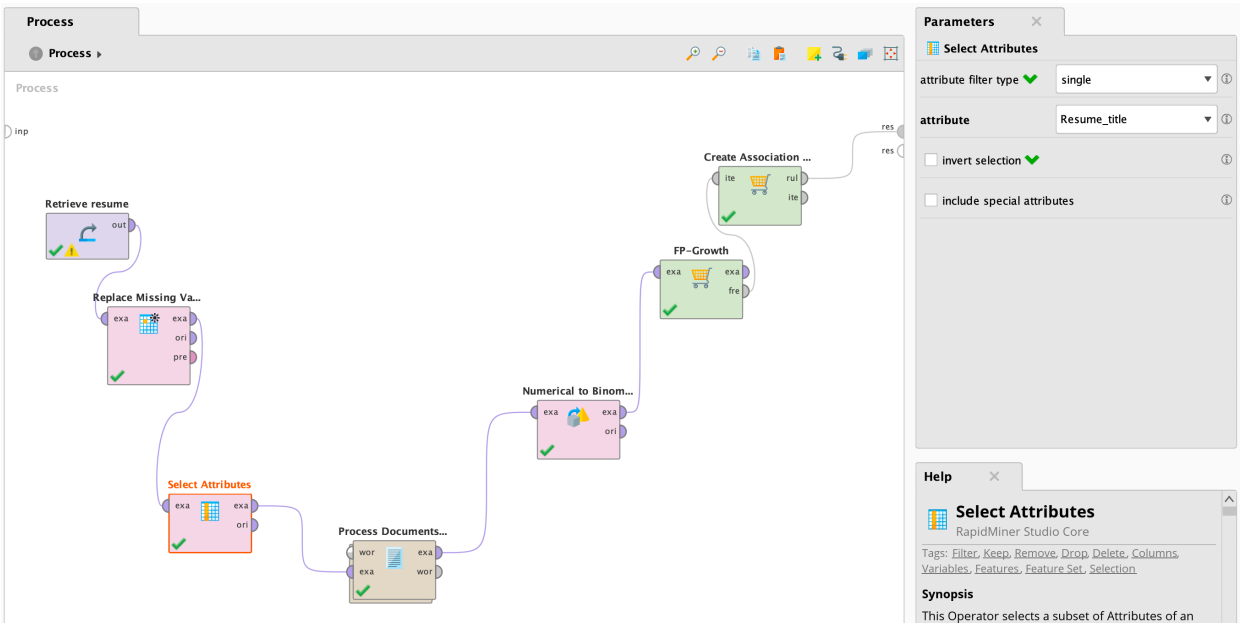


ii) Method- Association analysis

Operators-

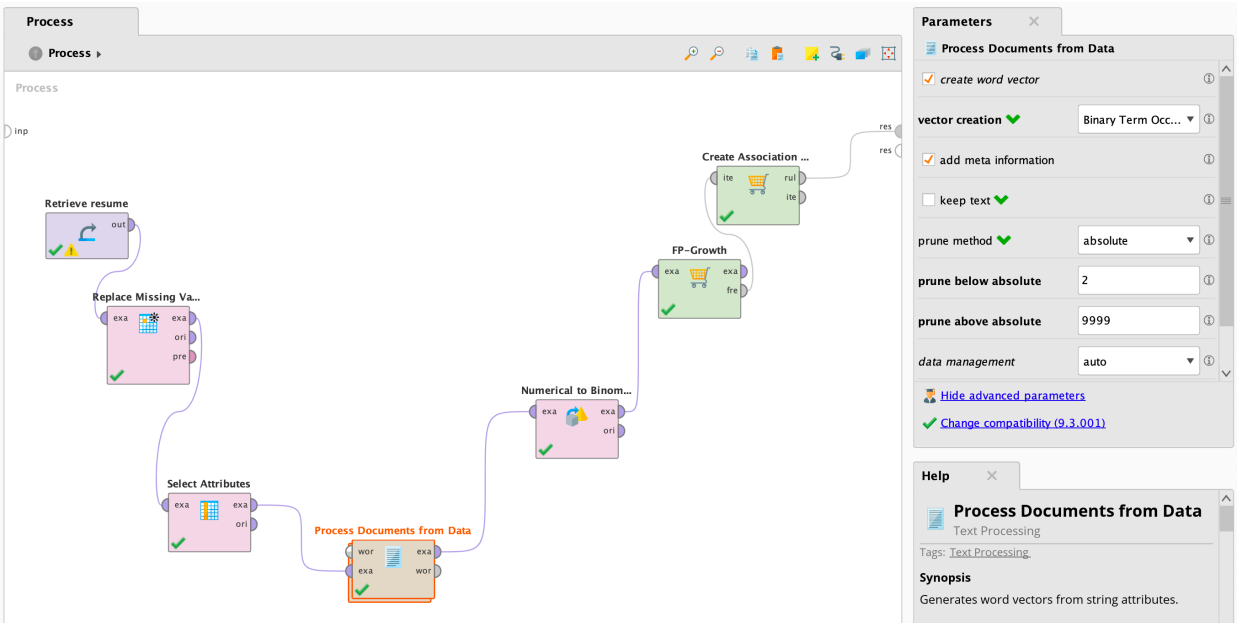
- *Select Attributes:*

  The operator selects a subset of the data set's attributes and deletes other attributes. To make attribute selection simple, the operator offers various filter types. We filter the data in the column 'Resume_title' for our project.



- *Process Documents from Data:*

The word vector from String Attributes is given by this operator. It is used to alter word vectors from string attributes in this assignment.



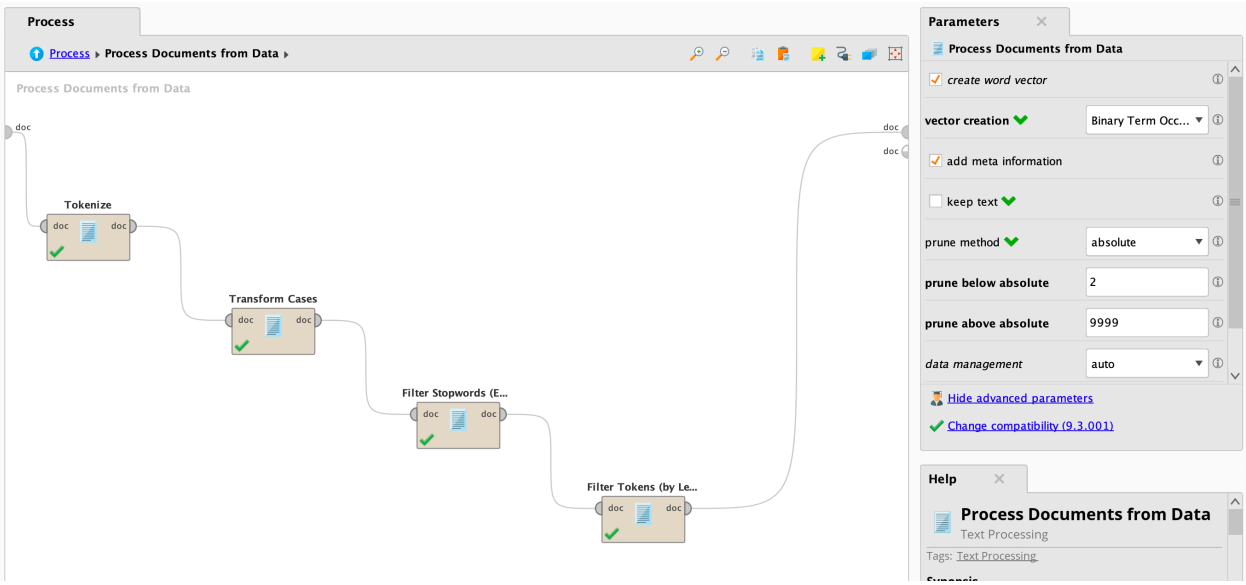
- *Tokenize:*

It tokenizes the documents and splits the document into token sequences. There are many ways of splitting points. In this project use all non-letter characters, each token composed of a single word to make the process quick and flexible, it is used to break large sentences into single words or tokens.

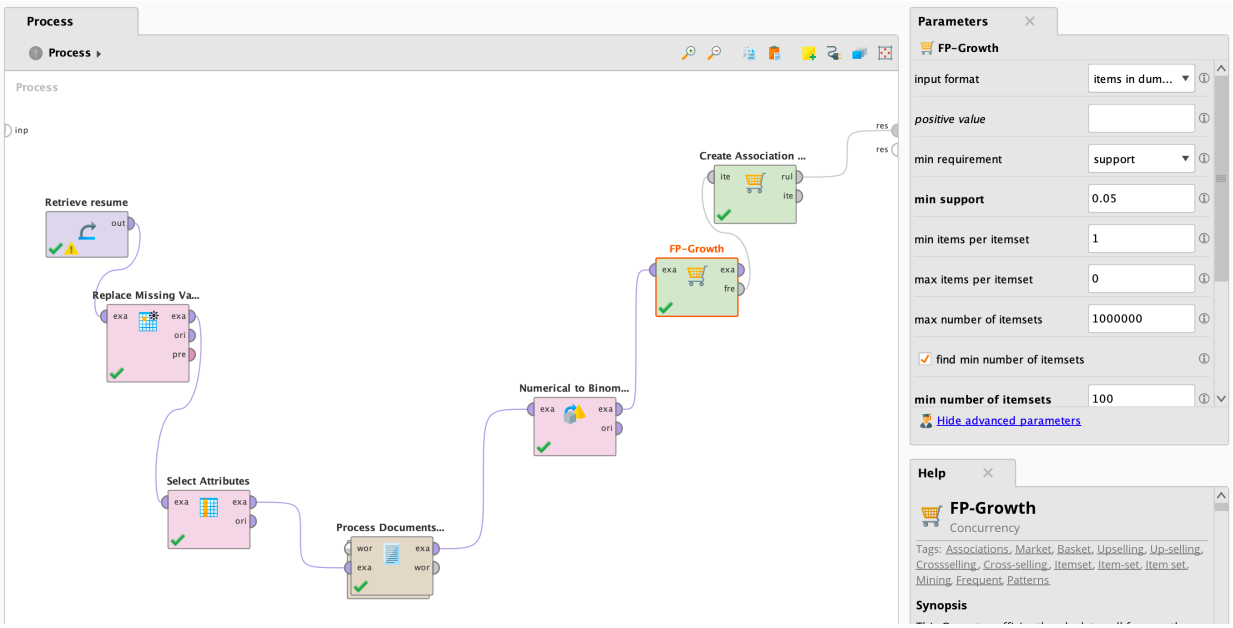
- *Transform Case:*
  This operator converts all the letters in the document into lower- or upper-case letters. And the 'transform to' parameter used shows whether the parameter is transformed to the lower case or the upper case. This operator is used to transform the data to lower case in this project.

- *Filter Stopwords:*
  This removes from the document the English stopwords. It operates perfectly when each token can only represent the single English word and before doing so, we apply the tokenize operator, to tokenize the document, then we can get a document representing a single word for each token.

- *Filter Tokens:*
  This operator filters the tokens according to their count, meaning the number of characters they contain. It is used to count the number of characters in this project.

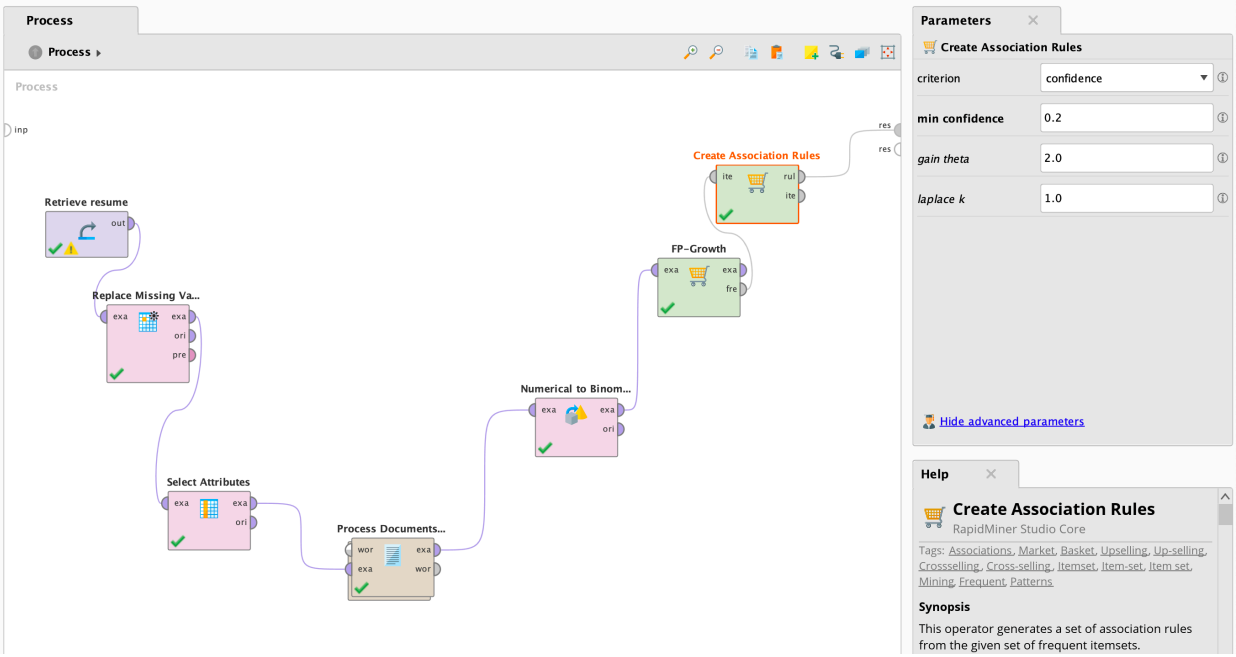


- *FP-Growth:*
  This operator uses the FP tree data structure specifically to measure all of the frequently occurring item sets in the data set. Basically, FP tree data structure can effectively construct, by sacrificing the data too much, in many instances even the massive datasets can fit into the main memory.



- *Create Association Rules:*

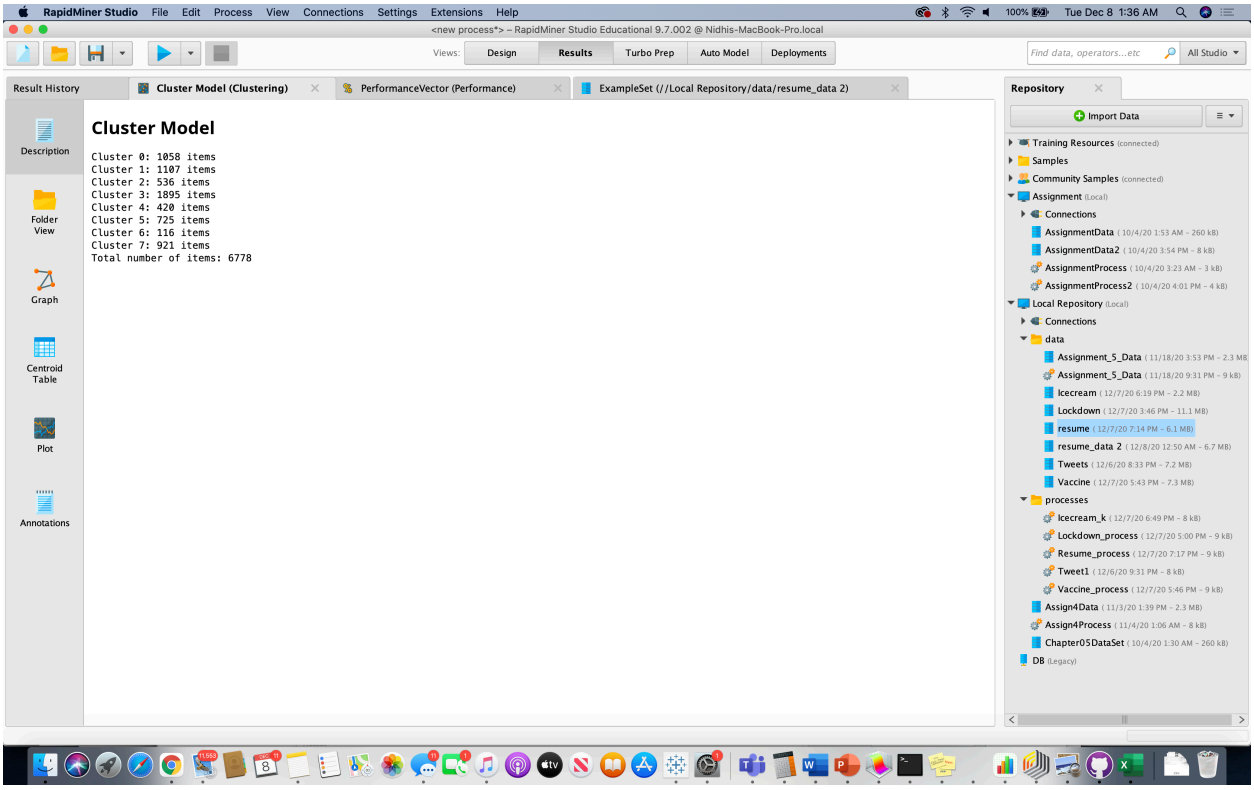
From the set of frequent item sets from the dataset, this operator produces a set of association rules. And it helps to uncover the link between seemingly unrelated details. By analyzing data from a frequent if / then pattern, the association rules are generated using the support and confidence criteria to classify the most relevant relationships.

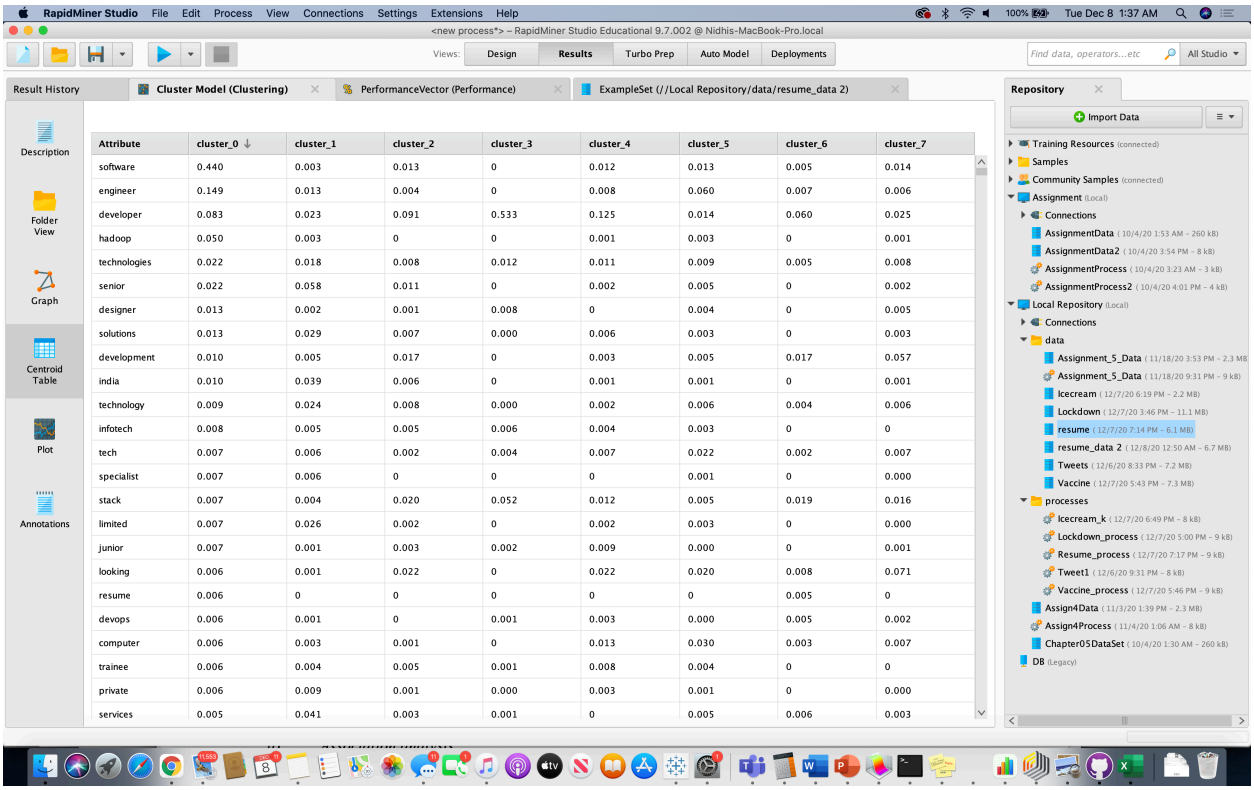


3. Results

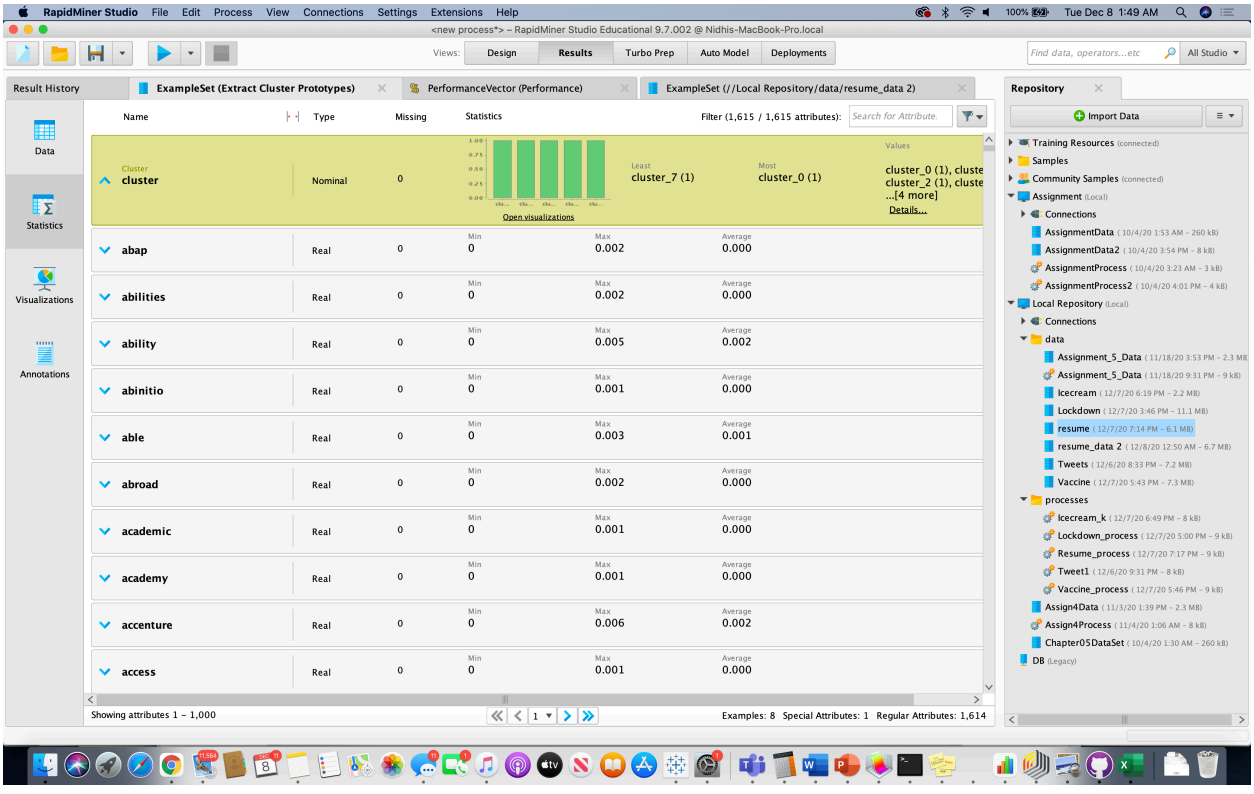
i) Textual Data Analysis using K-means Clustering

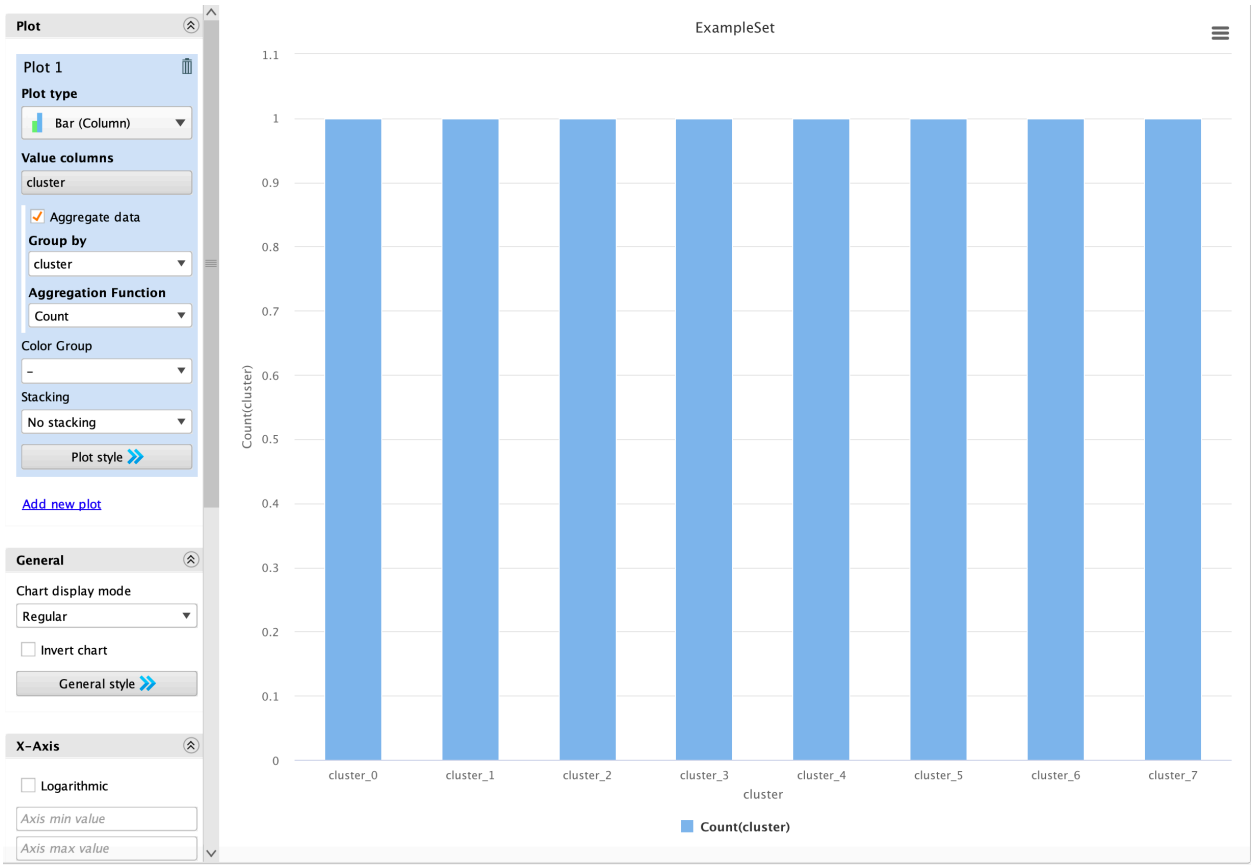
- *Centroid Table Analysis*





Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7
software	0.440	0.003	0.013	0	0.012	0.013	0.005	0.014
engineer	0.149	0.013	0.004	0	0.008	0.060	0.007	0.006
developer	0.083	0.023	0.091	0.533	0.125	0.014	0.060	0.025
hadoop	0.050	0.003	0	0	0.001	0.003	0	0.001
technologies	0.022	0.018	0.008	0.012	0.011	0.009	0.005	0.008
senior	0.022	0.058	0.011	0	0.002	0.005	0	0.002
designer	0.013	0.002	0.001	0.008	0	0.004	0	0.005
solutions	0.013	0.029	0.007	0.000	0.006	0.003	0	0.003
development	0.010	0.005	0.017	0	0.003	0.005	0.017	0.057
india	0.010	0.039	0.006	0	0.001	0.001	0	0.001
technology	0.009	0.024	0.008	0.000	0.002	0.006	0.004	0.006
infotech	0.008	0.005	0.005	0.006	0.004	0.003	0	0
tech	0.007	0.006	0.002	0.004	0.007	0.022	0.002	0.007
specialist	0.007	0.006	0	0	0	0.001	0	0.000
stack	0.007	0.004	0.020	0.052	0.012	0.005	0.019	0.016
limited	0.007	0.026	0.002	0	0.002	0.003	0	0.000
junior	0.007	0.001	0.003	0.002	0.009	0.000	0	0.001
looking	0.006	0.001	0.022	0	0.022	0.020	0.008	0.071
resume	0.006	0	0	0	0	0	0.005	0
devops	0.006	0.001	0	0.001	0.003	0.000	0.005	0.002
computer	0.006	0.003	0.001	0	0.013	0.030	0.003	0.007
trainee	0.006	0.004	0.005	0.001	0.008	0.004	0	0
private	0.006	0.009	0.001	0.000	0.003	0.001	0	0.000
services	0.005	0.041	0.003	0.001	0	0.005	0.006	0.003





- cluster_0: Based on the below attributes which are sorted according to their weights in the particular cluster. The story can be written as "Most Data Science freshers are looking for the analyst, scientist and engineer roles in the field of Machine learning and Deep learning."

Attribute	cluster_0 ↓
data	0.149
learning	0.126
machine	0.107
scientist	0.095
fresher	0.057
looking	0.054
science	0.053
engineer	0.043
analyst	0.038
deep	0.033

- cluster_1: Based on the below attributes which are sorted according to their weights in the particular cluster. The story can be written as “Computer software freshers who are looking for roles in Django full stack technologies are good Python developers.”

Attribute	cluster_0	cluster_1 ↓
python	0.022	0.621
django	0.002	0.130
developer	0.015	0.114
looking	0.054	0.019
stack	0.004	0.015
full	0.005	0.015
fresher	0.057	0.013
software	0.009	0.010
computer	0.021	0.010
technologies	0.006	0.009

- cluster_2: Based on the below attributes which are sorted according to their weights in the particular cluster. The story can be written as “Freshers looking for full stack developer roles have some years of experience in java development and spring framework.”

Attribute	cluster_0	cluster_1	cluster_2 ↓
java	0.002	0	0.640
developer	0.015	0.114	0.102
fresher	0.057	0.013	0.056
full	0.005	0.015	0.027
looking	0.054	0.019	0.026
stack	0.004	0.015	0.023
experience	0.015	0.007	0.023
years	0.010	0.003	0.022
development	0.007	0.005	0.020
spring	0	0	0.020

- cluster_3: Based on the below attributes which are sorted according to their weights in the particular cluster. The story can be written as “Individuals looking for the project management roles are mostly Architects, Technical Leads, Managers, Accountants and Account Assistants”

Attribute	cluster_0	cluster_1	cluster_2	cluster_3 ↓
manager	0.002	0	0	0.100
lead	0.002	0	0.002	0.077
auditor	0	0	0	0.056
technical	0.000	0.001	0.002	0.051
project	0.001	0.001	0.001	0.050
management	0.002	0.001	0	0.043
architect	0.002	0.002	0	0.041
accountant	0	0	0	0.035
assistant	0	0	0	0.033
accounts	0	0	0	0.031

- cluster_4: Based on the below attributes which are sorted according to their weights in the particular cluster. The story can be written as "Senior developers looking for application and solution development roles have some years of experience in HTML and Angular"

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4 ↓
experience	0.015	0.007	0.023	0.003	0.112
years	0.010	0.003	0.022	0.002	0.089
angular	0	0.001	0.006	0	0.050
development	0.007	0.005	0.020	0.010	0.047
application	0.002	0.003	0.010	0.001	0.046
senior	0.004	0.002	0.013	0.030	0.036
html	0.000	0.002	0.007	0.001	0.034
developer	0.015	0.114	0.102	0.008	0.030
year	0.002	0.005	0.008	0.002	0.030
solutions	0.002	0.005	0.009	0.008	0.026

- cluster_5: Based on the below attributes which are sorted according to their weights in the particular cluster. The story can be written as "Android developers looking for Frontend roles are Full stack and Mean stack developers with experience in technologies such as Laravel and Wordpress."

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5 ↓
developer	0.015	0.114	0.102	0.008	0.030	0.544
android	0.003	0.001	0.005	0.001	0.018	0.059
stack	0.004	0.015	0.023	0.001	0.013	0.054
full	0.005	0.015	0.027	0.002	0.009	0.045
front	0.002	0.000	0	0	0.010	0.020
frontend	0.003	0.001	0	0.001	0.003	0.020
wordpress	0.002	0.002	0	0	0.011	0.018
laravel	0	0	0	0.001	0.015	0.015
mean	0.002	0	0	0	0.004	0.013
technologies	0.006	0.009	0.010	0.011	0.013	0.013

- cluster_6: Based on the below attributes which are sorted according to their weights in the particular cluster. The story can be written as "Most developers looking for role in Tata Consultancy Services Limited in India are associates and consultants working in Oracle Labs"

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6 ↓
services	0.001	0.000	0.004	0.006	0.002	0	0.117
india	0.001	0.001	0.007	0.009	0.001	0	0.101
tata	0.000	0	0.002	0.000	0	0	0.086
consultancy	0.001	0	0	0	0	0	0.086
associate	0.004	0.001	0	0.003	0.000	0	0.071
limited	0.002	0.002	0.003	0.004	0.001	0	0.068
oracle	0.001	0.002	0.002	0.000	0.001	0	0.057
consultant	0.005	0.005	0	0.007	0.002	0	0.046
labs	0.001	0	0	0.001	0	0	0.043
developer	0.015	0.114	0.102	0.008	0.030	0.544	0.034

- cluster_7: Based on the below attributes which are sorted according to their weights in the particular cluster. The story can be written as "Software engineers who are senior developers at Infotech Solutions are looking for development roles in technologies such as Hadoop."

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7 ↓
software	0.009	0.010	0.015	0.002	0.013	0	0.017	0.496
engineer	0.043	0.006	0.005	0.014	0.005	0	0.013	0.171
developer	0.015	0.114	0.102	0.008	0.030	0.544	0.034	0.094
hadoop	0.002	0.001	0	0.001	0.001	0	0.006	0.058
senior	0.004	0.002	0.013	0.030	0.036	0.000	0.020	0.028
technologies	0.006	0.009	0.010	0.011	0.013	0.013	0.021	0.024
solutions	0.002	0.005	0.009	0.008	0.026	0.000	0.009	0.015
technology	0.004	0.002	0.010	0.009	0.017	0.001	0.021	0.011
development	0.007	0.005	0.020	0.010	0.047	0	0.005	0.010
infotech	0.002	0.003	0.006	0.003	0.002	0.006	0.005	0.009

- Performance vector:

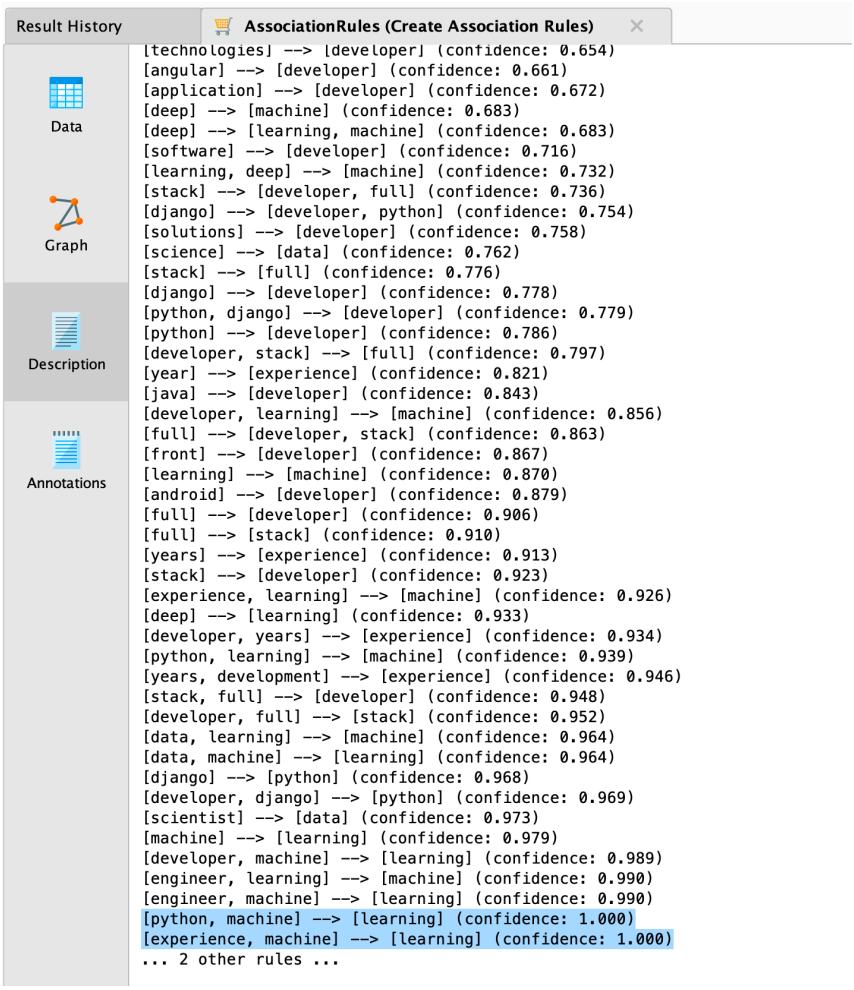
The performance of the clustering is as shown below

PerformanceVector

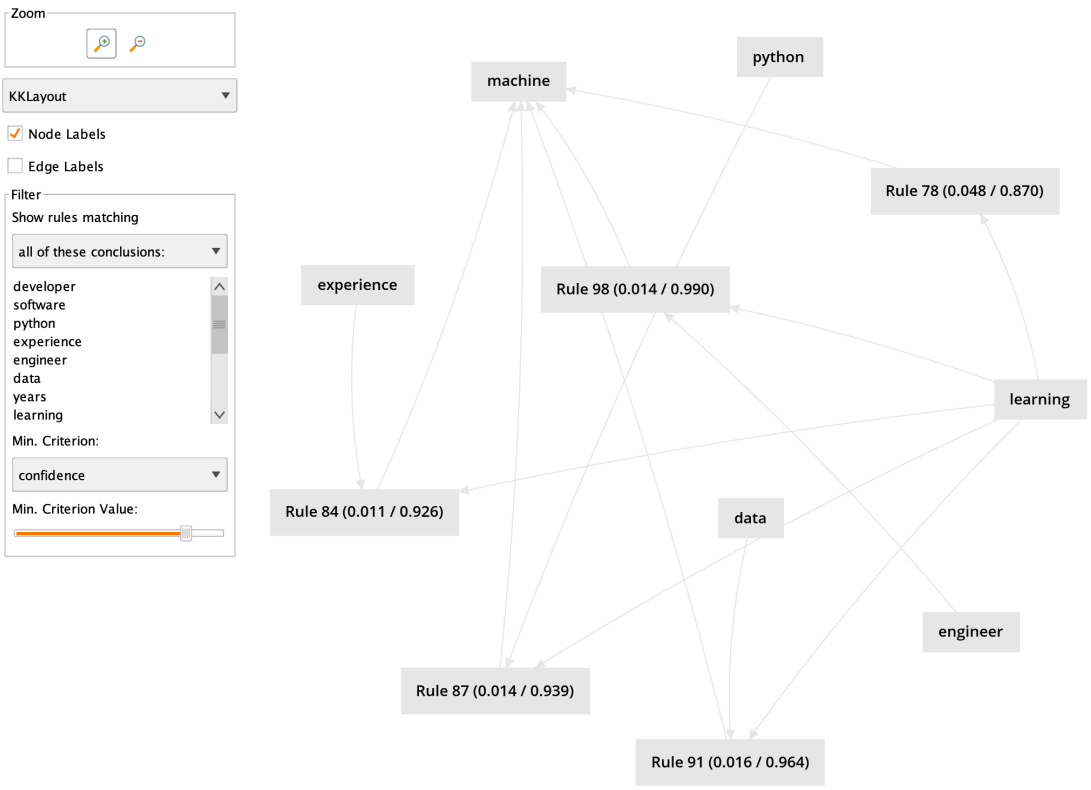
```
PerformanceVector:
Avg. within centroid distance: -0.783
Avg. within centroid distance_cluster_0: -0.877
Avg. within centroid distance_cluster_1: -0.581
Avg. within centroid distance_cluster_2: -0.570
Avg. within centroid distance_cluster_3: -0.960
Avg. within centroid distance_cluster_4: -0.956
Avg. within centroid distance_cluster_5: -0.693
Avg. within centroid distance_cluster_6: -0.933
Avg. within centroid distance_cluster_7: -0.709
Davies Bouldin: -4.802
```

ii) Association analysis

- *Association Rules:* Based on the results we can say that the entities which are intactly associated are 'experience', 'python', 'machine', 'learning'. With confidence of 1.000 for both the association rules. Thus we can say "Experience in the Python programming language is the most common requirement in the field on Machine Learning."



- *Plot:*



4. Evaluation and conclusion

- The selection of dataset has been the most difficult part of this project as extraction of a meaningful context from any textual data is not easily possible.
- The results for the project came out well as expected.
- There was a lot of scope to perform the K-means clustering technique and association analysis on this data to showcase the current industry trend.

References

Amala Deshpande, Deepika Khatri, Divya Deshpande, Prarthita Das, & Sujata Khedkar. (2016). Proposed System for Resume Analytics. *International Journal Of Engineering Research And*, *V5*(11). https://doi.org/10.17577/ijertv5is110274

Frequent Item based Text Clustering: Big data Analytics. (2020), *14*(4). https://doi.org/10.37896/jxu14.4/400

GmbH, R. (2020). *k-Means - RapidMiner Documentation*. Docs.rapidminer.com. Retrieved 6 December 2020, from https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/k_means.html.