Analysis Report on Movies’ Overviews Dataset

Introduction:

The success of a movie plays a crucial role in movie industry because of the huge investments and the fame. In recent days, movies are produced and released in a high rate. Even though the movies are costing over million dollars to produce, most of them are still unsuccessful or worse, flop. Movie success depends upon the various attributes. So, to analyse this success, different predictions are made, and new patterns are discovered in data. These results can be used by the moviemakers before planning a movie.

In this project, the three analysis methods are performed on the dataset: Correlation analysis, Association analysis, and K-means clustering analysis.

From the correlation analysis, we want to know how much the quantitative attributes: the popularity, vote_count and vote_average, are correlated to each other. This correlation will be helped in determining the success of the movie.

From the Association analysis, we want to find the data items that have affinity for each other, using the association rules. Depending upon the rules, the data items that have high probability are taken into consideration while planning out a movie.

From the k-means clustering analysis, we need to know the categorized groups of data items from the overview of the movie story. This clustering helps in which kind of movie stories that audience are more interested in and what kind of stories are popular.

Dataset:

The dataset used in this Term project, is TMDB movies’ overview dataset. This dataset is taken from https://www.kaggle.com/writuparnabanerjee/hollywood-movies and was generated from the TMDB Movie Database API. This dataset contains the information about popular movies along with their overviews and the popularity, explained in 7 columns and 10,001 rows.

The columns of the dataset include:

Title: The title of the movie

Overview: The overview of the movie

Original language: The language in which the movie was originally released

Release date: The date on which the movie was released

Popularity: The factor shows how popular the movie is.

Vote_count: The number of votes for the movie

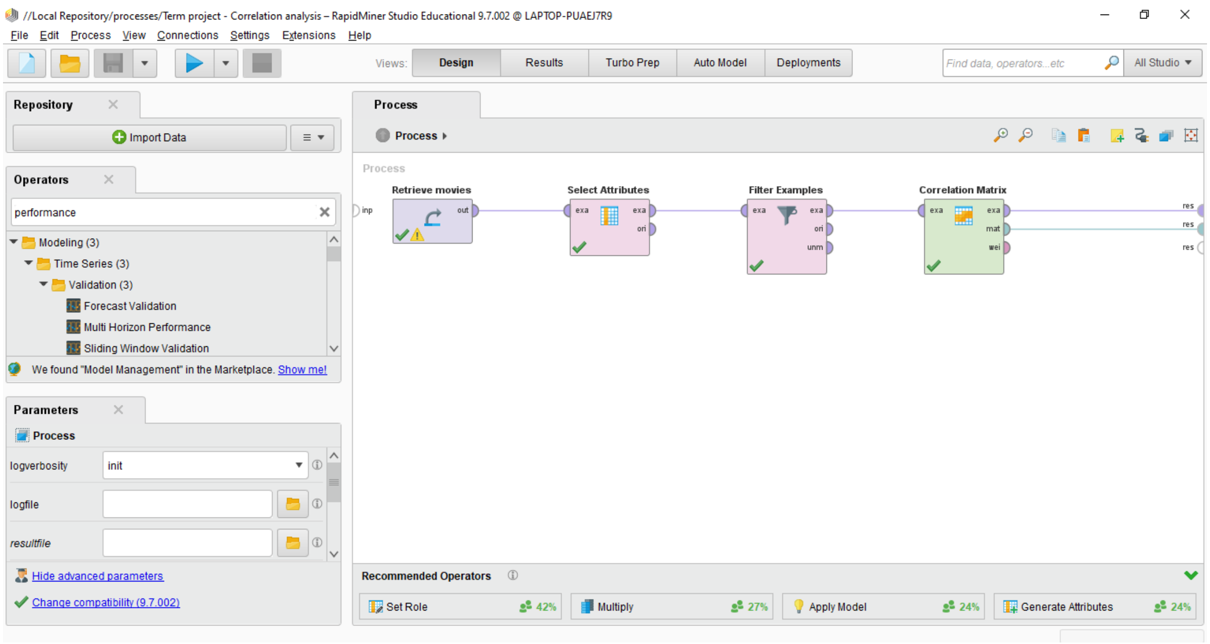
Vote_average: The average vote

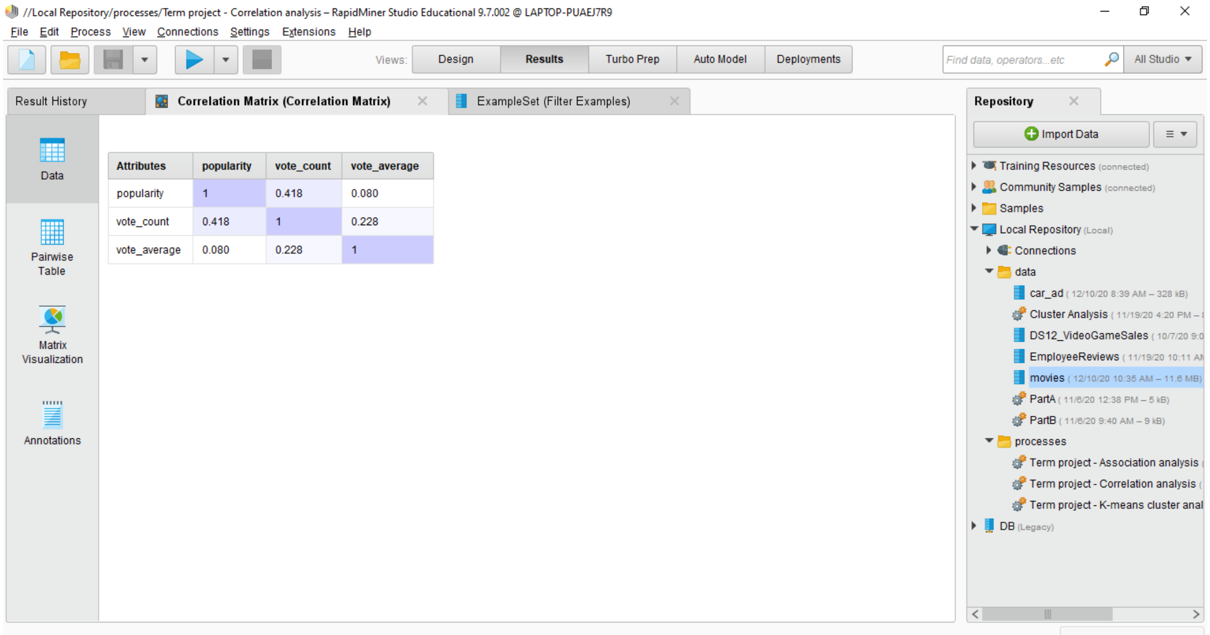
Data Cleaning:

The dataset used in this project, is cleaned by removing the first column since that column does not have any useful information. This dataset is already free from missing values and special characters so there is no requirement to replace them.

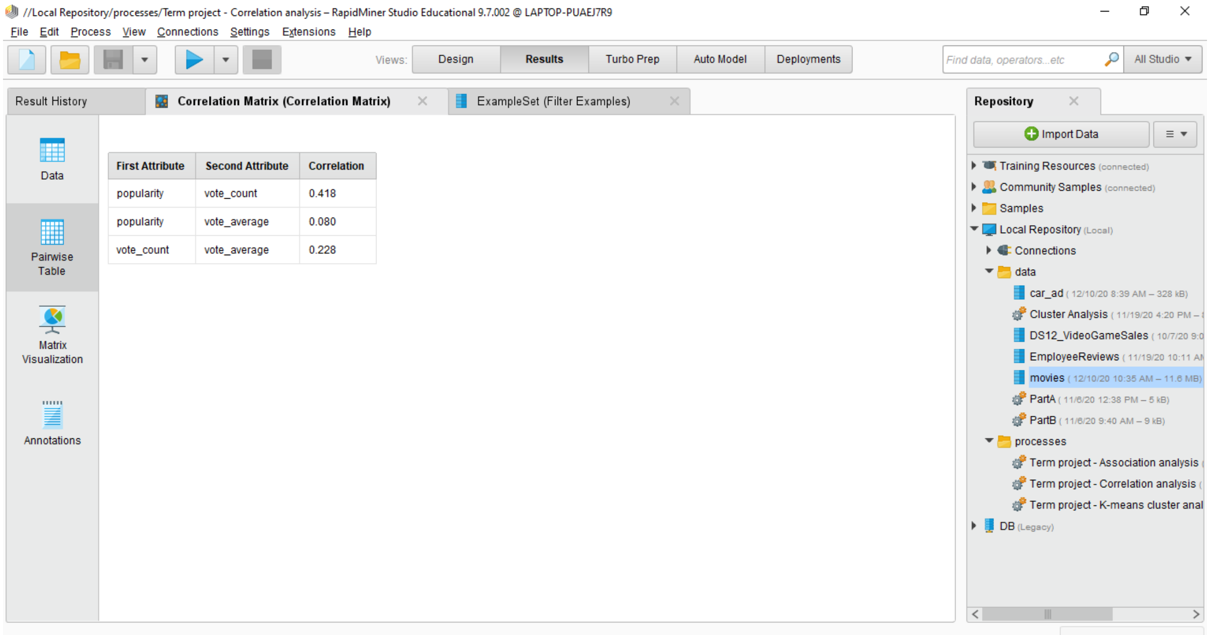
Datamining techniques:

Correlation analysis:

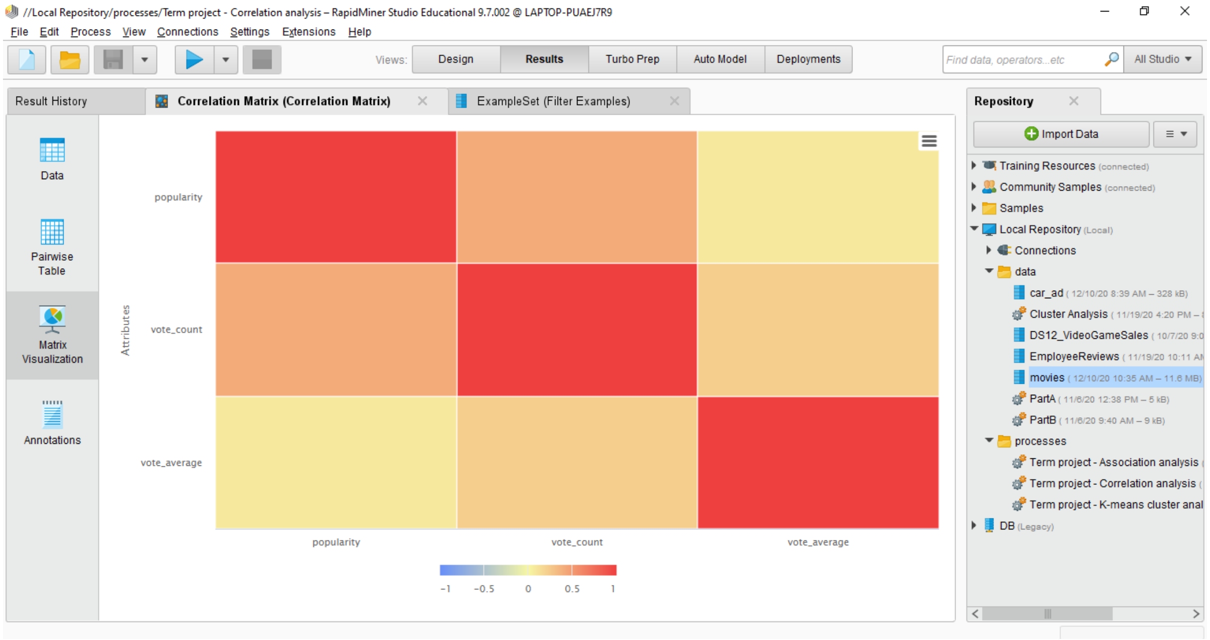

//Local Repository/processes/Term project - Correlation analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Operators
performance
Modeling (3)
Time Series (3)
Validation (3)
Forecast Validation
Multi Horizon Performance
Sliding Window Validation
We found "Model Management" in the Marketplace. Show me!
Parameters
Process
logverbosity init
logfile
resultfile
Hide advanced parameters
Change compatibility (9.7.002)
Process
Process
Retrieve movies
inp
out
Select Attributes
exa
exa
ori
Filter Examples
exa
exa
ori
unm
Correlation Matrix
exa
exa
mat
wei
res
res
res
Recommended Operators
Set Role 42%
Multiply 27%
Apply Model 24%
Generate Attributes 24%


//Local Repository/processes/Term project - Correlation analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
Correlation Matrix (Correlation Matrix)
ExampleSet (Filter Examples)
Data
Pairwise Table
Matrix Visualization
Annotations
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad (12/10/20 8:39 AM – 328 kB)
Cluster Analysis (11/19/20 4:20 PM –
DS12_VideoGameSales (10/7/20 9:0
EmployeeReviews (11/19/20 10:11 AM
movies (12/10/20 10:35 AM – 11.6 MB)
PartA (11/6/20 12:38 PM – 5 kB)
PartB (11/6/20 9:40 AM – 9 kB)
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)

Attributes	popularity	vote_count	vote_average
popularity	1	0.418	0.080
vote_count	0.418	1	0.228
vote_average	0.080	0.228	1

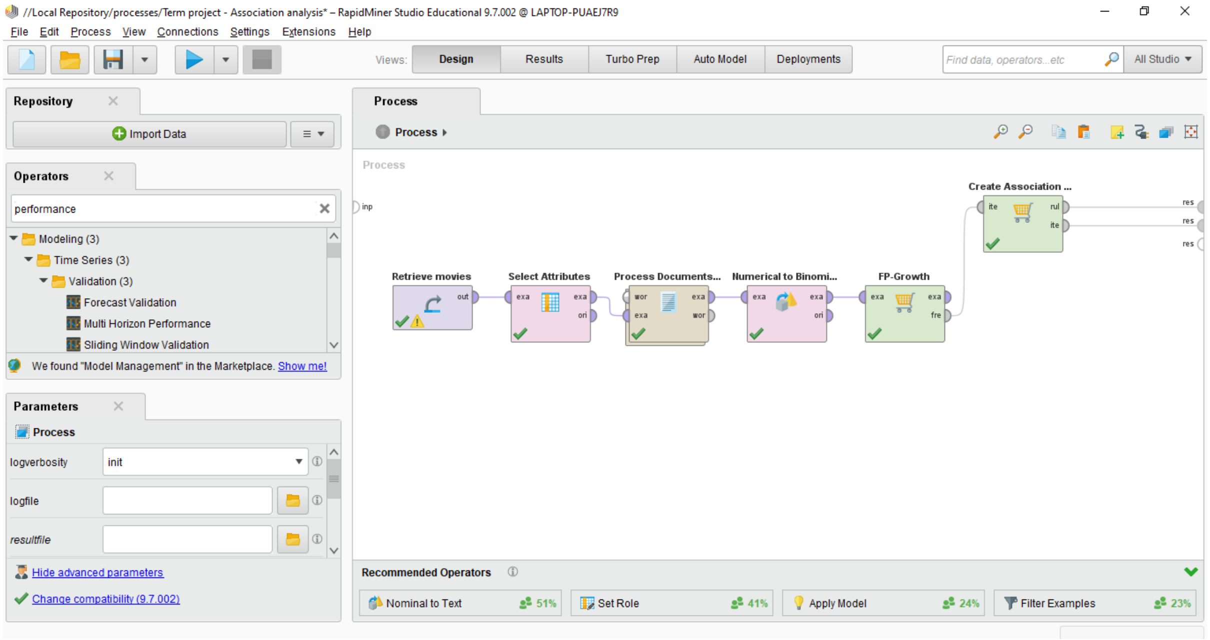

//Local Repository/processes/Term project - Correlation analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
Correlation Matrix (Correlation Matrix)
ExampleSet (Filter Examples)
Data
Pairwise Table
Matrix Visualization
Annotations
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad (12/10/20 8:39 AM – 328 kB)
Cluster Analysis (11/19/20 4:20 PM –
DS12_VideoGameSales (10/7/20 9:0
EmployeeReviews (11/19/20 10:11 AM
movies (12/10/20 10:35 AM – 11.6 MB)
PartA (11/6/20 12:38 PM – 5 kB)
PartB (11/6/20 9:40 AM – 9 kB)
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)

First Attribute	Second Attribute	Correlation
popularity	vote_count	0.418
popularity	vote_average	0.080
vote_count	vote_average	0.228


//Local Repository/processes/Term project - Correlation analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
Correlation Matrix (Correlation Matrix)
ExampleSet (Filter Examples)
Data
Pairwise Table
Matrix Visualization
Annotations
Attributes
popularity
vote_count
vote_average
popularity
vote_count
vote_average
-1
-0.5
0
0.5
1
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad (12/10/20 8:39 AM – 328 kB)
Cluster Analysis (11/19/20 4:20 PM –
DS12_VideoGameSales (10/7/20 9:0
EmployeeReviews (11/19/20 10:11 AM
movies (12/10/20 10:35 AM – 11.6 MB)
PartA (11/6/20 12:38 PM – 5 kB)
PartB (11/6/20 9:40 AM – 9 kB)
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)

The variable Popularity is positively correlated with the variable vote_count with coefficient: 0.418 and is also correlated with the variable vote_average with coefficient: 0.080. The variable vote_count is positively correlated with the variable vote_average like coefficient: 0.228. These correlations are explained through matrix visualization also. It is seen that there is least correlation between popularity and vote_average.

Association analysis:



//Local Repository/processes/Term project - Association analysis* – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9

Operators search: performance

- Modeling (3)
  - Time Series (3)
    - Validation (3)
      - Forecast Validation
      - Multi Horizon Performance
      - Sliding Window Validation

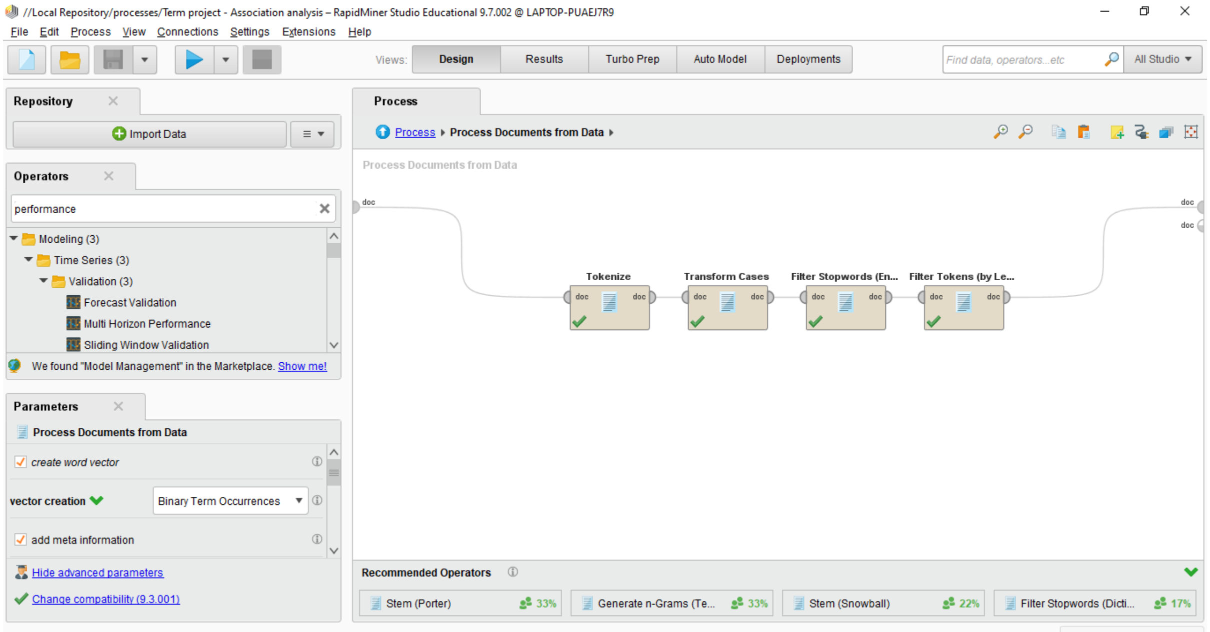
We found "Model Management" in the Marketplace. Show me!

Parameters – Process

logverbosity	init
logfile	
resultfile	

Hide advanced parameters

Change compatibility (9.7.002)


File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Operators
Process
Process ▸ Process Documents from Data ▸
Process Documents from Data
Tokenize
Transform Cases
Filter Stopwords (En...
Filter Tokens (by Le...
doc
Recommended Operators
Stem (Porter) 33%
Generate n-Grams (Te... 33%
Stem (Snowball) 22%
Filter Stopwords (Dicti... 17%

//Local Repository/processes/Term project - Association analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9

Operators search: performance

- Modeling (3)
  - Time Series (3)
    - Validation (3)
      - Forecast Validation
      - Multi Horizon Performance
      - Sliding Window Validation

We found "Model Management" in the Marketplace. Show me!

Parameters – Process Documents from Data

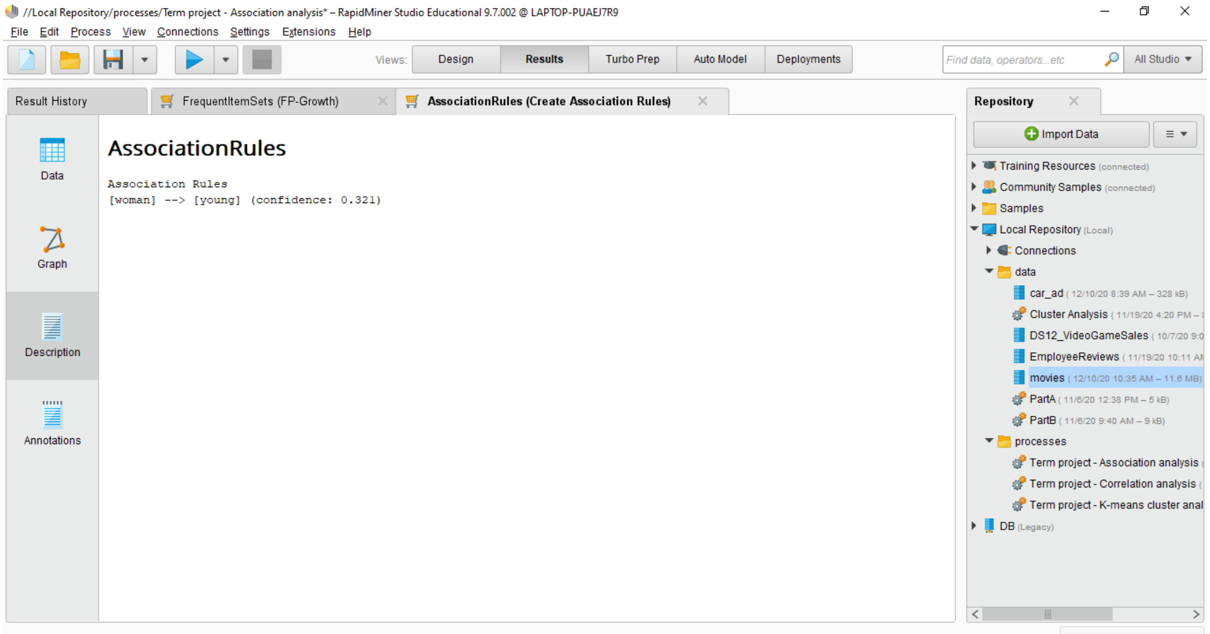
☑ create word vector

vector creation ✓ Binary Term Occurrences

☑ add meta information

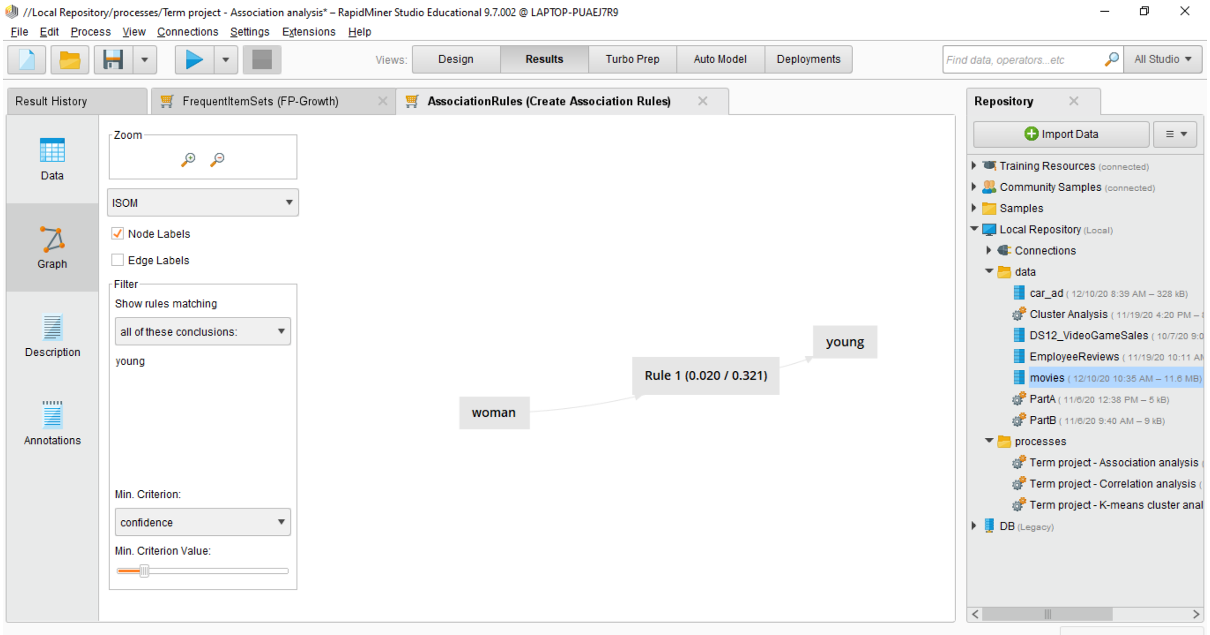
Hide advanced parameters

Change compatibility (9.3.001)


//Local Repository/processes/Term project - Association analysis* – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
FrequentItemSets (FP-Growth)
AssociationRules (Create Association Rules)
Data
Graph
Description
Annotations
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad (12/10/20 8:39 AM – 328 kB)
Cluster Analysis (11/19/20 4:20 PM –
DS12_VideoGameSales (10/7/20 9:0
EmployeeReviews (11/19/20 10:11 AM
movies (12/10/20 10:35 AM – 11.6 MB)
PartA (11/6/20 12:38 PM – 5 kB)
PartB (11/6/20 9:40 AM – 9 kB)
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)

AssociationRules

```
Association Rules
[woman] --> [young] (confidence: 0.321)
```


//Local Repository/processes/Term project - Association analysis* – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
FrequentItemSets (FP-Growth)
AssociationRules (Create Association Rules)
Data
Graph
Description
Annotations
Zoom
Filter
young
Rule 1 (0.020 / 0.321)
woman
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad (12/10/20 8:39 AM – 328 kB)
Cluster Analysis (11/19/20 4:20 PM –
DS12_VideoGameSales (10/7/20 9:0
EmployeeReviews (11/19/20 10:11 AM
movies (12/10/20 10:35 AM – 11.6 MB)
PartA (11/6/20 12:38 PM – 5 kB)
PartB (11/6/20 9:40 AM – 9 kB)
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)

ISOM

☑ Node Labels

□ Edge Labels

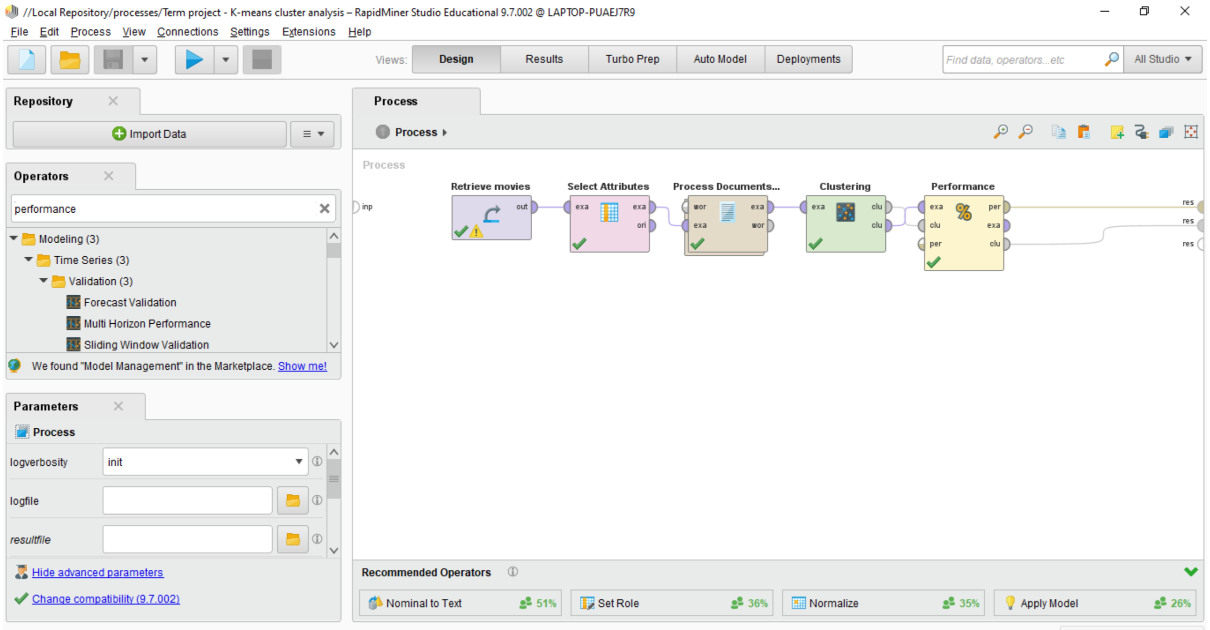
Show rules matching: all of these conclusions: young

Min. Criterion: confidence

Min. Criterion Value:

From the association rules, it is known that only data items ‘woman’ and ‘young’ have affinity for each other, with less confidence of 0.321. These rules depend upon the criteria I have given earlier while performing the analysis.

K-means clustering analysis:



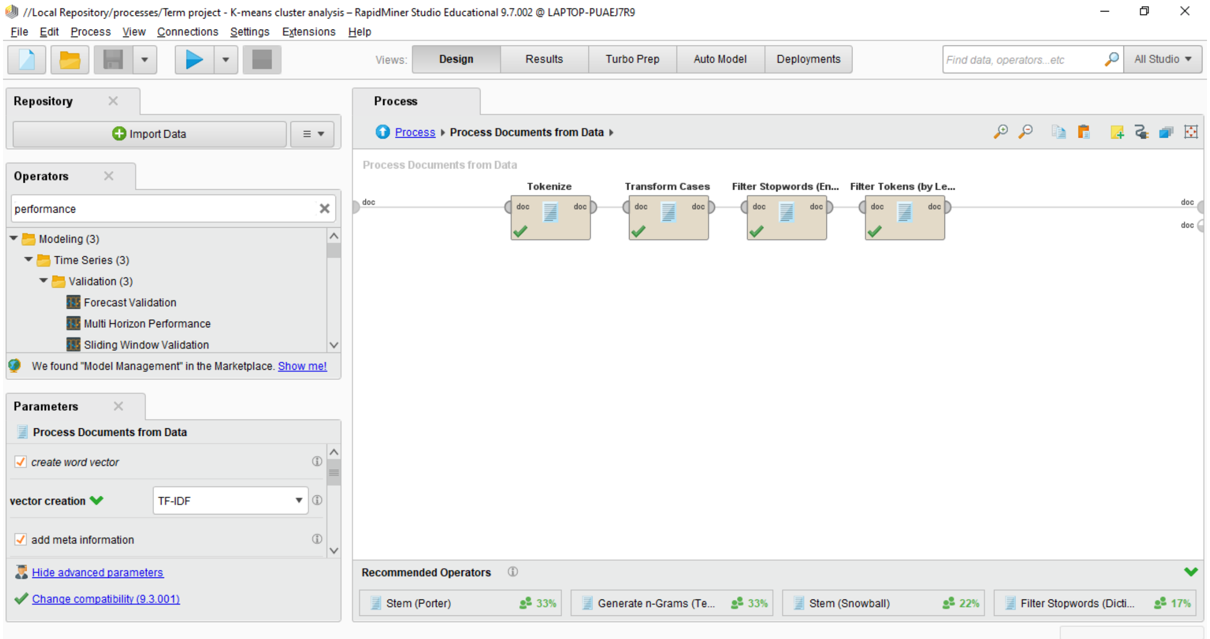
//Local Repository/processes/Term project - K-means cluster analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9

Parameters — Process

logverbosity: init

logfile:

resultfile:



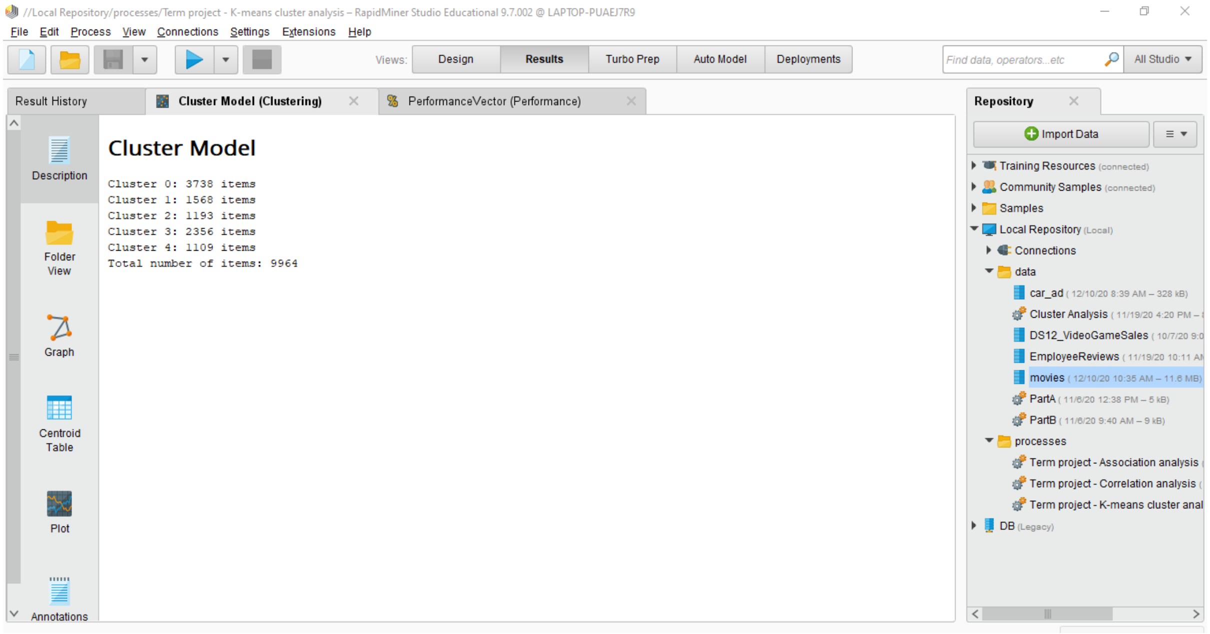
//Local Repository/processes/Term project - K-means cluster analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9

Parameters — Process Documents from Data

☑ create word vector

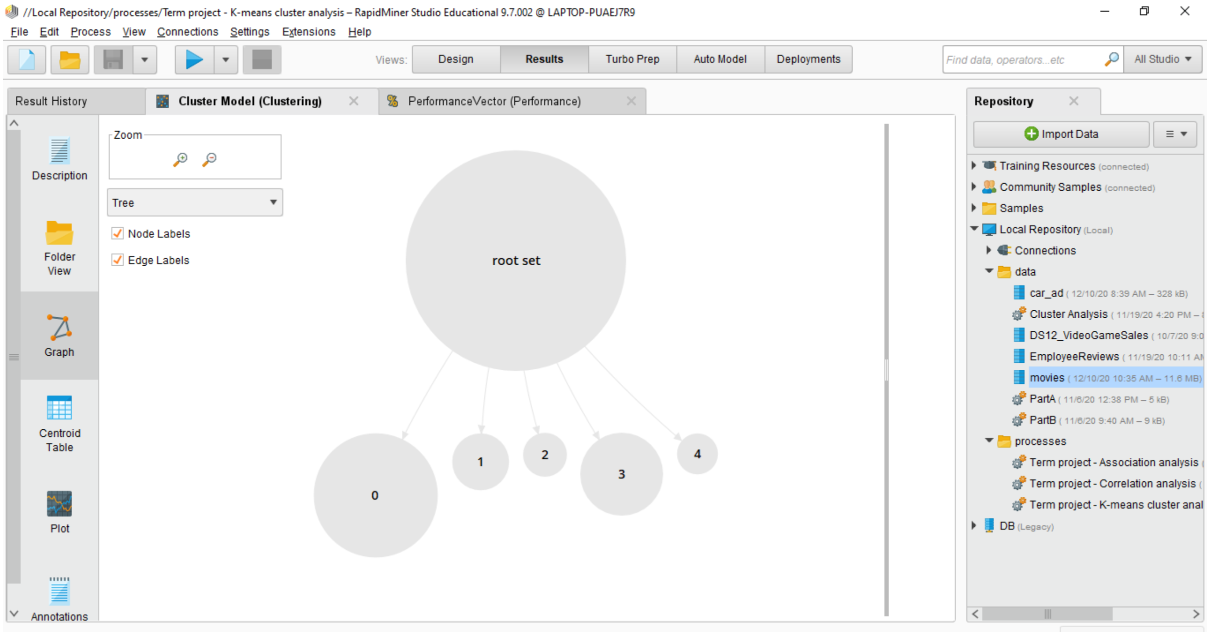
vector creation: TF-IDF

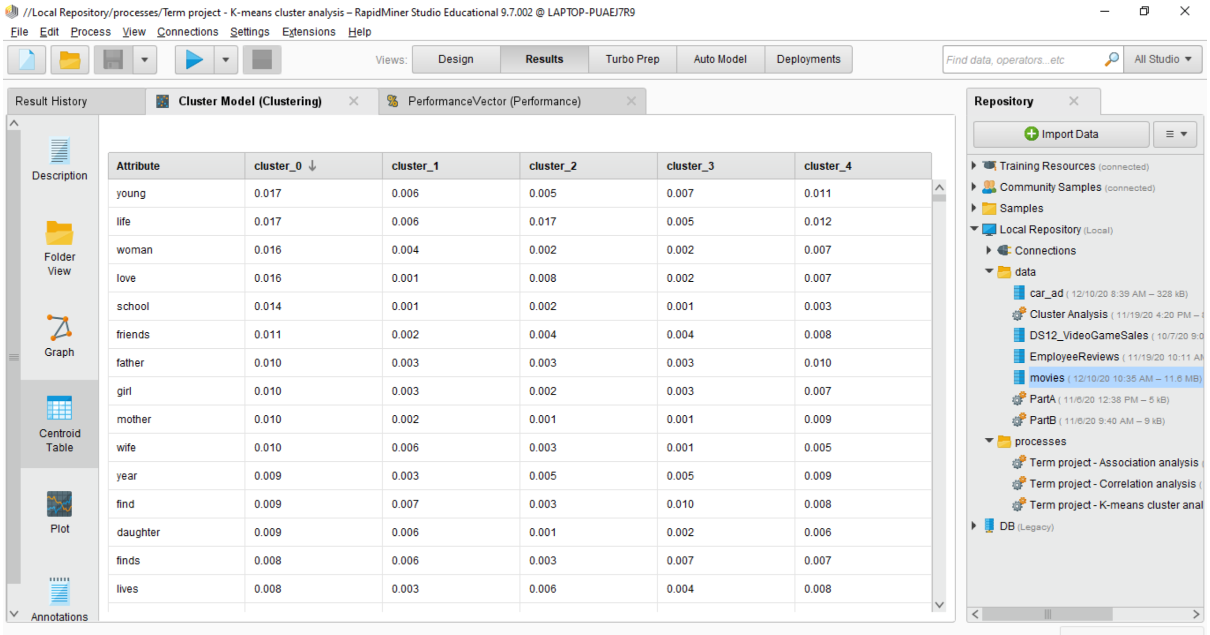
☑ add meta information


//Local Repository/processes/Term project - K-means cluster analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
PerformanceVector (Performance)
Description
Folder View
Graph
Centroid Table
Plot
Annotations
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad (12/10/20 8:39 AM – 328 kB)
Cluster Analysis (11/19/20 4:20 PM –
DS12_VideoGameSales (10/7/20 9:0
EmployeeReviews (11/19/20 10:11 A
movies (12/10/20 10:35 AM – 11.6 MB)
PartA (11/6/20 12:38 PM – 5 kB)
PartB (11/6/20 9:40 AM – 9 kB)
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)

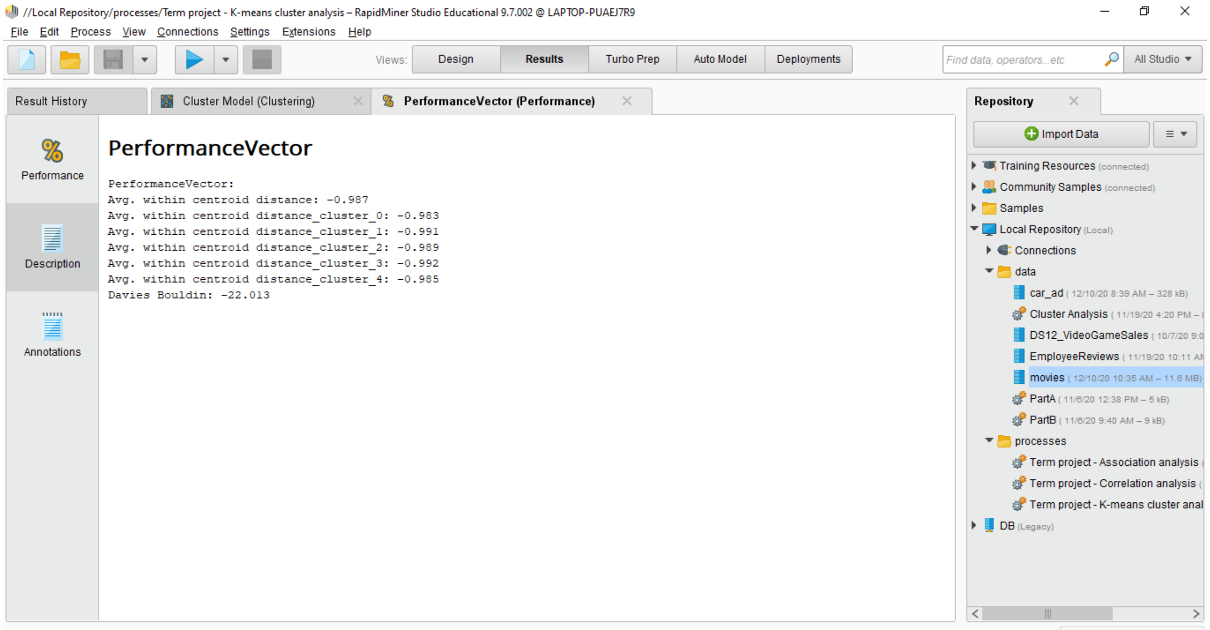
Cluster Model

```
Cluster 0: 3738 items
Cluster 1: 1568 items
Cluster 2: 1193 items
Cluster 3: 2356 items
Cluster 4: 1109 items
Total number of items: 9964
```


//Local Repository/processes/Term project - K-means cluster analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
PerformanceVector (Performance)
Description
Folder View
Graph
Centroid Table
Plot
Annotations
Zoom
Tree
☑ Node Labels
☑ Edge Labels
root set
0
1
2
3
4
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad (12/10/20 8:39 AM – 328 kB)
Cluster Analysis (11/19/20 4:20 PM –
DS12_VideoGameSales (10/7/20 9:0
EmployeeReviews (11/19/20 10:11 A
movies (12/10/20 10:35 AM – 11.6 MB)
PartA (11/6/20 12:38 PM – 5 kB)
PartB (11/6/20 9:40 AM – 9 kB)
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)


//Local Repository/processes/Term project - K-means cluster analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
PerformanceVector (Performance)
Description
Folder View
Graph
Centroid Table
Plot
Annotations
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad
Cluster Analysis
DS12_VideoGameSales
EmployeeReviews
movies
PartA
PartB
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)

Attribute	cluster_0 ↓	cluster_1	cluster_2	cluster_3	cluster_4
young	0.017	0.006	0.005	0.007	0.011
life	0.017	0.006	0.017	0.005	0.012
woman	0.016	0.004	0.002	0.002	0.007
love	0.016	0.001	0.008	0.002	0.007
school	0.014	0.001	0.002	0.001	0.003
friends	0.011	0.002	0.004	0.004	0.008
father	0.010	0.003	0.003	0.003	0.010
girl	0.010	0.003	0.002	0.003	0.007
mother	0.010	0.002	0.001	0.001	0.009
wife	0.010	0.006	0.003	0.001	0.005
year	0.009	0.003	0.005	0.005	0.009
find	0.009	0.007	0.003	0.010	0.008
daughter	0.009	0.006	0.001	0.002	0.006
finds	0.008	0.006	0.003	0.007	0.007
lives	0.008	0.003	0.006	0.004	0.008


//Local Repository/processes/Term project - K-means cluster analysis – RapidMiner Studio Educational 9.7.002 @ LAPTOP-PUAEJ7R9
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
PerformanceVector (Performance)
Performance
Description
Annotations
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Connections
data
car_ad
Cluster Analysis
DS12_VideoGameSales
EmployeeReviews
movies
PartA
PartB
processes
Term project - Association analysis
Term project - Correlation analysis
Term project - K-means cluster anal
DB (Legacy)

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.987
Avg. within centroid distance_cluster_0: -0.983
Avg. within centroid distance_cluster_1: -0.991
Avg. within centroid distance_cluster_2: -0.989
Avg. within centroid distance_cluster_3: -0.992
Avg. within centroid distance_cluster_4: -0.985
Davies Bouldin: -22.013
```

Centroid table gives the results showing the attributes and its cluster values, taken from the output of the performance operator.

From the performance vector, the average within the centroid distance is -0.0987.

Since the ‘k’ value is given as 5, there are four clusters: cluster_0, cluster_1, cluster_2, cluster_3 and cluster_4.

In Cluster_0, there are 3738 items and the series of first five words are young, life, woman, love, and school. This cluster may be about the movies with teenage love stories.

In Cluster_1, there are 1568 items and the series of first five words are police, agent, murder, detective, and killer. This cluster may be about the Crime thriller movies.

In Cluster_2, there are 1193 items and the series of first five words are film, story, movie, based, and documentary. This cluster may be about the biopic movies of film actors.

In Cluster_3, there are 2356 items and the series of first words are world, earth, group, team, evil, battle, and mission. This cluster may be about fictional character movies.

In Cluster_4, there are 1109 items and the series of first words are family, town, life, young, father, and mother. This cluster may be about the family drama movies.

Conclusion:

From the TMDB dataset, the popular movies are analysed based on various attributes. With the help of those attributes, the required criteria are calculated.

In this project, three different analysis methods are performed on the dataset to discover various patterns in the data. The predictions are made to know what attributes are correlated and associated. The overview of the movie story will be clustered to identify the discrete groups. However, success cannot be predicted based on a particular attribute. So, these models are based on interesting relations between attributes.

The movie industry persons like the directors, the script writers, the producers so on, can use these results to modify the movie criteria and plan a movie in advance to attain the successful blockbusters.

Not only the movie makers, even the audience can benefit from these models. Before buying a ticket and investing time, the movie watchers can know about the film according to their preferences and determine whether it is good movie, or a flop one.