TEXT MINING WITH RAPIDMINER

G. Ertek, D. Tapucu, and I. Arın
Sabancı University, Istanbul, Turkey

The goal of this chapter is to introduce the text mining capabilities of RAPIDMINER through a use case. The use case involves mining reviews for hotels at TripAdvisor.com, a popular web portal. We will be demonstrating basic text mining in RAPIDMINER using the text mining extension. We will present two different RAPIDMINER processes, namely Process01 andProcess02, which respectively describe how text mining can be combined with association mining and cluster modeling. While it is possible to construct each of these processes from scratch by inserting the appropriate operators into the process view, we will instead import these two processes readily from existing model files. Throughout the chapter, we will at times deliberately instruct the reader to take erroneous steps that result in undesired outcomes. We believe that this is a very realistic way of learning to use RAPIDMINER, since in practice, the modeling process frequently involves such steps that are later corrected.

USE CASES WITH RAPIDMINER

USE CASES WITH RAPIDMINER
Working Title

Dr. Markus Hofmann
Institute of Technology Blanchardstown, Ireland

Ralf Klinkenberg
Rapid-i

A JOHN WILEY & SONS, INC., PUBLICATION

Library of Congress Cataloging-in-Publication Data:

Use Cases with RapidMiner / Markus Hofmann. . . [et al.].
Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

to be added

CONTRIBUTORS

GURDAL ERTEK, Sabancı University, Istanbul, Turkey
DILEK TAPUCU, Sabancı University, Istanbul, Turkey
INANC ARIN, Sabancı University, Istanbul, Turkey

CONTENTS IN BRIEF

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

TEXT MINING WITH RAPIDMINER

G. Ertek, D. Tapucu, and I. Arin

Sabancı University, Istanbul, Turkey

1.1 INTRODUCTION

The goal of this chapter is to introduce the text mining capabilities of RapidMiner through a use case. The use case involves mining reviews for hotels at TripAdvisor.com, a popular web portal. We will be demonstrating basic text mining in RapidMiner using the text mining extension. We will present two different RapidMiner processes, namely Process01 and Process02, which respectively describe how text mining can be combined with association mining and cluster modeling. While it is possible to construct each of these processes from scratch by inserting the appropriate operators into the process view, we will instead import these two processes readily from existing model files.

Throughout the chapter, we will at times deliberately instruct the reader to take erroneous steps that result in undesired outcomes. We believe that this is a very realistic way of learning to use RapidMiner, since in practice, the modeling process frequently involves such steps that are later corrected.

1.1.1 Text Mining

Text mining (also referred to as *text data mining* or *knowledge discovery from textual databases*), refers to the process of discovering interesting and non-trivial knowledge from text documents. The common practice in text mining is the analysis of the information extracted through text processing to form new facts and new hypotheses, that can be explored further with other data mining algorithms. Text mining applications typically deal with large and complex data sets of textual documents that contain significant amount of irrelevant and noisy information. Feature selection aims to remove this irrelevant and noisy information by focusing only on relevant and informative data for use in text mining. Some of the topics within text mining include feature extraction, text categorization, clustering, trends analysis, association mining and visualization.

1.1.2 Data Description

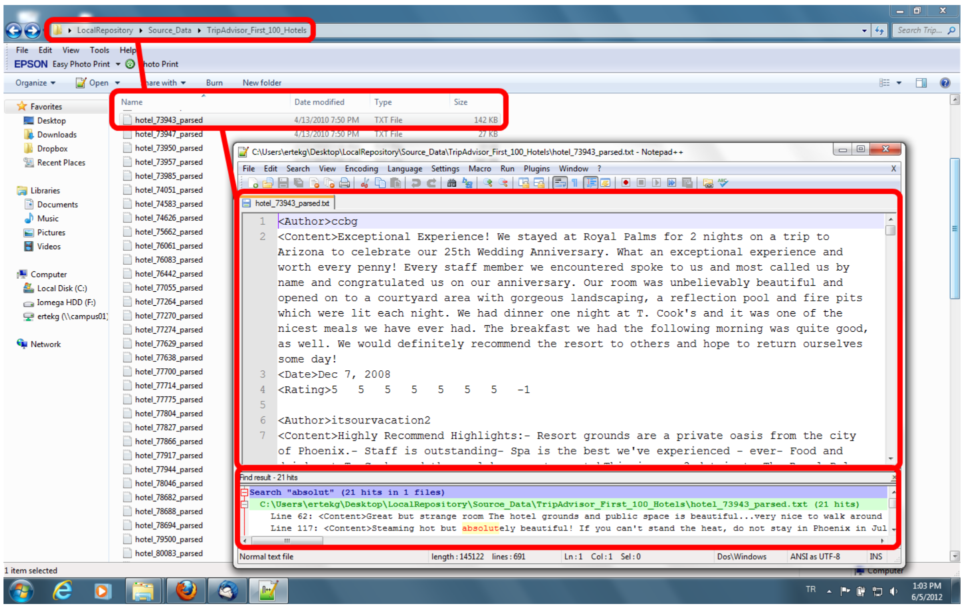
The files required for this chapter, including the data and pre-built processes reside within a folder titled `LocalRepository`. The data used in this chapter comes from TripAdvisor.com, a popular web portal in the hospitality industry, and is shown in Figure 1.1. This publicly available data set contains the reviews and ratings (1 through 5) of clients or customers for 1850 hotels. The original data was extracted by The Database and Systems Information Laboratory at the University of Illinois at Urbana-Champaign,and is available under `http://sifaka.cs.uiuc.edu/~wang296/Data/`. There are 1850 text documents in the original data set,corresponding to the reviews of 1850 hotels. Each document contains all the reviews for that hotel. While it is possible to run text mining processes with the original data, we will be using a subset of the data containing only the first 100 text documents. The data set used in this chapter may be downloaded from `http://people.sabanciuniv.edu/ertekg/papers/supp/09.zip` or the short web link `http://bit.ly/Kzdn5x`, as well as http://research.sabanciuniv.edu .

1.1.3 Running RapidMiner

When running RapidMiner, it is strongly recommended to right click the mouse button on the start menu and Run as administrator, as shown in Figure 1.2, rather than simply clicking or double clicking the executable or shortcut icon. By running as administrator, we are granted the permissions to install extension packages. Running the software without the administrator rights may cause errors when trying to install updates, including the extension packages. Figure 1.3 shows the splash screen displayed when RapidMiner is run for the first time or when it is run without any extensions installed. In Figure 1.3, the highlighted region is currently blank but it will later contain the icons for the installed extensions. When RapidMiner is loaded, the user starts with the *welcome perspective* (a *perspective* is a particular combination of *views*), as shown in Figure 1.4.

1.1.4 RapidMiner Text Processing Extension Package

RapidMiner is the most popular open source software in the world for data mining, and strongly supports text mining and other data mining techniques that are applied in combination with text mining. The power and flexibility of RapidMiner is due to the GUI-based IDE (integrated development environment) it provides for rapid prototyping



```
C:\Users\ertekg\Desktop\LocalRepository\Source_Data\TripAdvisor_First_100_Hotels\hotel_73943_parsed.txt - Notepad++

<Author>ccbg
<Content>Exceptional Experience! We stayed at Royal Palms for 2 nights on a trip to
Arizona to celebrate our 25th Wedding Anniversary. What an exceptional experience and
worth every penny! Every staff member we encountered spoke to us and most called us by
name and congratulated us on our anniversary. Our room was unbelievably beautiful and
opened on to a courtyard area with gorgeous landscaping, a reflection pool and fire pits
which were lit each night. We had dinner one night at T. Cook's and it was one of the
nicest meals we have ever had. The breakfast we had the following morning was quite good,
as well. We would definitely recommend the resort to others and hope to return ourselves
some day!
<Date>Dec 7, 2008
<Rating>5   5   5   5   5   5   5   -1

<Author>itsourvacation2
<Content>Highly Recommend Highlights:- Resort grounds are a private oasis from the city
of Phoenix.- Staff is outstanding- Spa is the best we've experienced - ever- Food and

Find result - 21 hits
Search "absolut" (21 hits in 1 files)
  C:\Users\ertekg\Desktop\LocalRepository\Source_Data\TripAdvisor_First_100_Hotels\hotel_73943_parsed.txt (21 hits)
    Line 62: <Content>Great but strange room The hotel grounds and public space is beautiful...very nice to walk around
    Line 117: <Content>Steaming hot but absolutely beautiful! If you can't stand the heat, do not stay in Phoenix in Jul

Normal text file    length : 145122    lines : 691    Ln : 1    Col : 1    Sel : 0    Dos\Windows    ANSI as UTF-8    INS
```

Figure 1.1 The TripAdvisor data set

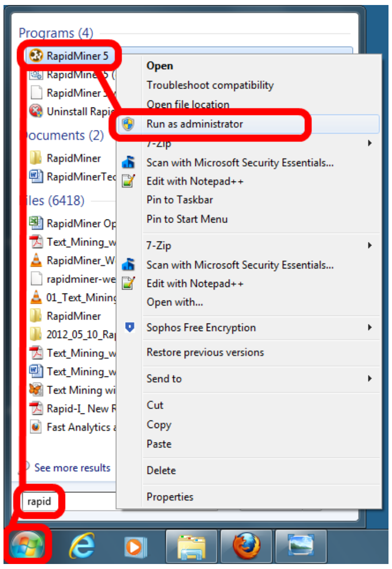
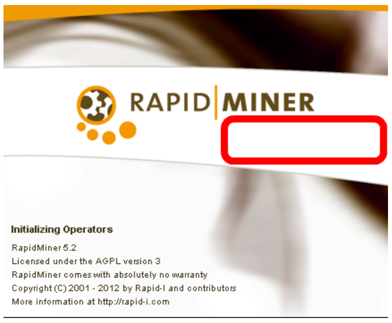


Figure 1.2 Running RAPIDMINER as administrator

and development of data mining models, as well as its strong support for scripting based on XML (extensible mark-up language). The visual modeling in the RAPIDMINER IDE is based on the defining of the data mining process in terms of operators and the flow of



Initializing Operators
RapidMiner 5.2
Licensed under the AGPL version 3
RapidMiner comes with absolutely no warranty
Copyright (C) 2001 - 2012 by Rapid-I and contributors
More information at http://rapid-i.com

Figure 1.3 RAPIDMINER splash screen when no extension packages are installed

process through these operators. Users specify the expected inputs, the delivered outputs, the mandatory and optional parameters, and the core functionalities of the operators, and the complete process is automatically executed by RAPIDMINER. Many packages are available for RAPIDMINER, such as text processing, Weka extension, parallel processing, web mining, reporting extension, series processing, PMML, community, and R extension packages. The package that is needed and used for text mining is the *Text Processing* package, which can be installed and updated through the Update RapidMiner menu item under the Help menu.

1.1.5 Installing Text Mining Extensions

We will initiate our text mining analysis by importing the two previously built processes. However, even before that we have to check and make sure that the extensions required for text mining are installed within RAPIDMINER. To manage the extensions, select the Help menu and then manage Extensions menu item, as shown in Figure 1.4. The dialog box that comes up, as shown in Figure 1.5, does not list any extensions. Thus, the extensions required have to be installed.

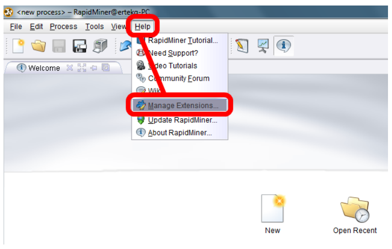
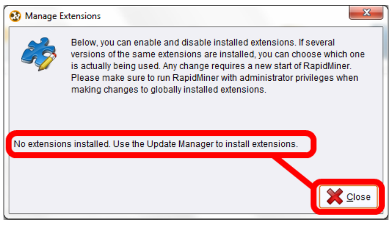


Figure 1.4 Managing RAPIDMINER extension packages


Manage Extensions

Below, you can enable and disable installed extensions. If several versions of the same extensions are installed, you can choose which one is actually being used. Any change requires a new start of RapidMiner. Please make sure to run RapidMiner with administrator privileges when making changes to globally installed extensions.

No extensions installed. Use the Update Manager to install extensions.

Figure 1.5 Dialog box stating that no extension packages are yet installed

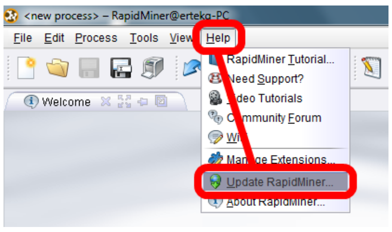
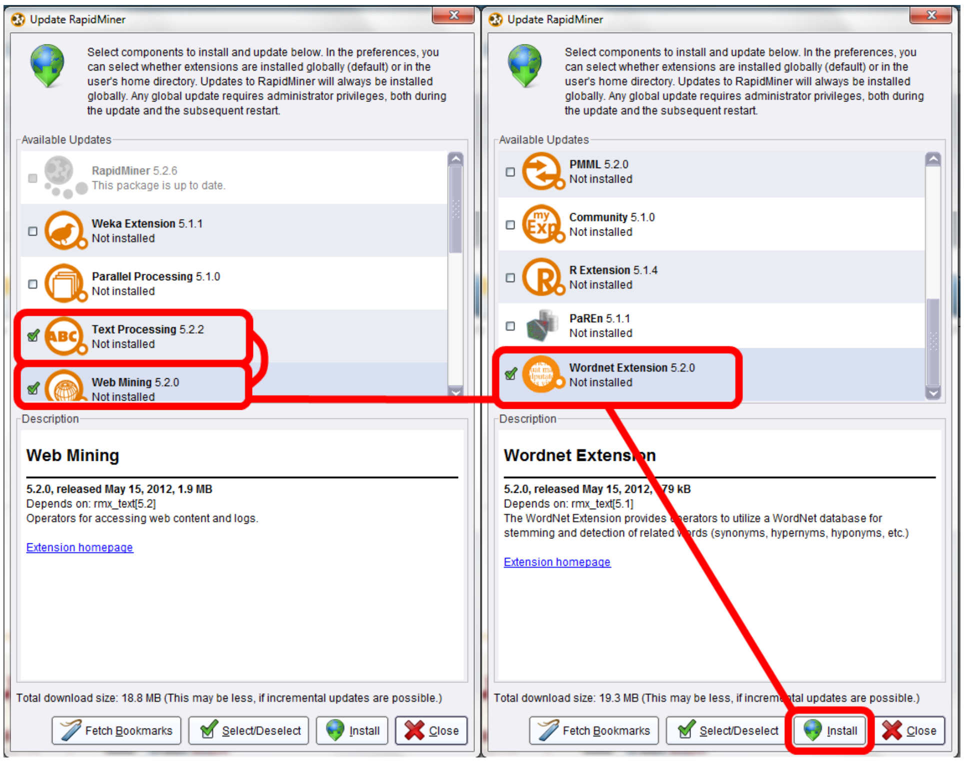


Figure 1.6 Updating RAPIDMINER


Update RapidMiner

Select components to install and update below. In the preferences, you can select whether extensions are installed globally (default) or in the user's home directory. Updates to RapidMiner will always be installed globally. Any global update requires administrator privileges, both during the update and the subsequent restart.

Available Updates

- RapidMiner 5.2.6 — This package is up to date.
- Weka Extension 5.1.1 — Not installed
- Parallel Processing 5.1.0 — Not installed
- ✓ Text Processing 5.2.2 — Not installed
- ✓ Web Mining 5.2.0 — Not installed

Description

Web Mining

5.2.0, released May 15, 2012, 1.9 MB
Depends on: rmx_text[5.2]
Operators for accessing web content and logs.

Extension homepage

Total download size: 18.8 MB (This may be less, if incremental updates are possible.)

Update RapidMiner

Select components to install and update below. In the preferences, you can select whether extensions are installed globally (default) or in the user's home directory. Updates to RapidMiner will always be installed globally. Any global update requires administrator privileges, both during the update and the subsequent restart.

Available Updates

- PMML 5.2.0 — Not installed
- Community 5.1.0 — Not installed
- R Extension 5.1.4 — Not installed
- PaREn 5.1.1 — Not installed
- ✓ Wordnet Extension 5.2.0 — Not installed

Description

Wordnet Extension

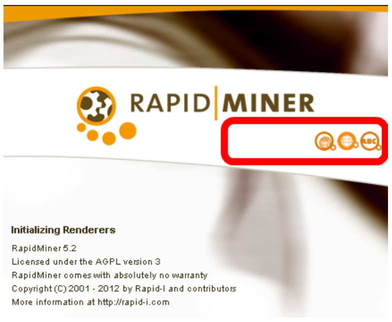
5.2.0, released May 15, 2012, 79 kB
Depends on: rmx_text[5.1]
The WordNet Extension provides operators to utilize a WordNet database for stemming and detection of related words (synonyms, hypernyms, hyponyms, etc.)

Extension homepage

Total download size: 19.3 MB (This may be less, if incremental updates are possible.)

Figure 1.7 Installing extension packages for text mining

Click the Close button and select Help menu and then Update RapidMiner menu item as shown in Figure 1.6. RAPIDMINER will connect to the internet and fetch the list of available updates, eventually displaying all the available updates, as in Figure 1.7. In this window, the selection of an update for installation can be made only by double clicking the check box on the left hand side of that update's name. When an update is selected for installation, a small green check sign appears on the check box, as in Figure 1.7. The extension packages (available as updates) needed for text mining are Text Processing, Web Mining, and Wordnet Extension. Select these updates as shown in Figure 1.7. Then, click the install button. When the Terms of Use window appears, click the radio button for I accept the terms of use, and click Ok . During the installation of updates, the installation process can be stopped by first clicking inside the Progress window, and clicking Stop , but don't do this now. When the Update Complete dialog box appears, click Yes to restart RAPIDMINER with the newly installed updates available . When the splash screen for RAPIDMINER is displayed, you will notice that the icons for the installed extensions appear in the previously blank region (Figure 1.8).



Initializing Renderers

RapidMiner 5.2
Licensed under the AGPL version 3
RapidMiner comes with absolutely no warranty
Copyright (C) 2001 - 2012 by Rapid-I and contributors
More information at http://rapid-i.com

Figure 1.8 RAPIDMINER splash screen with the three extension packages installed for text mining

1.2 ASSOCIATION MINING OF TEXT DOCUMENT COLLECTION (Process01)

1.2.1 Importing Process01

We will now initiate the text mining analysis by importing the processes supplied with this chapter. For this, select Files and then the Import Process menu item as in Figure 1.9. Go to the `LocalRepository` folder (supplied with this chapter), then to `Processes` folder, click on `Process01.rmp`, and click the Open button, as in Figure 1.10. The RapidMiner process `Process01` is now displayed in the *design perspective*, as shown in Figure 1.11.

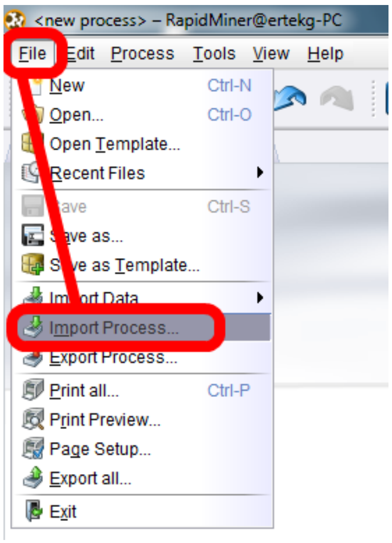
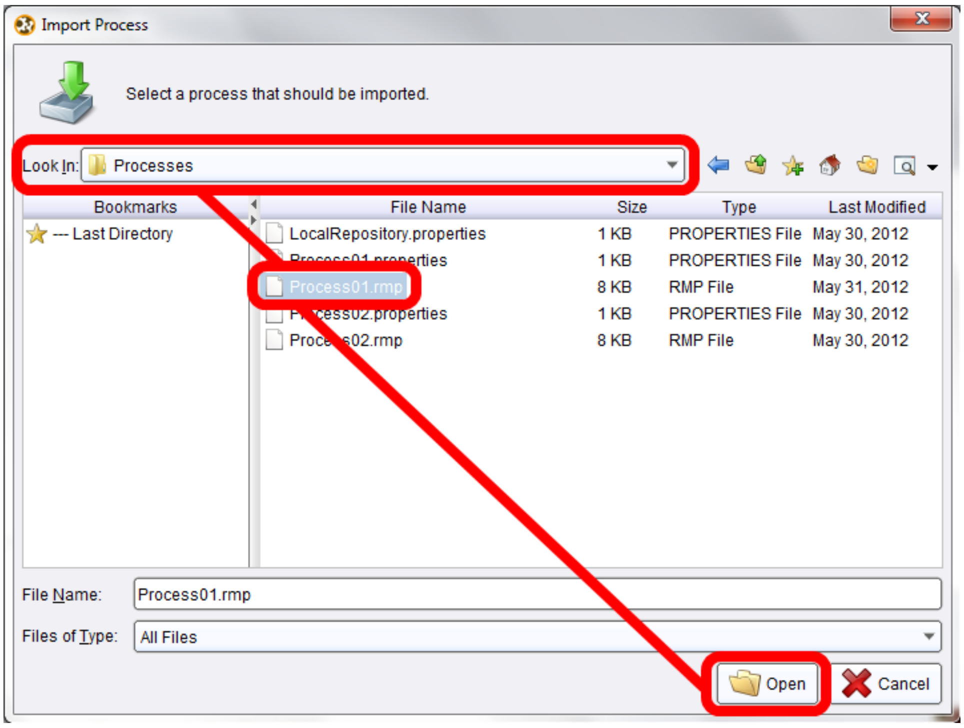


Figure 1.9 Importing an existing process


Import Process

Select a process that should be imported.

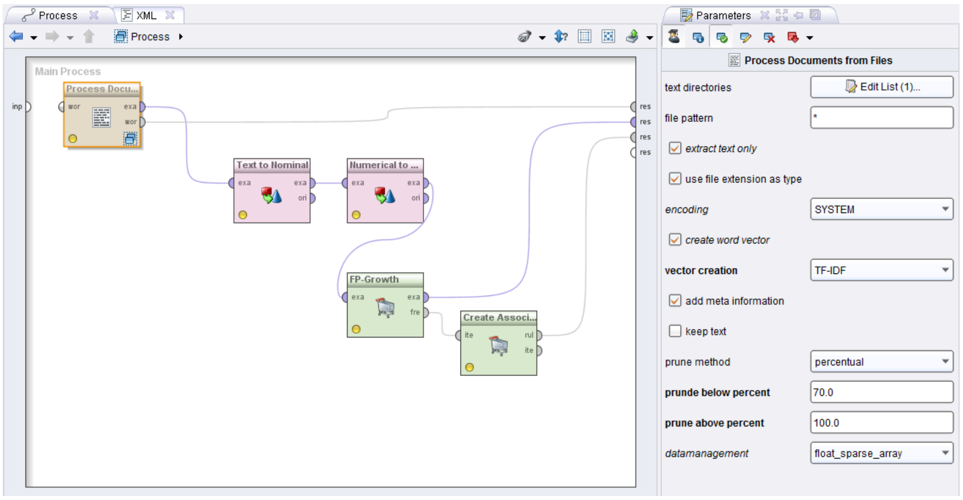
Look In: Processes

File Name	Size	Type	Last Modified
LocalRepository.properties	1 KB	PROPERTIES File	May 30, 2012
Process01.properties	1 KB	PROPERTIES File	May 30, 2012
Process01.rmp	8 KB	RMP File	May 31, 2012
Process02.properties	1 KB	PROPERTIES File	May 30, 2012
Process02.rmp	8 KB	RMP File	May 30, 2012

File Name: Process01.rmp

Files of Type: All Files

Figure 1.10 Selecting the process to import



Process Documents from Files

text directories	Edit List (1)...
file pattern	*
extract text only	✓
use file extension as type	✓
encoding	SYSTEM
create word vector	✓
vector creation	TF-IDF
add meta information	✓
keep text	☐
prune method	percentual
prunde below percent	70.0
prune above percent	100.0
datamanagement	float_sparse_array

Figure 1.11 Operators for Process01 and the Parameters for Process Documents from Files operator

1.2.2 Operators in Process01

The parameters for the operators in this process are given on the right-hand side of the process, listed under the Parameters tab. For the first process, the parameter text directories specifies where to read the text data from. A very important parameter is the vector creation used. In Process01, the selected vector creation method is TF-IDF *TF-IDF* (*Term Frequency-Inverse Document Frequency*) is a term weighting method. It gives higher weight to terms that appear frequently in the document but not many times in other documents. However, this may yield too many words, even up to tens of thousands of them. Too many words would make it prohibitive to carry out the successive data mining steps due to lengthy running times for the data mining algorithms. Hence, it is a very good idea to prune the resulting word set using a prune method, by selecting the method and its parameters within the Parameters view.

Click on the Process Documents from Files operator, as in Figure 1.11. In Process01, words that appear in less than 70.0 % of the documents are pruned, as can be seen from the value of 70.0 for the prune below percent parameter. It is also possible to prune the words that appear in too many documents, but this was not done in this example, as can be seen from the value of 100.0 for the prune above percent parameter. In association mining, we are interested in the items (words in text mining context) that appear in 100.0% of the transactions (documents in our text mining context), since they can form interesting frequent itemsets (word lists) and association rules, that provide actionable insights. Thus, it is appropriate to set prune above percent to 100.0 in Process01 (Figure 1.11), including the items (words) that appear in every document.

Process01 consists of five operators (Figure 1.11). Firstly, the Process Documents from Files operator performs text processing which involves preparing the text data for the application of conventional data mining techniques. Process Documents from Files operator reads data from a collection of text files and manipulates this data using text processing algorithms. This is a nested operator, meaning that it can contain a sub-process consisting of a multitude of operators. Indeed, in `Process01`, this nested operator contains other operators inside. Double click on this operator, and you will see the sub-process inside it, as outlined in Figure 1.12. This sub-process consists of six operators that are serially linked (Figure 1.12):

- Tokenize Non-letters (Tokenize)
- Tokenize Linguistic (Tokenize)
- Filter Stopwords (English)
- Filter Tokens (by Length)
- Stem (Porter)
- Transform Cases (2) (Transform Cases)

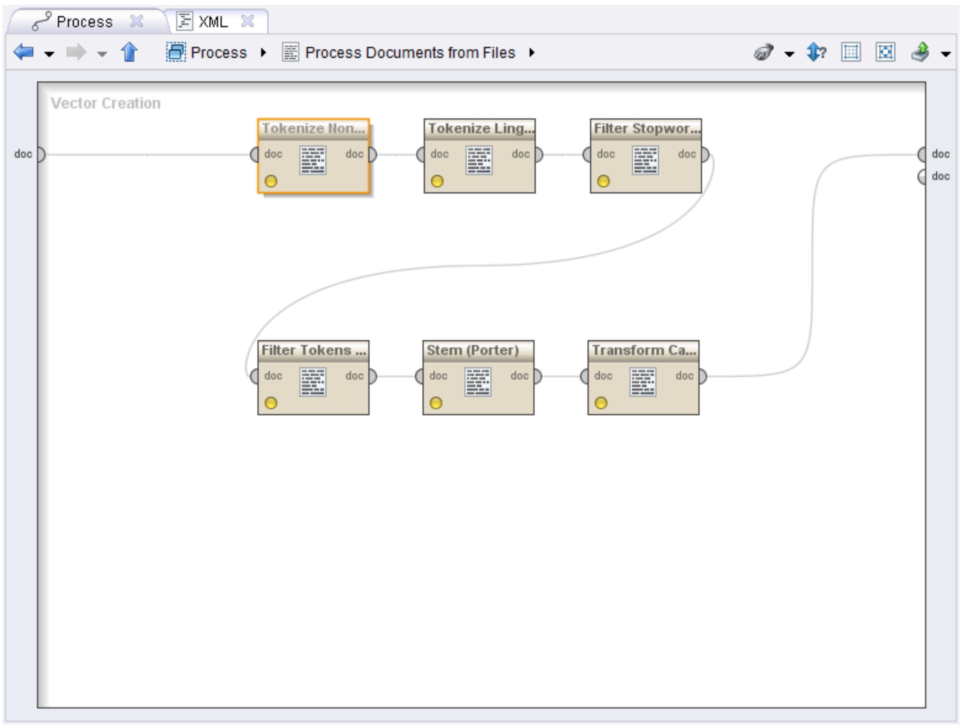
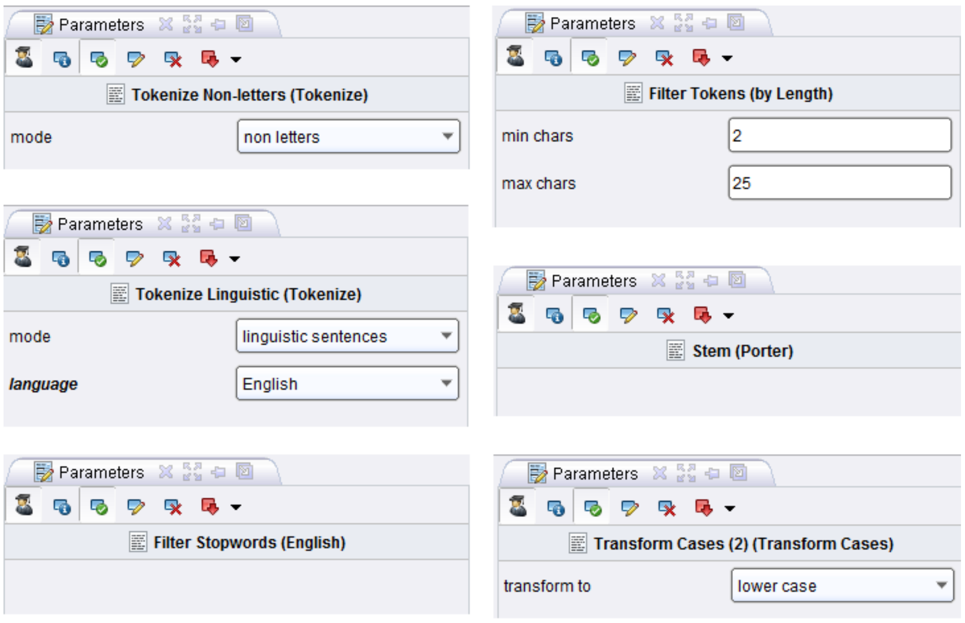


Figure 1.12 Operators within the Process Documents from Files nested operator


Parameters

Tokenize Non-letters (Tokenize)

mode	non letters

Parameters

Filter Tokens (by Length)

min chars	2
max chars	25

Parameters

Tokenize Linguistic (Tokenize)

mode	linguistic sentences
language	English

Parameters

Stem (Porter)

Parameters

Filter Stopwords (English)

Parameters

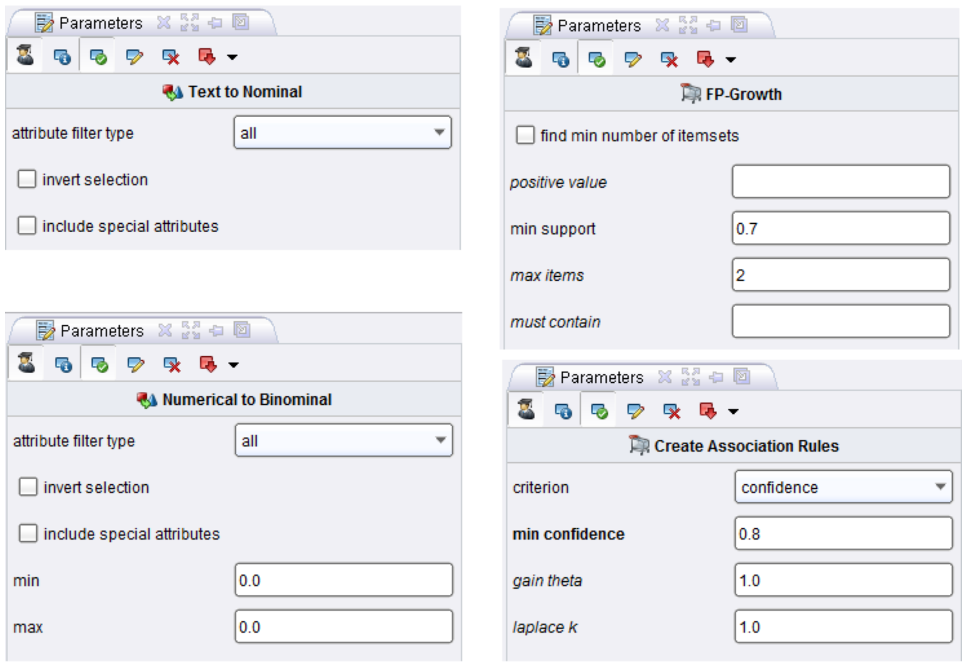
Transform Cases (2) (Transform Cases)

transform to	lower case

Figure 1.13 Parameters for the operators within the Process Documents from Files operator

The sub-process basically transforms the text data into a format that can be easily analyzed using conventional data mining techniques such as association mining and cluster modeling. The parameters for each of the operators in this sub-process (within the Process Documents from Files operator) are displayed in Figure 1.13. Notice that Figure 1.13 was created manually by combining multiple snapshots.

In this sub-process, the Tokenize Non-letters (Tokenize) and Tokenize Linguistic (Tokenize) operators are both created by selecting the Tokenize operator, but with different parameter selections. The former operater tokenizes based on non letters whereas the latter operater tokenizes based on the linguistic sentences within the English language. The Filter Stopwords (English) operator removes the stop words in the English language from the text data set. The Filter Tokens (by Length) operator removes all the words composed of less than min chars characters and more than max chars characters. In this example, words that have less than 2 characters or more than 25 characters are removed from the data set. The Stem (Porter) operator performs stemming and the Transform Cases(2) (Transform Cases) operator transforms all the characters in the text into lower case. It should be noted that the name of this last operator is not a good name, and we, as the modelers have forgotten to rename the operator after its name was automatically assigned by RapidMiner. This mistake should be avoided in constructing the processes.



Text to Nominal

attribute filter type	all
☐ invert selection	
☐ include special attributes	

Numerical to Binominal

attribute filter type	all
☐ invert selection	
☐ include special attributes	
min	0.0
max	0.0

FP-Growth

☐ find min number of itemsets	
positive value	
min support	0.7
max items	2
must contain	

Create Association Rules

criterion	confidence
min confidence	0.8
gain theta	1.0
laplace k	1.0

Figure 1.14 Parameters for the operators in Process01

The parameters for each of the operators in Process01 (Figure 1.11) are displayed in Figure 1.14. Notice that Figure 1.14 was created manually by combining multiple snapshots. The Text to Nominal operator transforms the text data into nominal (categorical) data. The Numerical to Binomial operator then transforms the data into binominal form. This means that each row represents a document, a few columns provide meta-data about that document and the remaining columns represent the words appearing in all the documents, with the cell contents telling (true or false) whether that word exist in that document or not. FP-Growth algorithm is used for identifying the frequent item sets. In this example, the min support parameter is 0.7 (Figure 1.14), meaning that the operator generates a list of the frequent sets of words (itemsets) that appear in at least 70 % of the documents. Notice that, it will be computationally efficient to select the min support value in the FP-Growth operator to be equal to prune below percent value for the Process Documents from File operator (Figure 1.11) divided by 100. Also, the max items parameter is 2, meaning that the generated list is limited to pairs of words (2-itemsets), and the list will not contain frequent word sets (itemsets) with 3 or more words in them. The final operator in Process01, namely Create Association Rules, receives the list of frequent word sets from the FP-Growth operator, and computes the rules that satisfy the specified constraints on selected association mining criteria. In this example, the association rules are computed according to the the criterion of confidence, as well as gain theta and laplace k. The specified minimal values for these 3 criteria are 0.8, 1.0 and 1.0, respectively.

1.2.3 Saving Process01

So far Process01 has been imported in to the workspace of RapidMiner. Now it is a good time to keep it in the LocalRepository so that we will not have to import it again next time we work with RapidMiner. Click on the Repositories view, right-click on the LocalRepository and select Configure Repository as shown in Figure 1.15. This is actually an initialization step before using RapidMiner for the first time, but we will still go through this step to ensure that we are saving everything to a preferred folder directory in our computer. In this chapter, we will be saving everything under the Root directory of `C:\RapidMiner`, as in Figure 1.16. Next, we can save Process01 in to the LocalRepository. Right-click on the LocalRepository text and select Store Process Here, as in Figure 1.17. When the Store Process dialog window appears, click Ok. Our importing and saving of Process01 is now completed.

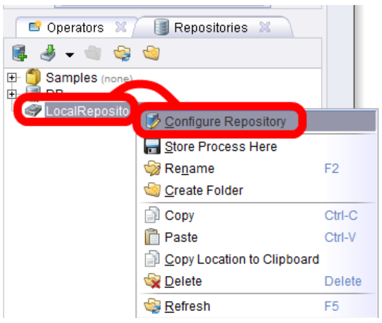
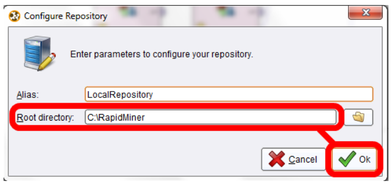


Figure 1.15 Configuring LocalRepository



Enter parameters to configure your repository.

Alias: LocalRepository

Root directory: C:\RapidMiner

Figure 1.16 Specifying the Root directory for LocalRepository

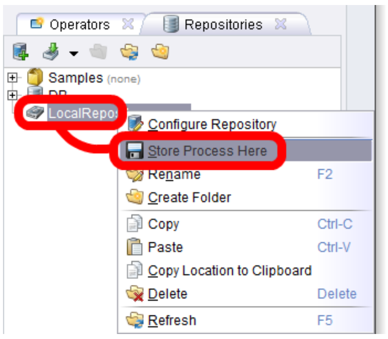


Figure 1.17 Storing Process01 in LocalRepository

1.3 CLUSTERING TEXT DOCUMENTS (Process02)

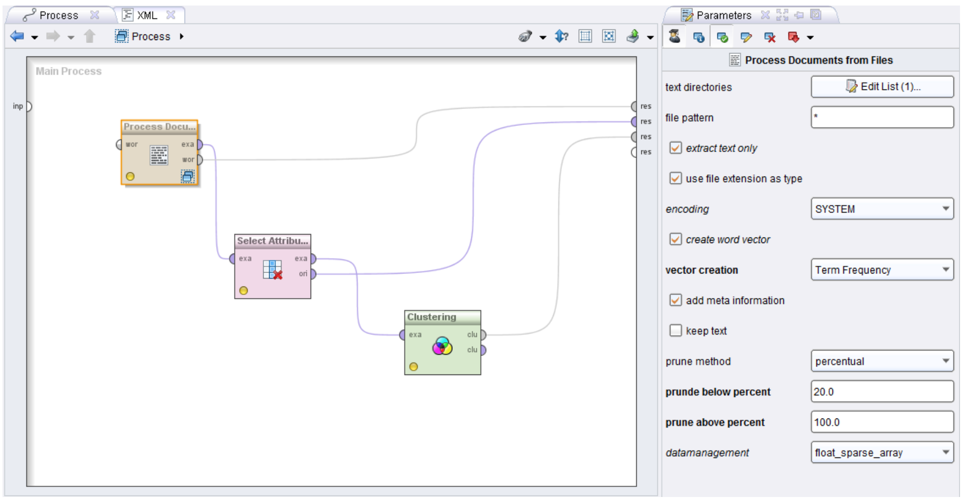
The second text mining process that we will introduce in this chapter is Process02, and involves the clustering of the 100 documents in the text collection.

1.3.1 Importing Process02

To import this process, select File menu, Import Process menu item, select `Process02.rmp` within the Import Process window, and click Open.

1.3.2 Operators in Process02

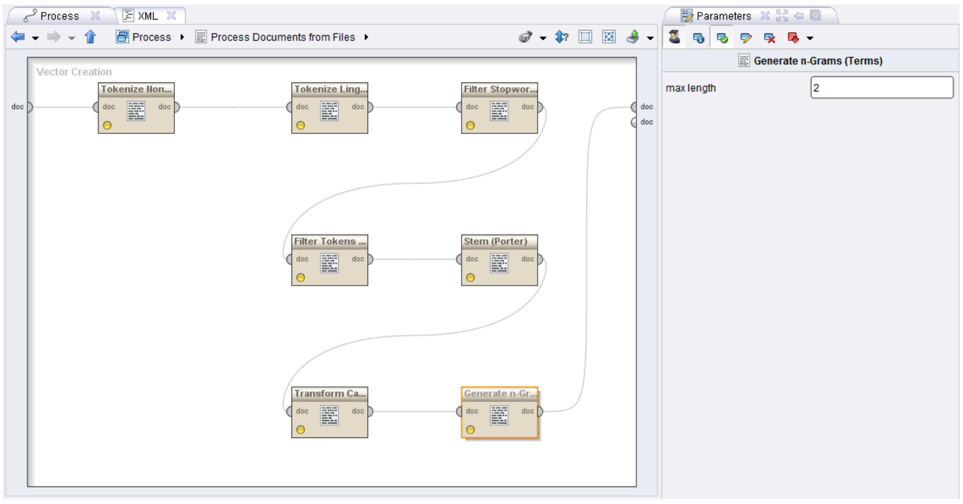
Process02 is given in Figure 1.18. Similar to Process01, Process02 begins with the Process Documents from Files operator, whose parameters are given in Figure 1.18. In Process02, the selected vector creation method is different; it is Term Frequency. The impact of this new selection will be illustrated later. For now, it should be noted that this selection results in the computation of the relative frequencies of each of the words in each of the documents in the data set. For example, if a word appears 5 times within a document that consist of 200 words, then the relative frequency of that word will be $5/200 = 0.025$. This value of 0.025 will appear under the column for that word, at the row for that document. In Process02, the prune method is again percentual, just as in Process01 (Figure 1.11). However, the value for the prune below percent parameter is now different.



Process Documents from Files

text directories	Edit List (1)...
file pattern	*
✓ extract text only	
✓ use file extension as type	
encoding	SYSTEM
✓ create word vector	
vector creation	Term Frequency
✓ add meta information	
☐ keep text	
prune method	percentual
prunde below percent	20.0
prune above percent	100.0
datamanagement	float_sparse_array

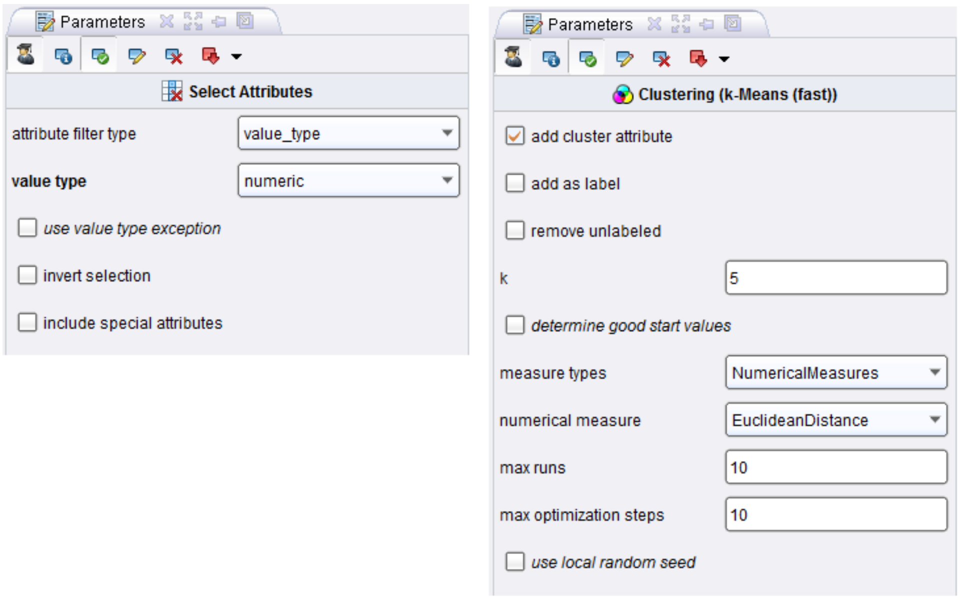
Figure 1.18 Operators in Process02 and the Parameters for Process Documents from Files operator



Generate n-Grams (Terms)

max length	2

Figure 1.19 Operators within the Process Documents from Files nested operator and the Parameters for the Generate n-Grams (Terms) operator



Select Attributes

attribute filter type	value_type
value type	numeric
☐ *use value type exception*	
☐ invert selection	
☐ include special attributes	

Clustering (k-Means (fast))

✓ add cluster attribute	
☐ add as label	
☐ remove unlabeled	
k	5
☐ *determine good start values*	
measure types	NumericalMeasures
numerical measure	EuclideanDistance
max runs	10
max optimization steps	10
☐ *use local random seed*	

Figure 1.20 Parameters for the operators in Process02

The prune below percent parameter in Process01 was 70.0, whereas it is now 20.0 in Process02. The reason for this change is due to the fact that the applied data mining technique in Process01 is association mining, whereas in Process02 it is clustering.

The former technique is computationally much more expensive then the latter, meaning that running association mining algorithms on a data set takes much longer running time compared to the k-Means clustering algorithm on the same data set. Therefore, with the same amount of time available for our data mining process, we have to work with much smaller data sets if we are carrying out association mining, as in Process01. The values displayed in Figures 1.11 and 1.18 have been determined through trial and error, such that the running time for the processes do not exceed 30 seconds on a laptop with Intel i5 processor and 4GB RAM.

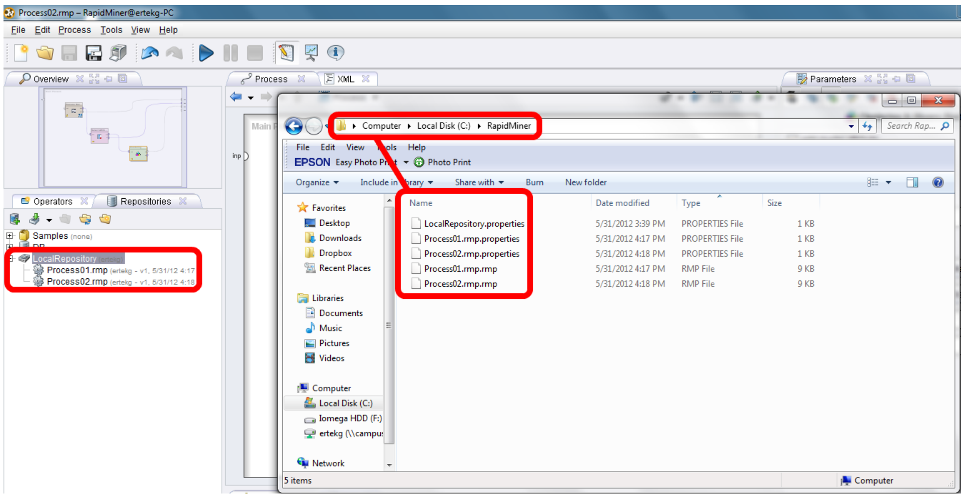
Figure 1.19 shows the contents of this operator. While there were six operators within this nested operator in Process01, there are now seven operators in Process02. The newly-added operator is Generate n-Grams (Terms). The only parameter for this operator is max length which is set equal to 2 in our example (Figure 1.19).

The parameters for the Select Attributes and Clustering (k-Means(fast)) operators within Process02 are displayed in Figure 1.20. The Select Attributes operator takes the complete data set and transforms it into a new one by selecting only the columns with numeric values, i.e. columns corresponding to the words. This transformation is required for the next operator, which performs clustering based on numerical values. Clustering (k-Means (fast)) operator carries out the k-Means clustering algorithm on the numerical data set. Since each row of the data (each document in the text collection) is characterized in terms of the occurrence frequency of words in it, this operator will place together the documents that have a similar distribution of the word frequencies. The k value, which denotes the number of clusters to be constructed, is set to 5 in our example.

1.3.3 Saving Process02

Now, let us save Process02 into LocalRepository: Click on Repositories view, right click on LocalRepository view and select Store Process Here. When the store process dialog box appears click Ok. As shown in Figure 1.21, both processes are now saved under local repository. Since we had earlier defined the directory root for local repository as `C:\RapidMiner`, the saved processes will appear as `.rpm` (RAPIDMINER process) files under that directory.

Unfortunately, there is a problem with the naming of the files: the file names also contain the extension .rpm, in addition to the `.rpm` extension itself. For example, the correct name for `Process01rmp.rmp` under `C:\RapidMiner` should be `Process01.rmp`. Right click on Process01.rmp text within RAPIDMINER, and select rename as in Figure 1.22 then, as in Figure 1.23 remove the .rmp extension from the file name and click OK. Do the same for renaming `Process02` and you will obtain the results in Figure 1.24.


Process02.rmp – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Overview
Operators
Repositories
Samples (none)
LocalRepository
Process01.rmp (ertekg - v1, 5/31/12 4:17
Process02.rmp (ertekg - v1, 5/31/12 4:18
Process
XML
Parameters
Computer ▸ Local Disk (C:) ▸ RapidMiner
File Edit View Tools Help
EPSON Easy Photo Print ▾ Photo Print
Organize ▾ Include in library ▾ Share with ▾ Burn New folder
Favorites
Desktop
Downloads
Dropbox
Recent Places
Libraries
Documents
Music
Pictures
Videos
Computer
Local Disk (C:)
Iomega HDD (F:)
ertekg (\\campus
Network
5 items
Computer

Name	Date modified	Type	Size
LocalRepository.properties	5/31/2012 3:39 PM	PROPERTIES File	1 KB
Process01.rmp.properties	5/31/2012 4:17 PM	PROPERTIES File	1 KB
Process02.rmp.properties	5/31/2012 4:18 PM	PROPERTIES File	1 KB
Process01.rmp.rmp	5/31/2012 4:17 PM	RMP File	9 KB
Process02.rmp.rmp	5/31/2012 4:18 PM	RMP File	9 KB

Figure 1.21 Saved processes within LocalRepository

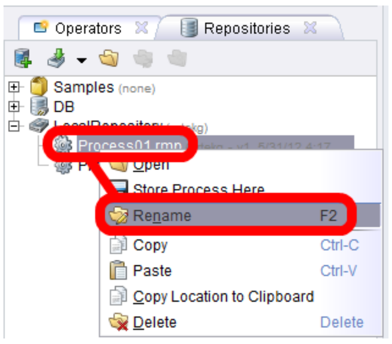
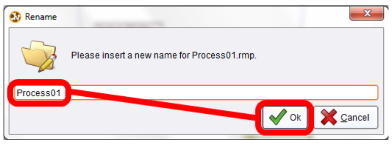


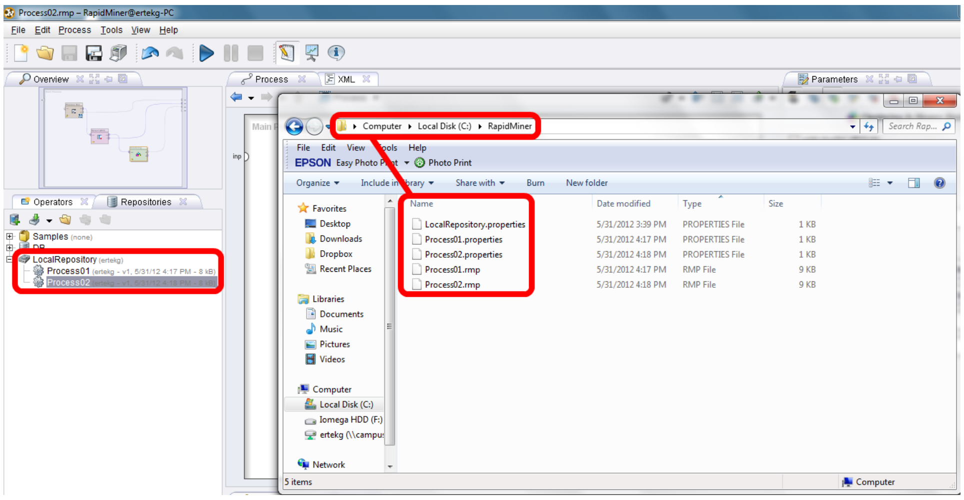
Figure 1.22 Renaming a process



Please insert a new name for Process01.rmp.

Process01

Figure 1.23 Renaming Process01 correctly



Name	Date modified	Type	Size
LocalRepository.properties	5/31/2012 3:39 PM	PROPERTIES File	1 KB
Process01.properties	5/31/2012 4:17 PM	PROPERTIES File	1 KB
Process02.properties	5/31/2012 4:18 PM	PROPERTIES File	1 KB
Process01.rmp	5/31/2012 4:17 PM	RMP File	9 KB
Process02.rmp	5/31/2012 4:18 PM	RMP File	9 KB

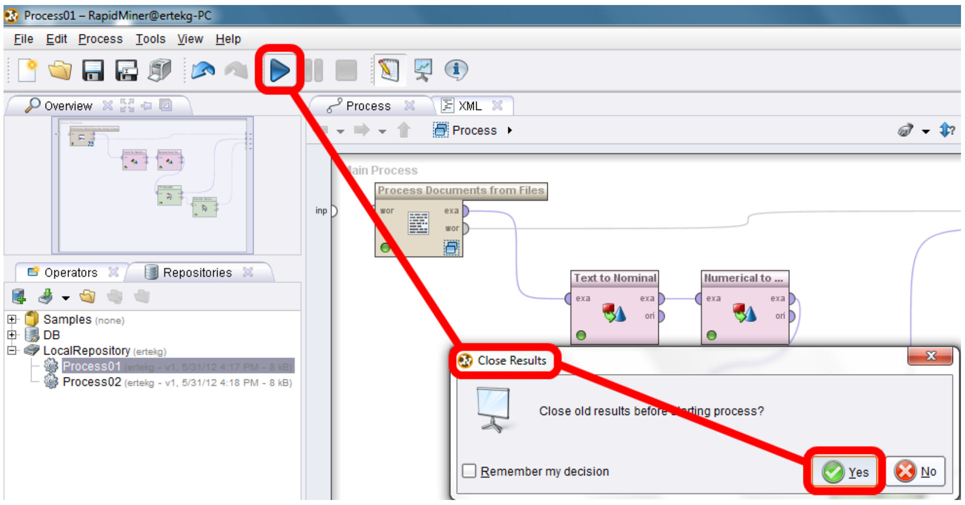
Figure 1.24 Corrected names for Process01 and Process02 in LocalRepository and the corresponding files

1.4 RUNNING Process01 AND ANALYZING THE RESULTS

Having imported and saved the processes into local repository, now we will run these processes.

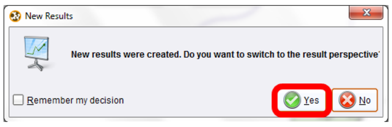
1.4.1 Running Process01

For running Process01, double click on Process01.rmp under the Repositories view. Process01 will be loaded into the Process view. Click the run button in the tool bar, as in Figure 1.25. You may be prompted and asked whether you would like to close all results before starting process (Figure 1.25). Click Yes. When the process run is completed, you will be prompted again and asked whether you would like to switch to the result perspective (Figure 1.26). Click Yes and you will have the screen in Figure 1.27. This is indeed the *result perspective*, displaying the result obtained through your process. Until now, you worked on your process in the *design perspective*, where we could change your process by adding/removing/editing operators. Now, in the result perspective, we are not able to modify our process. In the tool bar, we can switch to the design perspective or the result perspective by clicking their respective buttons.


Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Overview
Process
XML
Process
Main Process
Process Documents from Files
Text to Nominal
Numerical to ...
Operators
Repositories
Samples (none)
DB
LocalRepository (ertekg)
Process01 (ertekg - v1, 5/31/12 4:17 PM - 8 kB)
Process02 (ertekg - v1, 5/31/12 4:18 PM - 8 kB)
Close Results
Remember my decision
Yes
No

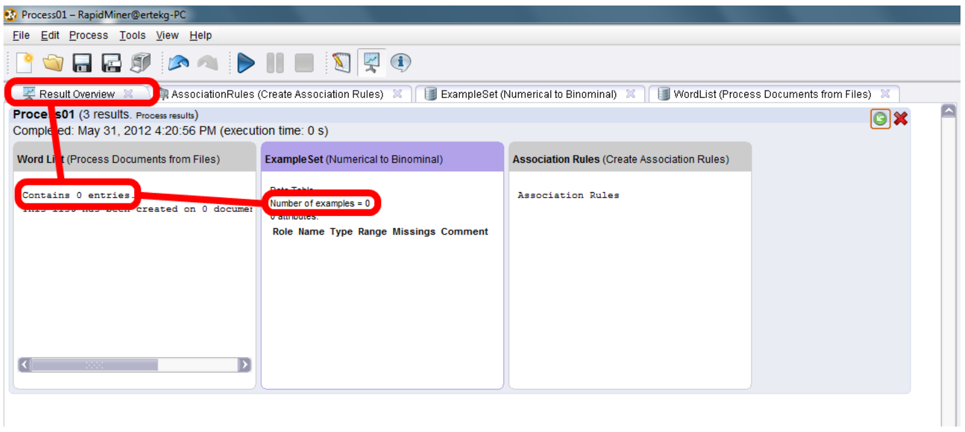
Close old results before starting process?

Figure 1.25 Opening and running Process01



New results were created. Do you want to switch to the result perspective?

Figure 1.26 Dialog box alerting the switch to the result perspective


Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)

Proce[illegible]s01 (3 results. Process results)
Comple[illegible]ed: May 31, 2012 4:20:56 PM (execution time: 0 s)

Word Li[illegible]t (Process Documents from Files)

```
Contains 0 entries.
[illegible] created on 0 documen
```

ExampleSet (Numerical to Binominal)

Data Table
Number of examples = 0
0 attributes:

Role Name Type Range Missings Comment

Association Rules (Create Association Rules)

```
Association Rules
```

Figure 1.27 Result Overview for Process01 results

1.4.2 Empty Results for Process01

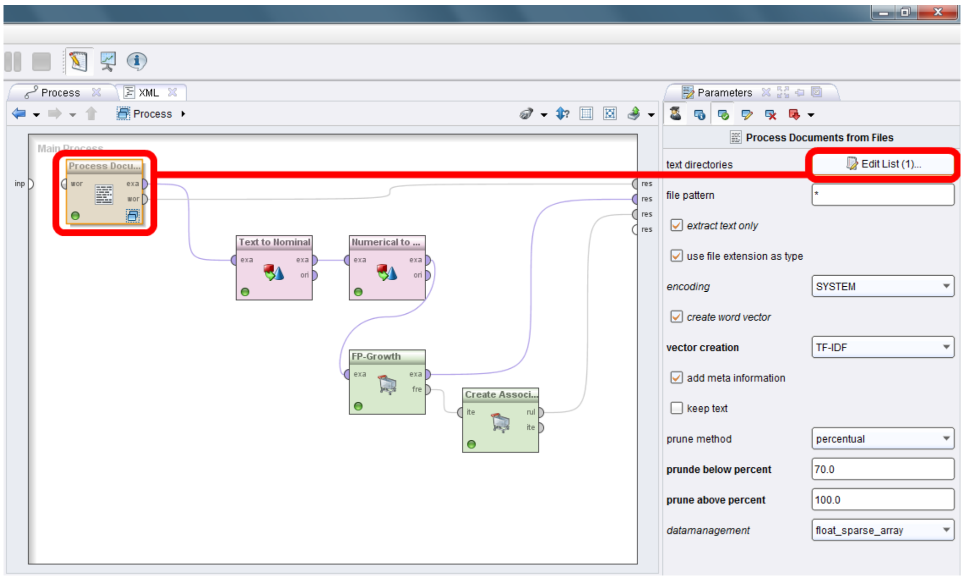
In result perspective, we can see the latest results at the bottom of the result overview. To see the details click inside the bar with the most recent date and time and you will see an expanded view as in Figure 1.27. In the extended view you will note that there are three types of results obtained as a result of running Proces01. These three results correspond precisely to the three result ports for Process01 which are represented by the three half-circles on the right of the process view in Figure 1.11 and are labeled with res. The first result is the wordlist (wor) arising from the Process Documents From Files operator (Figure 1.11). The second result is the example set exa, which arises from the FP-Growth operator. The third and final result is the list of association rules (rul) which arises from the Create Association Rules.

In Figure 1.27 the word list contains 0 entries and the number of examples is 0. But why? Was there something wrong with our processes? We will answer this question next.

1.4.3 Specifying the Source Data for Process01

Running Process01 resulted in no tangible results and the reason is very simple. We have not yet specified the text data set that will be processed and mined in Process01. To specify the data source, as in Figure 1.28 switch to the design perspective by clicking the notepad button, click on the Process Documents from Files operator and click on the box to the right of the parameter text directories.

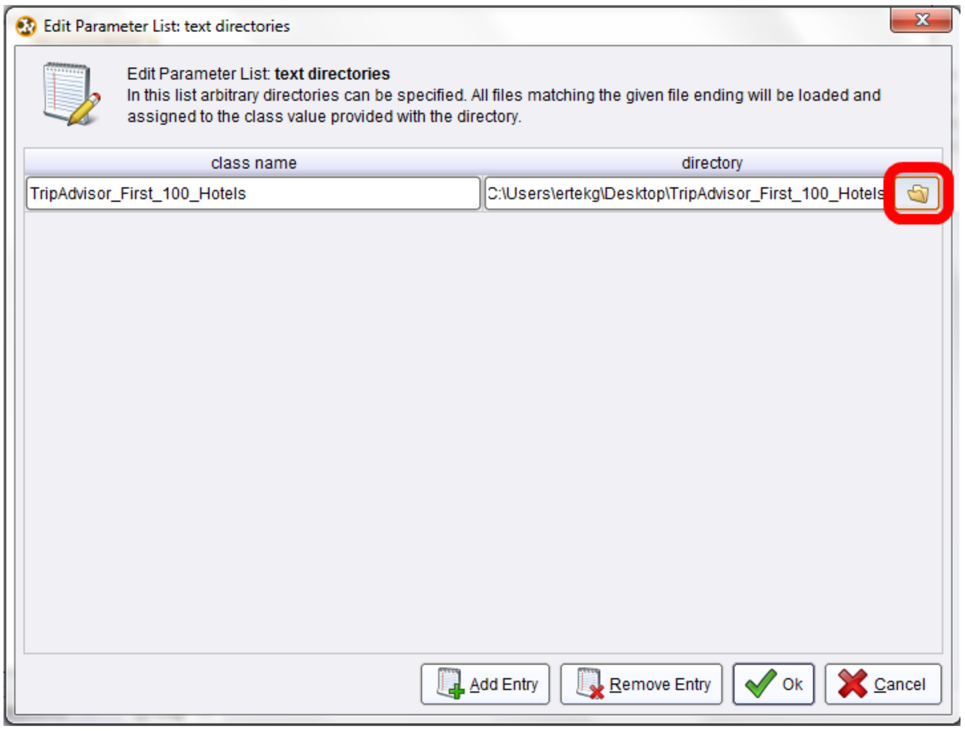
In the dialog box click on the directory button, (see Figure 1.29) select the source data folder, double click TripAdvisor_First_100_Hotels folder, and click the open button (Figure 1.30). Now the directory for the source text data appears in the dialog box, as in Figure 1.31. Now click Ok.



Process Documents from Files

text directories	Edit List (1)...
file pattern	*
✓ *extract text only*	
✓ use file extension as type	
encoding	SYSTEM
✓ *create word vector*	
vector creation	TF-IDF
✓ add meta information	
☐ keep text	
prune method	percentual
prunde below percent	70.0
prune above percent	100.0
datamanagement	float_sparse_array

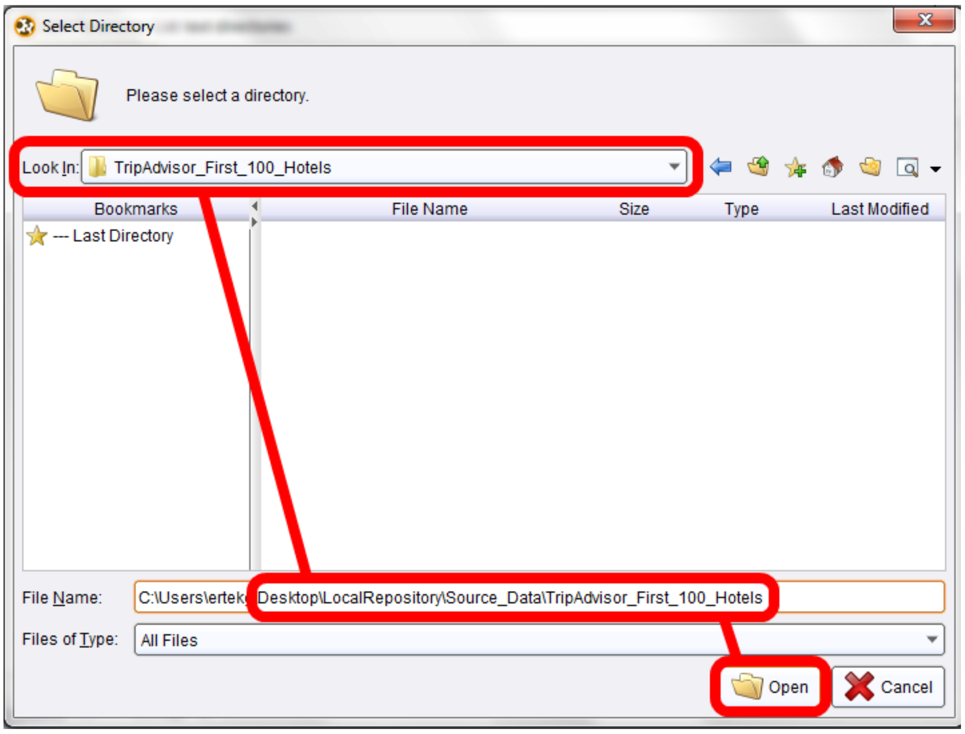
Figure 1.28 Specifying the data source text directories for Process01


Edit Parameter List: text directories

Edit Parameter List: **text directories**
In this list arbitrary directories can be specified. All files matching the given file ending will be loaded and assigned to the class value provided with the directory.

class name	directory
TripAdvisor_First_100_Hotels	C:\Users\ertekg\Desktop\TripAdvisor_First_100_Hotels

Figure 1.29 Specifying the text directories

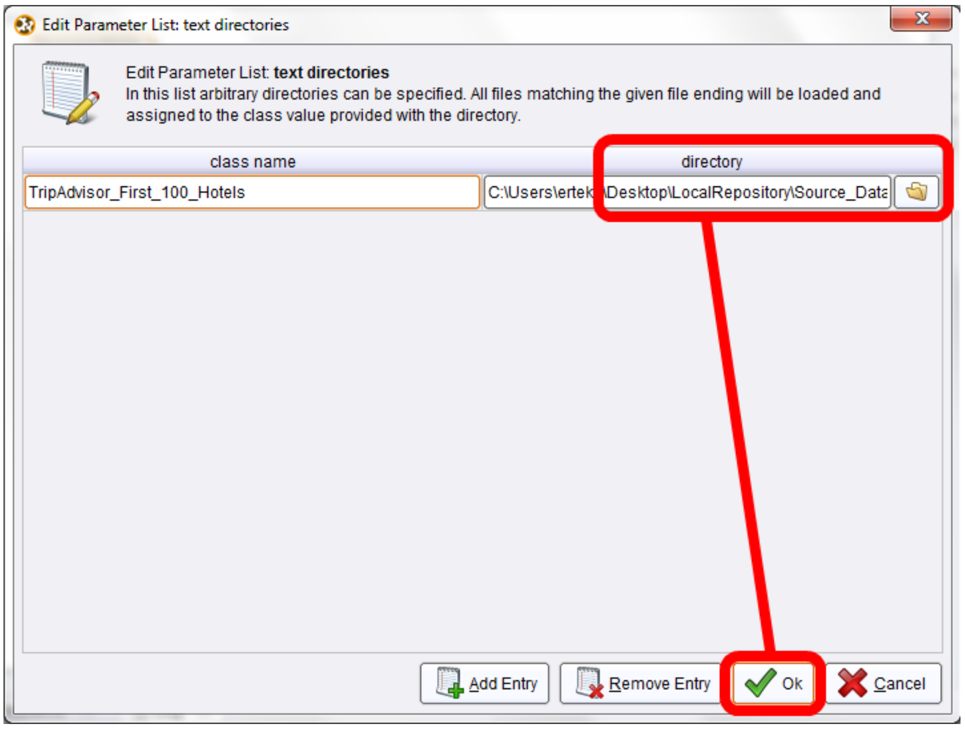


Look In: TripAdvisor_First_100_Hotels

File Name: C:\Users\ertek\Desktop\LocalRepository\Source_Data\TripAdvisor_First_100_Hotels

Files of Type: All Files

Figure 1.30 Specifying the text directories


Edit Parameter List: text directories

Edit Parameter List: **text directories**
In this list arbitrary directories can be specified. All files matching the given file ending will be loaded and assigned to the class value provided with the directory.

class name	directory
TripAdvisor_First_100_Hotels	C:\Users\ertek\Desktop\LocalRepository\Source_Data

Figure 1.31 Specified text directories

1.4.4 Re-Running Process01

Having specified the data source, we can run Process01 again. Click on the run button (see Figure 1.32) and click Yes when you are asked. You should click Save Process Before Start?, and click Yes when you are asked whether to Close old results before starting process. When the process is running you can observe which operator is currently being used. For example, in Figure 1.33 the FP-Growth algorithm is running, as can be understood from the small green triangle on its lower left corner. From Figure 1.33 we can also read (from the lower left corner of the RAPIDMINER window) that Process01 has been running for 14 seconds (14 s) and the FP-Growth operator has been running for 7 seconds (7 s).

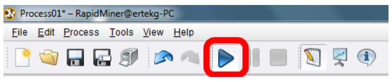

Process01* – RapidMiner@ertekg-PC
File Edit Process Tools View Help

Figure 1.32 Running Process01 again

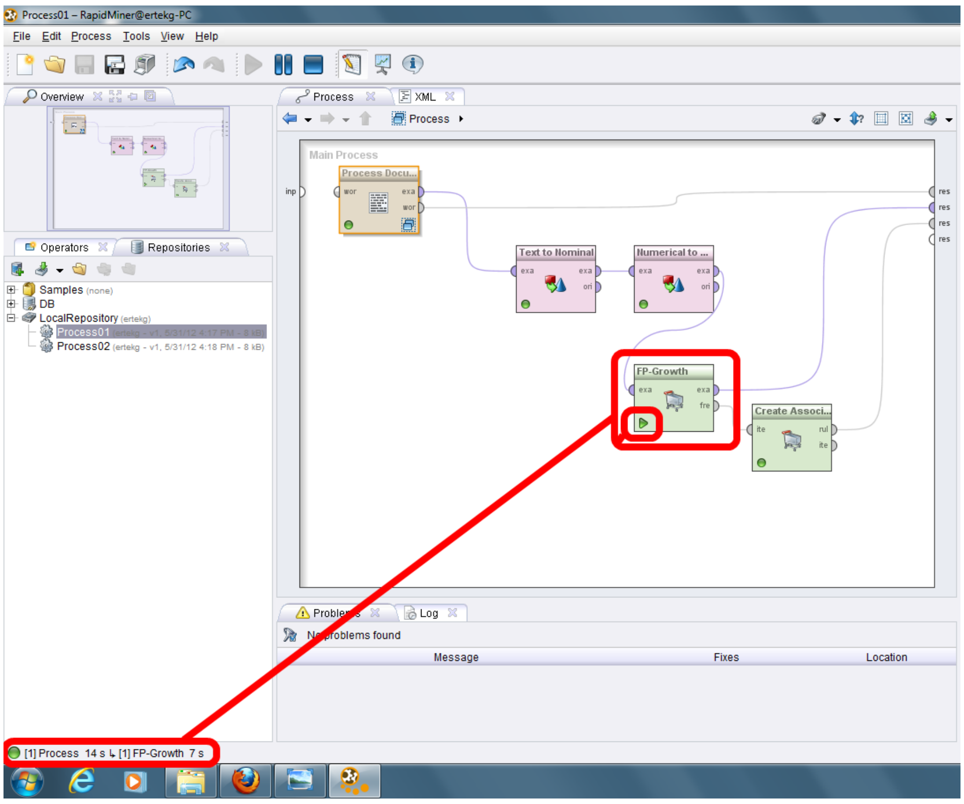
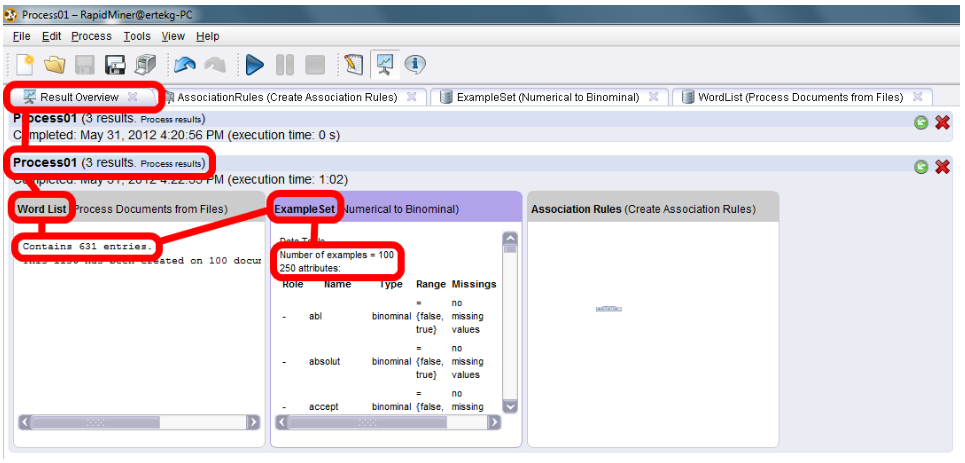


Figure 1.33 Running of Process01

1.4.5 Process01 Results

When the new results are created, click Yes to the result perspective. In the result perspective you will now see a see a blue bar. Click inside the bottommost blue bar and you will see an overview of the results, as in Figure 1.34. This time the world list is not empty, it Contains 631 entries. The Number of examples in the example set is equal to 100, and there are 250 attributes (Figure 1.34).



Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)

Process01 (3 results. Process results)
Completed: May 31, 2012 4:20:56 PM (execution time: 0 s)

Process01 (3 results. Process results)
(execution time: 1:02)

Word List (Process Documents from Files)

```
Contains 631 entries.
                    ated on 100 docur
```

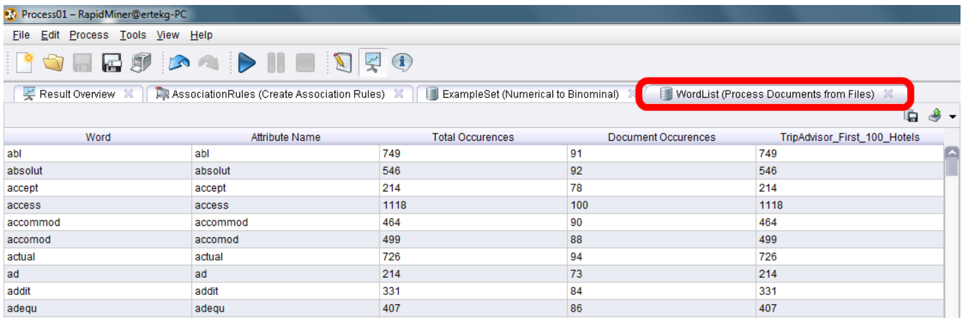
ExampleSet (Numerical to Binominal)

Number of examples = 100
250 attributes:

Role	Name	Type	Range	Missings
-	abl	binominal	= {false, true}	no missing values
-	absolut	binominal	= {false, true}	no missing values
-	accept	binominal	= {false,	no missing

Association Rules (Create Association Rules)

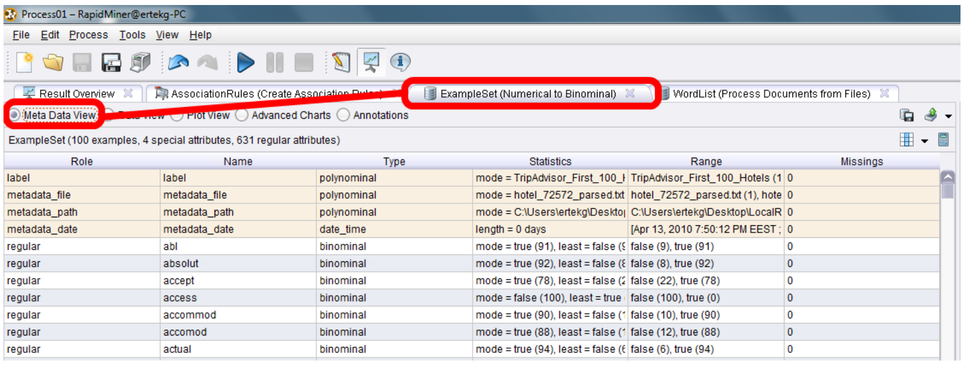
Figure 1.34 Result Overview for Process01 results



Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)

Word	Attribute Name	Total Occurences	Document Occurences	TripAdvisor_First_100_Hotels
abl	abl	749	91	749
absolut	absolut	546	92	546
accept	accept	214	78	214
access	access	1118	100	1118
accommod	accommod	464	90	464
accomod	accomod	499	88	499
actual	actual	726	94	726
ad	ad	214	73	214
addit	addit	331	84	331
adequ	adequ	407	86	407

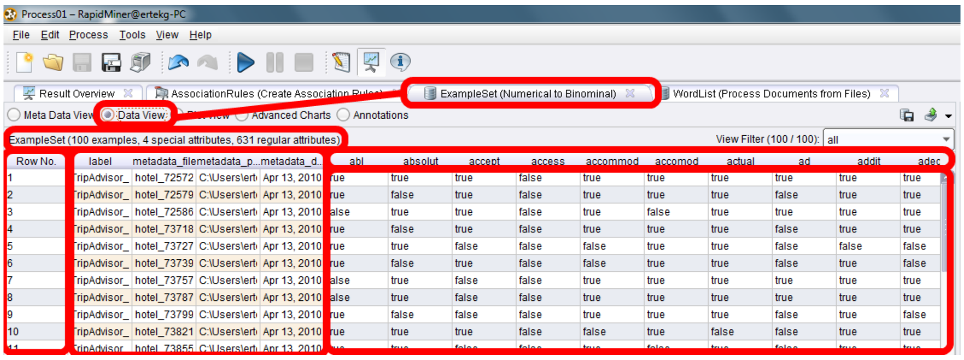
Figure 1.35 WordList generated by Process01


Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)
Meta Data View
Data View
Plot View
Advanced Charts
Annotations

ExampleSet (100 examples, 4 special attributes, 631 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	label	polynominal	mode = TripAdvisor_First_100_I	TripAdvisor_First_100_Hotels (1	0
metadata_file	metadata_file	polynominal	mode = hotel_72572_parsed.txt	hotel_72572_parsed.txt (1), hote	0
metadata_path	metadata_path	polynominal	mode = C:\Users\ertekg\Deskto	C:\Users\ertekg\Desktop\LocalR	0
metadata_date	metadata_date	date_time	length = 0 days	[Apr 13, 2010 7:50:12 PM EEST ;	0
regular	abl	binominal	mode = true (91), least = false (9	false (9), true (91)	0
regular	absolut	binominal	mode = true (92), least = false (8	false (8), true (92)	0
regular	accept	binominal	mode = true (78), least = false (2	false (22), true (78)	0
regular	access	binominal	mode = false (100), least = true	false (100), true (0)	0
regular	accommod	binominal	mode = true (90), least = false (1	false (10), true (90)	0
regular	accomod	binominal	mode = true (88), least = false (1	false (12), true (88)	0
regular	actual	binominal	mode = true (94), least = false (6	false (6), true (94)	0

Figure 1.36 Meta Data View for the ExampleSet generated by Process01


Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)
Meta Data View
Data View
Plot View
Advanced Charts
Annotations

ExampleSet (100 examples, 4 special attributes, 631 regular attributes)

View Filter (100 / 100): all

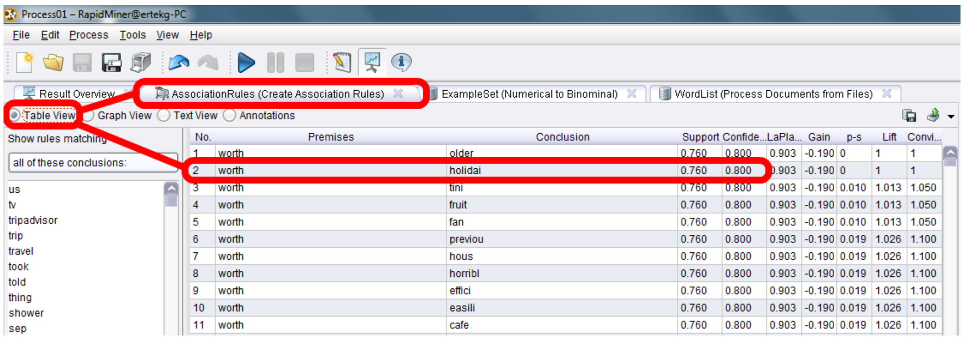
Row No.	label	metadata_file	metadata_p...	metadata_d...	abl	absolut	accept	access	accommod	accomod	actual	ad	addit	adec
1	TripAdvisor_	hotel_72572	C:\Users\ert	Apr 13, 2010	rue	true	true	false	true	true	true	true	true	true
2	TripAdvisor_	hotel_72579	C:\Users\ert	Apr 13, 2010	rue	false	true	false	true	true	true	false	true	true
3	TripAdvisor_	hotel_72586	C:\Users\ert	Apr 13, 2010	alse	true	true	false	true	false	true	true	true	true
4	TripAdvisor_	hotel_73718	C:\Users\ert	Apr 13, 2010	rue	false	true	false	true	true	true	false	true	true
5	TripAdvisor_	hotel_73727	C:\Users\ert	Apr 13, 2010	rue	true	false	false	false	true	true	false	false	false
6	TripAdvisor_	hotel_73739	C:\Users\ert	Apr 13, 2010	rue	false	true	false	false	true	true	false	true	false
7	TripAdvisor_	hotel_73757	C:\Users\ert	Apr 13, 2010	alse	true	false	false	true	true	true	true	true	true
8	TripAdvisor_	hotel_73787	C:\Users\ert	Apr 13, 2010	alse	true	false	false	true	true	true	true	true	true
9	TripAdvisor_	hotel_73799	C:\Users\ert	Apr 13, 2010	rue	false	false	false	true	true	true	false	true	false
10	TripAdvisor_	hotel_73821	C:\Users\ert	Apr 13, 2010	rue	true	true	false	false	true	false	false	true	true
11	TripAdvisor	hotel_73855	C:\Users\ert	Apr 13, 2010	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Figure 1.37 Data View for the ExampleSet generated by Process01

Click on the world list view to display the words found by the Process Documents from Files operator, as in Figure 1.35.

Click on the ExampleSet (Numerical to Binomial) view to display the example set. Figure 1.36 presents the Meta Data View, which gives the metadata (information on the data attributes) of the ExampleSet. Now click on the Data View button to display the data itself (Figure 1.37). Each row corresponds to a text document and the first column is an automatically generated key attribute. The extra four columns contain the label, file name, file location and the file creation data. The remaining columns correspond to the world list. The cells under the word attributes take the binominal value of true or false, denoting whether the word exists in that document.

The final result is the set of association rules. Click on the AssociationRules(Create Association Rules) tab to display the association rules (Figure 1.38). In the Table View, table grid presents the generated association rules, with one rule in each row. For example, the second row states “IF worth THEN holidai” with a Support level of 0.760 and Confidence level of 0.800. This rule means that in 76 of the 100 documents, words with stem worth and holidai appear together. Furthermore in 80% of the documents where a word derived from the stem worth appears, at least one word derived from the stem holidai is observed. On the left hand side of the grid we can select and Show rules matching a particular set of words (Figure 1.38).


Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)
[✓ Table View] Graph View Text View Annotations
Show rules matching
all of these conclusions:
us
tv
tripadvisor
trip
travel
took
told
thing
shower
sep

No.	Premises	Conclusion	Support	Confide...	LaPla...	Gain	p-s	Lift	Convi...
1	worth	older	0.760	0.800	0.903	-0.190	0	1	1
2	worth	holidai	0.760	0.800	0.903	-0.190	0	1	1
3	worth	tini	0.760	0.800	0.903	-0.190	0.010	1.013	1.050
4	worth	fruit	0.760	0.800	0.903	-0.190	0.010	1.013	1.050
5	worth	fan	0.760	0.800	0.903	-0.190	0.010	1.013	1.050
6	worth	previou	0.760	0.800	0.903	-0.190	0.019	1.026	1.100
7	worth	hous	0.760	0.800	0.903	-0.190	0.019	1.026	1.100
8	worth	horribl	0.760	0.800	0.903	-0.190	0.019	1.026	1.100
9	worth	effici	0.760	0.800	0.903	-0.190	0.019	1.026	1.100
10	worth	easili	0.760	0.800	0.903	-0.190	0.019	1.026	1.100
11	worth	cafe	0.760	0.800	0.903	-0.190	0.019	1.026	1.100

Figure 1.38 Table View for the AssociationRules generated by Process01

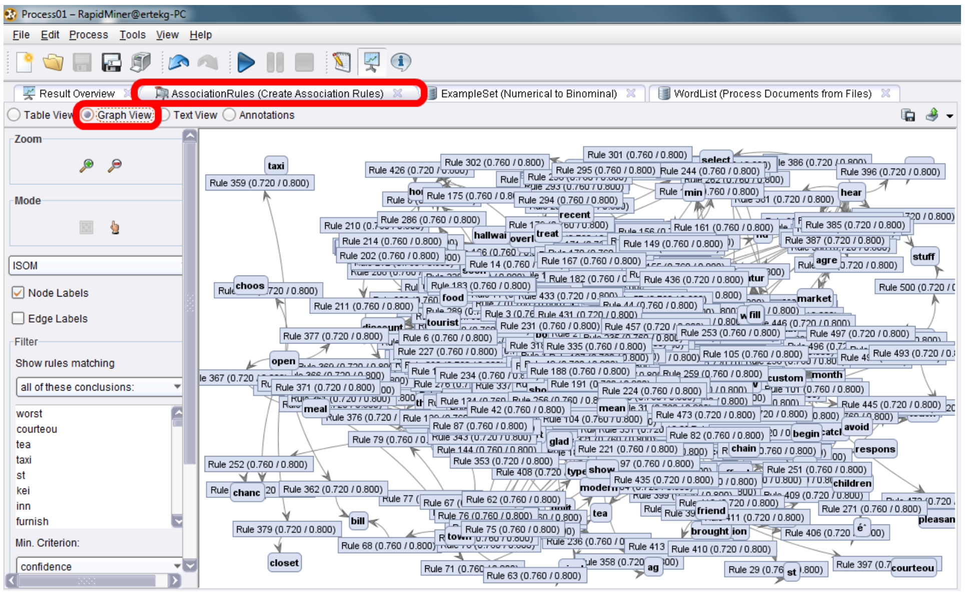

Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)
Table View
Graph View
Text View
Annotations
Zoom
Mode
ISOM
Node Labels
Edge Labels
Filter
Show rules matching
all of these conclusions:
worst
courteou
tea
taxi
st
kei
inn
furnish
Min. Criterion:
confidence

Figure 1.39 Graph View for the AssociationRules generated by Process01

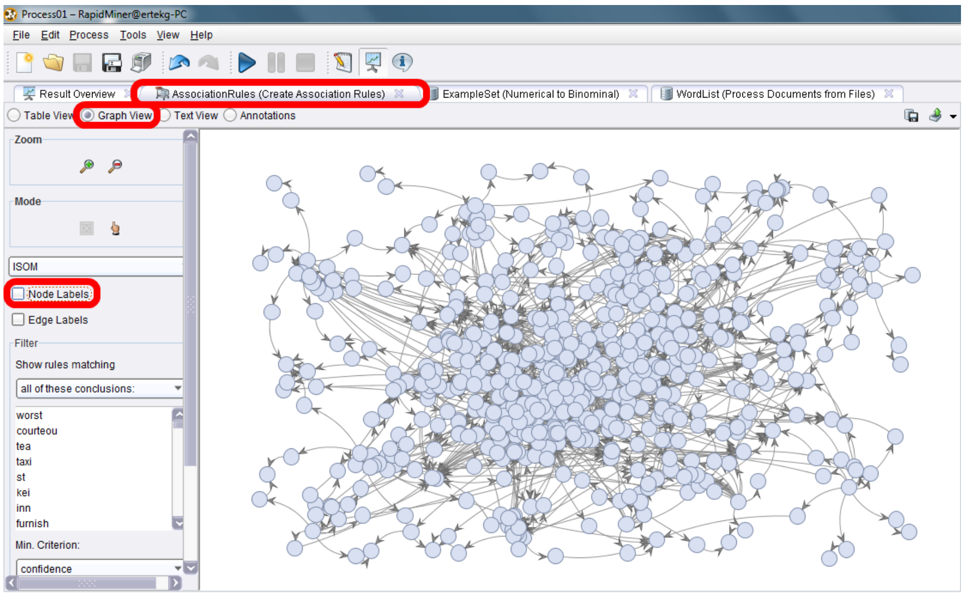

Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)
Table View
Graph View
Text View
Annotations
Zoom
Mode
ISOM
Node Labels
Edge Labels
Filter
Show rules matching
all of these conclusions:
worst
courteou
tea
taxi
st
kei
inn
furnish
Min. Criterion:
confidence

Figure 1.40 Graph View for the AssociationRules, without the node labels

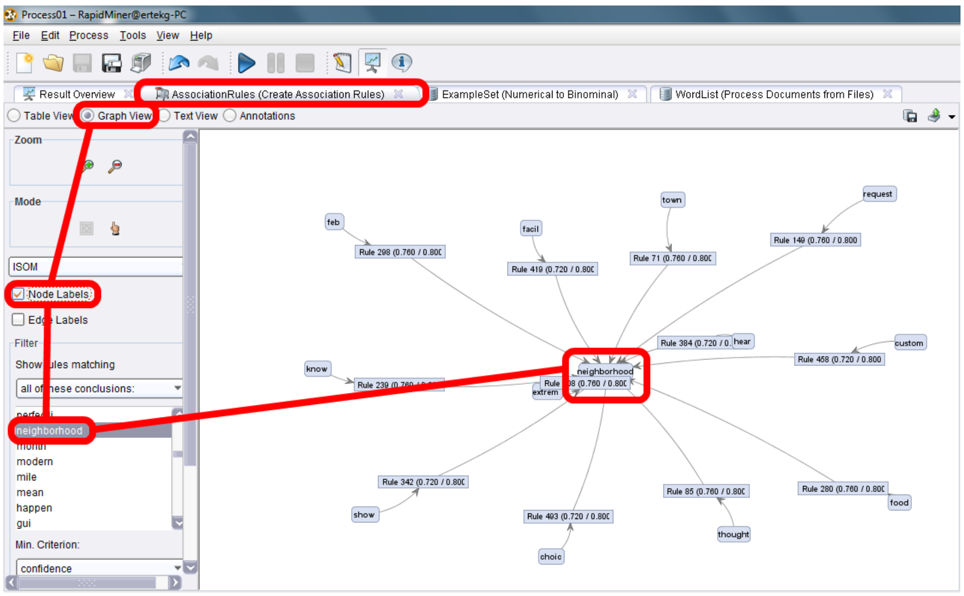
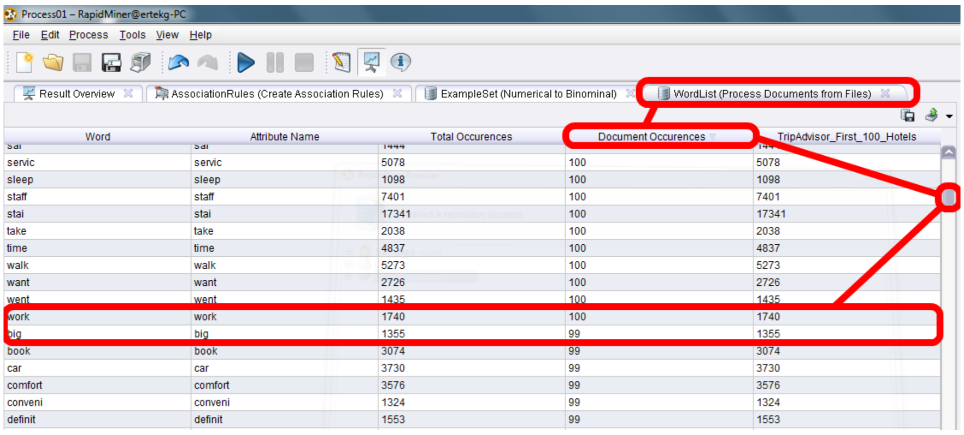


Figure 1.41 Filtering rules in the Graph View for the AssociationRules

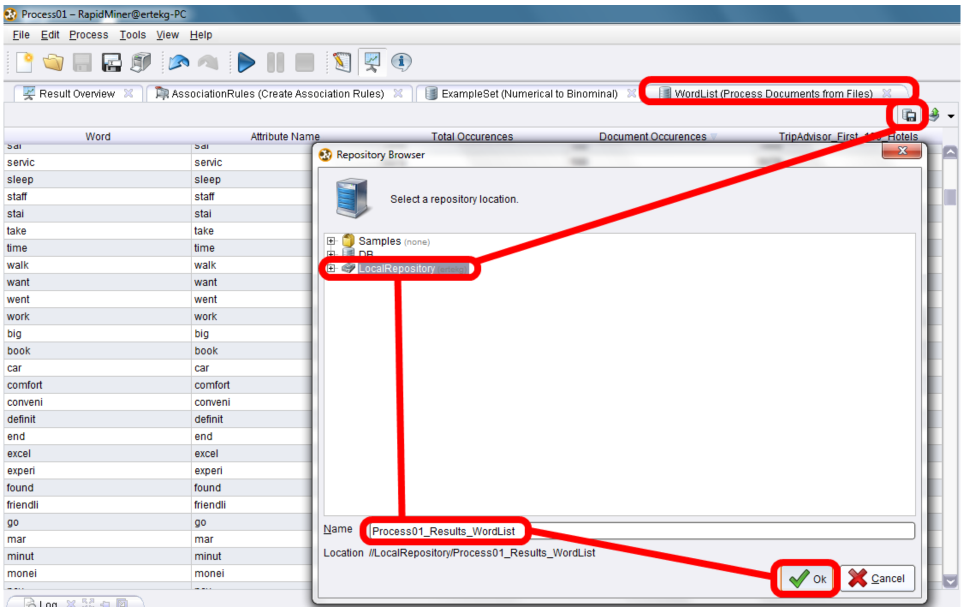
These association rules can also be displayed in the form of a graph. For this, click on the Graph View radio button as in Figure 1.39. The initial graph visualization may be cluttered due to the Node Labels. As Click on the check box Node Labels to uncheck that option and eliminate or reduce the clutter in (Figure 1.40). In this window, filtering can again be carried out by selecting words from the list box on the left hand side. Once a filtered word is selected you will want to again display the Node Labels as in Figure 1.41. At this point, select a word from the list box and check the Node Label check box (Figure 1.41).

The final interesting result that we would like to share is displayed in Figure 1.42. Click on WordList(Process Documents from Files) tab to return back to the word list. Click on the attribute label Document Occurences to obtain a descending order. Then move the vertical scroll bar to see the border line where the value changes from 100 to 99. From Figure 1.42 we can observe that the word work appears in all 100 documents, whereas the word big appears in 99 documents.


Process01 – RapidMiner@ertekg-PC
File
Edit
Process
Tools
View
Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)

Word	Attribute Name	Total Occurences	Document Occurences	TripAdvisor_First_100_Hotels
sai	sai	1444		144
servic	servic	5078	100	5078
sleep	sleep	1098	100	1098
staff	staff	7401	100	7401
stai	stai	17341	100	17341
take	take	2038	100	2038
time	time	4837	100	4837
walk	walk	5273	100	5273
want	want	2726	100	2726
went	went	1435	100	1435
work	work	1740	100	1740
big	big	1355	99	1355
book	book	3074	99	3074
car	car	3730	99	3730
comfort	comfort	3576	99	3576
conveni	conveni	1324	99	1324
definit	definit	1553	99	1553

Figure 1.42 Document Occurences of the words in the WordList

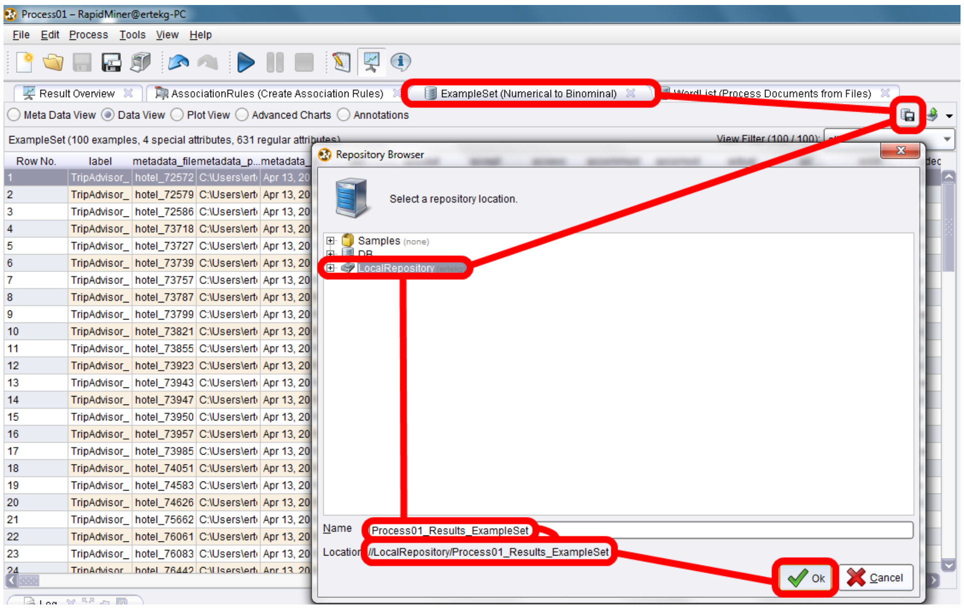

Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)
Repository Browser
Select a repository location.
Samples (none)
DB
LocalRepository
Ok
Cancel
Log

Word	Attribute Name	Total Occurences	Document Occurences	TripAdvisor_First_100_Hotels
servic	servic			
sleep	sleep			
staff	staff			
stai	stai			
take	take			
time	time			
walk	walk			
want	want			
went	went			
work	work			
big	big			
book	book			
car	car			
comfort	comfort			
conveni	conveni			
definit	definit			
end	end			
excel	excel			
experi	experi			
found	found			
friendli	friendli			
go	go			
mar	mar			
minut	minut			
monei	monei			

Name Process01_Results_WordList

Location //LocalRepository/Process01_Results_WordList

Figure 1.43 Saving the WordList for Process01


Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)
Meta Data View
Data View
Plot View
Advanced Charts
Annotations
Repository Browser
Select a repository location.
Samples (none)
DB
LocalRepository
Ok
Cancel
Log

ExampleSet (100 examples, 4 special attributes, 631 regular attributes)

View Filter (100 / 100):

Row No.	label	metadata_file	metadata_p...	metadata_
1	TripAdvisor_	hotel_72572	C:\Users\ert	Apr 13, 20
2	TripAdvisor_	hotel_72579	C:\Users\ert	Apr 13, 20
3	TripAdvisor_	hotel_72586	C:\Users\ert	Apr 13, 20
4	TripAdvisor_	hotel_73718	C:\Users\ert	Apr 13, 20
5	TripAdvisor_	hotel_73727	C:\Users\ert	Apr 13, 20
6	TripAdvisor_	hotel_73739	C:\Users\ert	Apr 13, 20
7	TripAdvisor_	hotel_73757	C:\Users\ert	Apr 13, 20
8	TripAdvisor_	hotel_73787	C:\Users\ert	Apr 13, 20
9	TripAdvisor_	hotel_73799	C:\Users\ert	Apr 13, 20
10	TripAdvisor_	hotel_73821	C:\Users\ert	Apr 13, 20
11	TripAdvisor_	hotel_73855	C:\Users\ert	Apr 13, 20
12	TripAdvisor_	hotel_73923	C:\Users\ert	Apr 13, 20
13	TripAdvisor_	hotel_73943	C:\Users\ert	Apr 13, 20
14	TripAdvisor_	hotel_73947	C:\Users\ert	Apr 13, 20
15	TripAdvisor_	hotel_73950	C:\Users\ert	Apr 13, 20
16	TripAdvisor_	hotel_73957	C:\Users\ert	Apr 13, 20
17	TripAdvisor_	hotel_73985	C:\Users\ert	Apr 13, 20
18	TripAdvisor_	hotel_74051	C:\Users\ert	Apr 13, 20
19	TripAdvisor_	hotel_74583	C:\Users\ert	Apr 13, 20
20	TripAdvisor_	hotel_74626	C:\Users\ert	Apr 13, 20
21	TripAdvisor_	hotel_75662	C:\Users\ert	Apr 13, 20
22	TripAdvisor_	hotel_76061	C:\Users\ert	Apr 13, 20
23	TripAdvisor_	hotel_76083	C:\Users\ert	Apr 13, 20
24	TripAdvisor_	hotel_76442	C:\Users\ert	Apr 13, 20

Name Process01_Results_ExampleSet

Location //LocalRepository/Process01_Results_ExampleSet

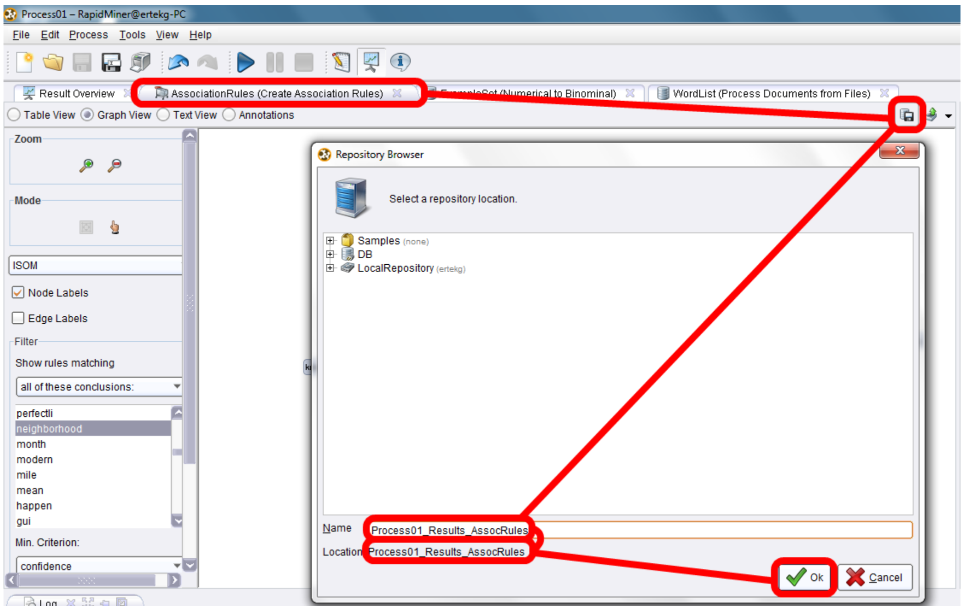
Figure 1.44 Saving the ExampleSet for Process01

1.4.6 Saving Process01 Results

Having generated results of the Process01 we will save these results into LocalRepository, so that we can save time when we want to analyze the results next time. The three types of results, namely the Word List, ExampleSet, AssociationRules have to be saved individually. First, click on the save button in the upper right corner of the results, click on LocalRepository in the appearing window, and type in an appropriate name inside the Name text box as in Figure 1.43, then click Ok.

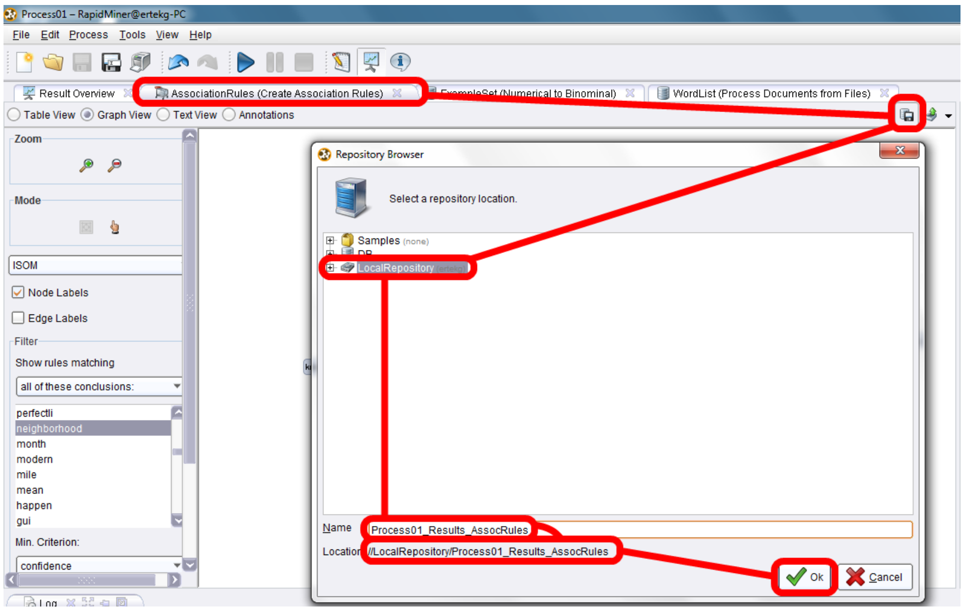
Next, click on the ExampleSet (Numerical to Binominal) tab, and repeat the same steps, as in Figure 1.44.

Finally, click on the AssociationRules, as in Figure 1.45. For this last result we would like to show you a common possible mistake. After clicking the save button, omit clicking on LocalRepository and click Ok as Figure 1.45. Please note that in this case we did not specify a location for the generated results. This will cause an error . When you see the error dialog box , click Close. Then, follow the correct steps; Click on the save button, click on LocalRepository, give an appropriate unique name and click Ok as in Figure 1.46.



Name: Process01_Results_AssocRules
Location: Process01_Results_AssocRules

Figure 1.45 Saving the AssociationRules incorrectly, without selecting LocalRepository or another directory



Name: Process01_Results_AssocRules
Location: //LocalRepository/Process01_Results_AssocRules

Figure 1.46 Saving the AssociationRules correctly, selecting the directory to be saved into

Now that all the data for the generated results are saved let us also see how we can save (export) the data visualizations. Click on the export button to the right of save button, as in Figure 1.47. Then specify the directory where the visualization will be saved (exported) and select the image file type as in Figure 1.48. In this example, select the .png image format, and then click Ok. You can check that the visualization is now saved as an image file, as specified.

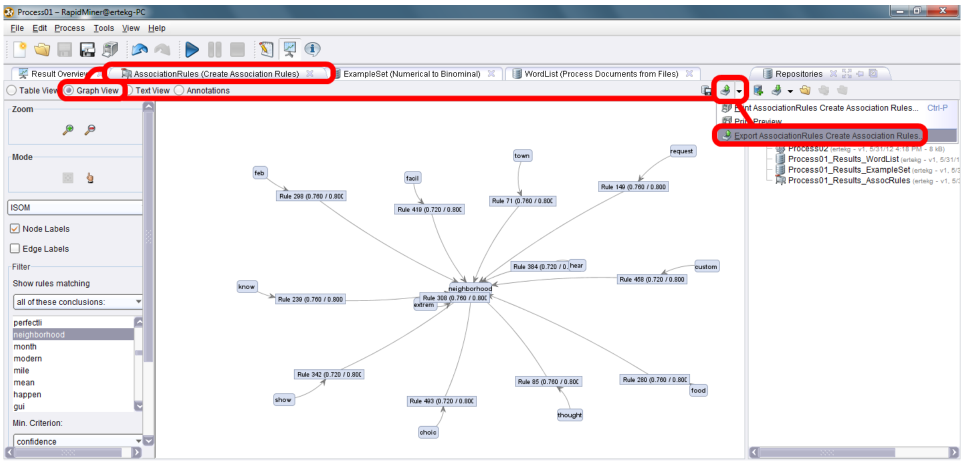
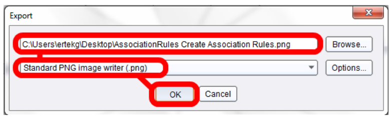

Process01 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
WordList (Process Documents from Files)
Repositories
Table View
Graph View
Text View
Annotations
Print AssociationRules Create Association Rules... Ctrl-P
Print Preview
Export AssociationRules Create Association Rules
Process02
Process01_Results_WordList
Process01_Results_ExampleSet
Process01_Results_AssocRules
Zoom
Mode
ISOM
Node Labels
Edge Labels
Filter
Show rules matching
all of these conclusions:
perfectli
neighborhood
month
modern
mile
mean
happen
gui
Min. Criterion:
confidence
neighborhood
feb
facil
town
request
know
custom
hear
extrem
show
choic
thought
food

Figure 1.47 Exporting the graph visualization as an image file



C:\Users\ertekg\Desktop\AssociationRules Create Association Rules.png

Standard PNG image writer (.png)

Figure 1.48 Specifying the directory to be saved into, the file name, and the file type

1.5 RUNNING Process02 AND ANALYZING THE RESULTS

In this section we will run Process02 and analyze the generated results.

1.5.1 Running Process02

To switch to the design view click on the design view button (Notepad icon) as in Figure 1.49. Then double click on Process02 inside the Repositories view, and click the run button as in Figure 1.50. You will be prompted with an error dialog box, as in Figure 1.51.

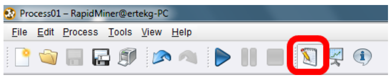


Figure 1.49 Switching to the design view

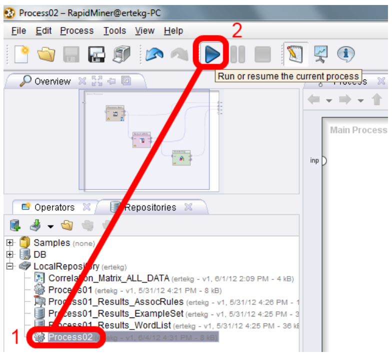
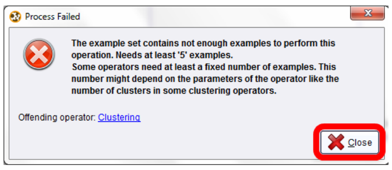

Process02 – RapidMiner@ertekg-PC
File
Edit
Process
Tools
View
Help
2
Run or resume the current process
Overview
Operators
Repositories
Main Process
inp
Samples (none)
DB
LocalRepository (ertekg)
Correlation_Matrix_ALL_DATA
Process01
Process01_Results_AssocRules
Process01_Results_ExampleSet
Process01_Results_WordList
1
Process02

Figure 1.50 Opening and running Process02

1.5.2 Specifying the Source Data for Process02

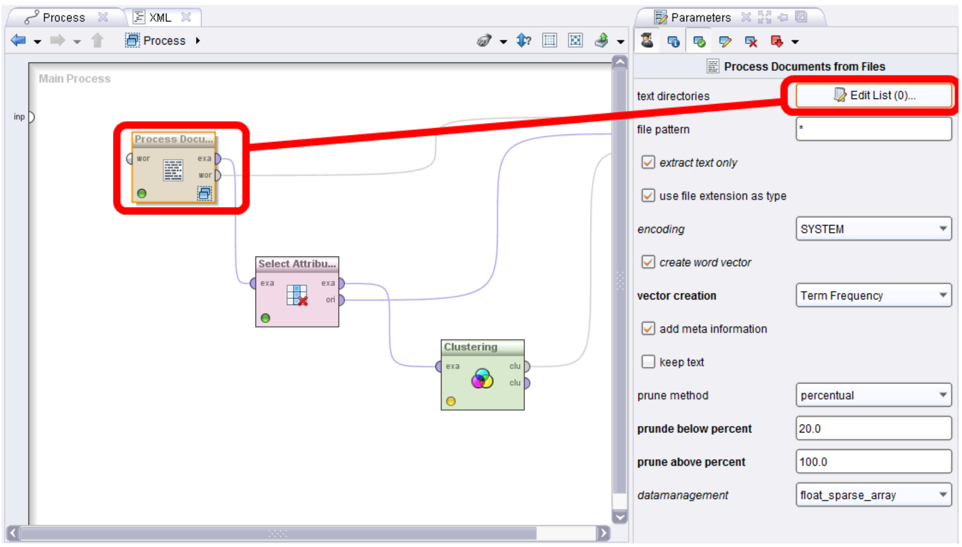
The reason for this error is that we have not specified the data set for Process02 to solve this problem. To resolve this, first click on Close (Figure 1.51). Then, as in Figure 1.52, click on Process Documents from Files operator and click on the button next to the text directories parameter inside the Parameters view. Then, select the same data set as you have selected for Process01, as shown in Figures 1.28, 1.29, 1.30 and 1.31. Now the data source has been specified and we can run Process02 to obtain tangible results. Click on the run button as in Figure 1.50, and answer Yes to the questions that are prompted.


Process Failed

The example set contains not enough examples to perform this operation. Needs at least '5' examples.
Some operators need at least a fixed number of examples. This number might depend on the parameters of the operator like the number of clusters in some clustering operators.

Offending operator: Clustering

Figure 1.51 Error message due to not specifying the source data for Process02



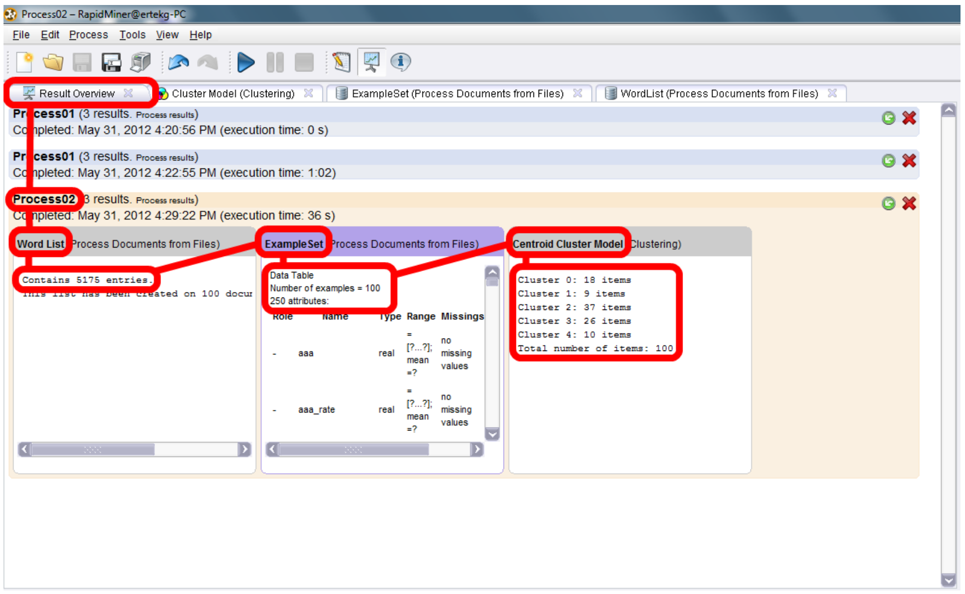
Process Documents from Files

text directories	Edit List (0)...
file pattern	*
☑ extract text only	
☑ use file extension as type	
encoding	SYSTEM
☑ create word vector	
vector creation	Term Frequency
☑ add meta information	
☐ keep text	
prune method	percentual
prunde below percent	20.0
prune above percent	100.0
datamanagement	float_sparse_array

Figure 1.52 Specifying the text directories for the source data of Process02

1.5.3 Process02 Results

The results of Process02 are shown in Figure 1.53. There are three types of results generated after running Process02:


Process02 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
Cluster Model (Clustering)
ExampleSet (Process Documents from Files)
WordList (Process Documents from Files)

Pr cess01 (3 results. Process results)
Co pleted: May 31, 2012 4:20:56 PM (execution time: 0 s)

Pr cess01 (3 results. Process results)
Co pleted: May 31, 2012 4:22:55 PM (execution time: 1:02)

Process02 3 results. Process results)
Co pleted: May 31, 2012 4:29:22 PM (execution time: 36 s)

Word List Process Documents from Files)

```
Contains 5175 entries.
This list has been created on 100 docur
```

ExampleSet Process Documents from Files)

Data Table
Number of examples = 100
250 attributes:

Role	Name	Type	Range	Missings
-	aaa	real	= [?...?]; mean =?	no missing values
-	aaa_rate	real	= [?...?]; mean =?	no missing values

Centroid Cluster Model Clustering)

```
Cluster 0: 18 items
Cluster 1: 9 items
Cluster 2: 37 items
Cluster 3: 26 items
Cluster 4: 10 items
Total number of items: 100
```

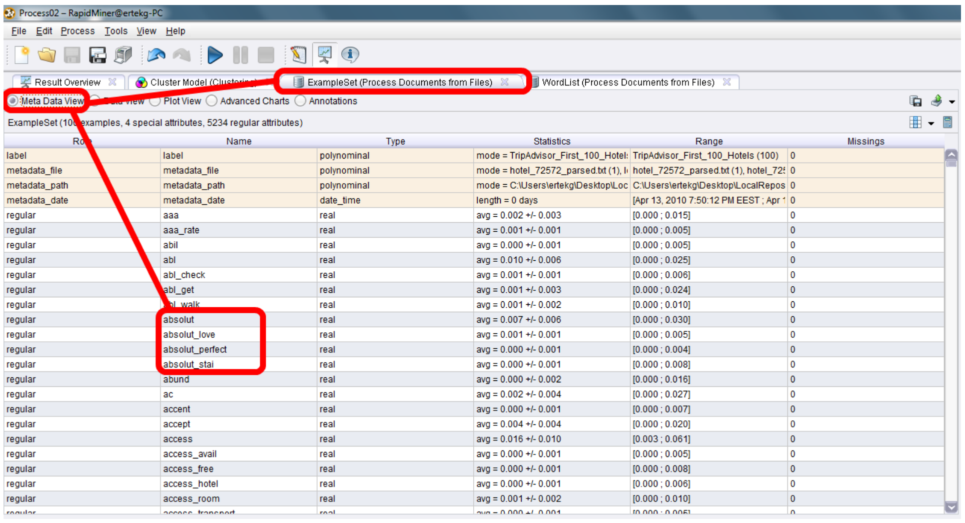
Figure 1.53 Result Overview for Process02 results

Firstly, the Word List contains the filtered words from the documents in our example, the Word List generated Contains 5175 entries.

The second result is the Example Set, which contains data requiring the word composition of each of the documents in the database.

The third result of Process02 is Centroid Cluster Model. In our example five clusters (Cluster 0, Cluster 1, Cluster2, Cluster 3, Cluster 4) have been generated, where each cluster contains a subset of 100 documents.

The Word List is very similar to the word list of Process01 word list. However, there is a difference due to the additional Generate n-Grams (Terms) operator (Figure 1.19). This additional operator extracts n-grams (sequences of words), as well as single words. To see these generated n-grams as a part of the word list you can click on the WordList view or the Example view. Click on the Example view, as in Figure 1.54 and you will see several of these n-grams. For example, besides the word absolut the n-gram words absolut_love, absolut_perfect, absolut_stai are part of the WordList, and appear as attributes in the ExampleSet (Figure 1.54). Now, to see the data itself, click on Data View radio button as in Figure 1.55. In this grid, each number tells the relative frequency of a word in a document. For example, the word absolut (the word that have absolut as their stem) constitutes 0.030 proportion (3%) of all the words in document number 13, that contains the reviews for hotel_73943.

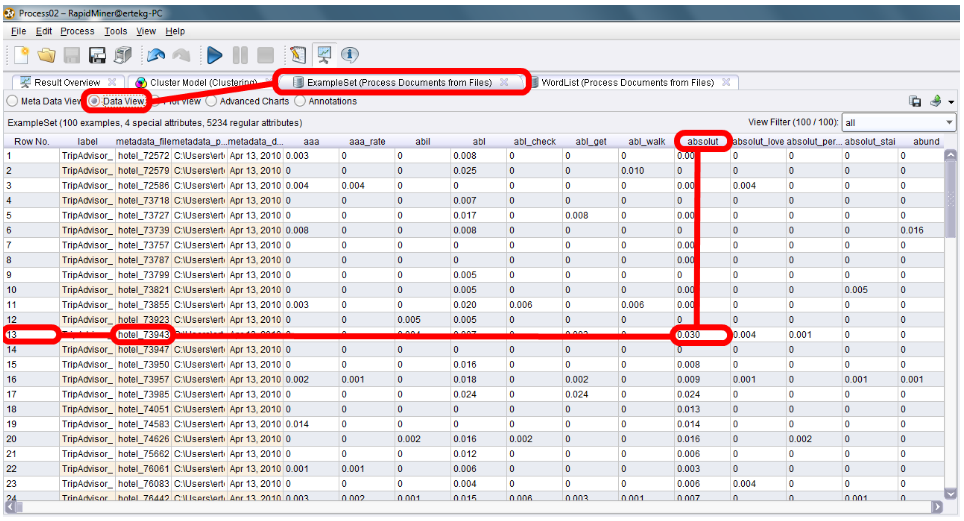

Process02 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
Cluster Model (Clustering)
ExampleSet (Process Documents from Files)
WordList (Process Documents from Files)
Meta Data View
Data View
Plot View
Advanced Charts
Annotations

ExampleSet (100 examples, 4 special attributes, 5234 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	label	polynominal	mode = TripAdvisor_First_100_Hotel:	TripAdvisor_First_100_Hotels (100)	0
metadata_file	metadata_file	polynominal	mode = hotel_72572_parsed.txt (1), l	hotel_72572_parsed.txt (1), hotel_72	0
metadata_path	metadata_path	polynominal	mode = C:\Users\ertekg\Desktop\Loc	C:\Users\ertekg\Desktop\LocalRepos	0
metadata_date	metadata_date	date_time	length = 0 days	[Apr 13, 2010 7:50:12 PM EEST ; Apr 1	0
regular	aaa	real	avg = 0.002 +/- 0.003	[0.000 ; 0.015]	0
regular	aaa_rate	real	avg = 0.001 +/- 0.001	[0.000 ; 0.005]	0
regular	abil	real	avg = 0.000 +/- 0.001	[0.000 ; 0.005]	0
regular	abl	real	avg = 0.010 +/- 0.006	[0.000 ; 0.025]	0
regular	abl_check	real	avg = 0.001 +/- 0.001	[0.000 ; 0.006]	0
regular	abl_get	real	avg = 0.001 +/- 0.003	[0.000 ; 0.024]	0
regular	abl_walk	real	avg = 0.001 +/- 0.002	[0.000 ; 0.010]	0
regular	absolut	real	avg = 0.007 +/- 0.006	[0.000 ; 0.030]	0
regular	absolut_love	real	avg = 0.001 +/- 0.001	[0.000 ; 0.005]	0
regular	absolut_perfect	real	avg = 0.000 +/- 0.001	[0.000 ; 0.004]	0
regular	absolut_stai	real	avg = 0.000 +/- 0.001	[0.000 ; 0.008]	0
regular	abund	real	avg = 0.000 +/- 0.002	[0.000 ; 0.016]	0
regular	ac	real	avg = 0.002 +/- 0.004	[0.000 ; 0.027]	0
regular	accent	real	avg = 0.000 +/- 0.001	[0.000 ; 0.007]	0
regular	accept	real	avg = 0.004 +/- 0.004	[0.000 ; 0.020]	0
regular	access	real	avg = 0.016 +/- 0.010	[0.003 ; 0.061]	0
regular	access_avail	real	avg = 0.000 +/- 0.001	[0.000 ; 0.005]	0
regular	access_free	real	avg = 0.000 +/- 0.001	[0.000 ; 0.008]	0
regular	access_hotel	real	avg = 0.000 +/- 0.001	[0.000 ; 0.006]	0
regular	access_room	real	avg = 0.001 +/- 0.002	[0.000 ; 0.010]	0
regular	access_transport	real	avg = 0.000 +/- 0.001	[0.000 ; 0.005]	0

Figure 1.54 Meta Data View for the ExampleSet generated by Process02, including the n-Grams

The main contribution of Process02 is the cluster model, which clusters documents, and thus the hotels, according to the similarity of the frequency distribution of the words contained in their TripAdvisor reviews. Now click on Cluster Model (clustering) view to view the cluster model results, as in Figure 1.56. Then, to view the content of each of the clusters, click on Folder View radio button, as in Figure 1.57. In our example the first cluster of hotels is cluster_0, which contains 28 hotels, including hotels with row numbers 47, 51, 55, 95, 100.


Process02 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
Cluster Model (Clustering)
ExampleSet (Process Documents from Files)
WordList (Process Documents from Files)
Meta Data View
Plot View
Advanced Charts
Annotations

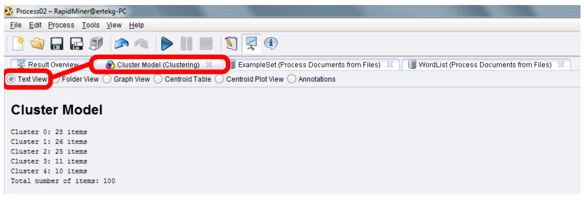
Meta Data View [✓ Data View] Plot View Advanced Charts Annotations

ExampleSet (100 examples, 4 special attributes, 5234 regular attributes)

View Filter (100 / 100): all

Row No.	label	metadata_file	metadata_p...	metadata_d...	aaa	aaa_rate	abil	abl	abl_check	abl_get	abl_walk	absolut	absolut_love	absolut_per...	absolut_stai	abund
1	TripAdvisor_	hotel_72572	C:\Users\ert	Apr 13, 2010	0.003	0	0	0.008	0	0	0	0.00	0	0	0	0
2	TripAdvisor_	hotel_72579	C:\Users\ert	Apr 13, 2010	0	0	0	0.025	0	0	0.010	0	0	0	0	0
3	TripAdvisor_	hotel_72586	C:\Users\ert	Apr 13, 2010	0.004	0.004	0	0	0	0	0	0.00	0.004	0	0	0
4	TripAdvisor_	hotel_73718	C:\Users\ert	Apr 13, 2010	0	0	0	0.007	0	0	0	0	0	0	0	0
5	TripAdvisor_	hotel_73727	C:\Users\ert	Apr 13, 2010	0	0	0	0.017	0	0.008	0	0.00	0	0	0	0
6	TripAdvisor_	hotel_73739	C:\Users\ert	Apr 13, 2010	0.008	0	0	0.008	0	0	0	0	0	0	0	0.016
7	TripAdvisor_	hotel_73757	C:\Users\ert	Apr 13, 2010	0	0	0	0	0	0	0	0.00	0	0	0	0
8	TripAdvisor_	hotel_73787	C:\Users\ert	Apr 13, 2010	0	0	0	0	0	0	0	0.00	0	0	0	0
9	TripAdvisor_	hotel_73799	C:\Users\ert	Apr 13, 2010	0	0	0	0.005	0	0	0	0	0	0	0	0
10	TripAdvisor_	hotel_73821	C:\Users\ert	Apr 13, 2010	0	0	0	0.005	0	0	0	0.00	0	0	0.005	0
11	TripAdvisor_	hotel_73855	C:\Users\ert	Apr 13, 2010	0.003	0	0	0.020	0.006	0	0.006	0.00	0	0	0	0
12	TripAdvisor_	hotel_73923	C:\Users\ert	Apr 13, 2010	0	0	0.005	0.005	0	0	0	0	0	0	0	0
13	TripAdvisor_	hotel_73943	C:\Users\ert	Apr 13, 2010	0	0	0.004	0.007	0	0.003	0	0.030	0.004	0.001	0	0
14	TripAdvisor_	hotel_73947	C:\Users\ert	Apr 13, 2010	0	0	0	0	0	0	0	0	0	0	0	0
15	TripAdvisor_	hotel_73950	C:\Users\ert	Apr 13, 2010	0	0	0	0.016	0	0	0	0.008	0	0	0	0
16	TripAdvisor_	hotel_73957	C:\Users\ert	Apr 13, 2010	0.002	0.001	0	0.018	0	0.002	0	0.009	0.001	0	0.001	0.001
17	TripAdvisor_	hotel_73985	C:\Users\ert	Apr 13, 2010	0	0	0	0.024	0	0.024	0	0.024	0	0	0	0
18	TripAdvisor_	hotel_74051	C:\Users\ert	Apr 13, 2010	0	0	0	0	0	0	0	0.013	0	0	0	0
19	TripAdvisor_	hotel_74583	C:\Users\ert	Apr 13, 2010	0.014	0	0	0	0	0	0	0.014	0	0	0	0
20	TripAdvisor_	hotel_74626	C:\Users\ert	Apr 13, 2010	0	0	0.002	0.016	0.002	0	0	0.016	0	0.002	0	0
21	TripAdvisor_	hotel_75662	C:\Users\ert	Apr 13, 2010	0	0	0	0.012	0	0	0	0.006	0	0	0	0
22	TripAdvisor_	hotel_76061	C:\Users\ert	Apr 13, 2010	0.001	0.001	0	0.006	0	0	0	0.003	0	0	0	0
23	TripAdvisor_	hotel_76083	C:\Users\ert	Apr 13, 2010	0	0	0	0.004	0	0	0	0.006	0.004	0	0	0
24	TripAdvisor_	hotel_76442	C:\Users\ert	Apr 13, 2010	0.003	0.002	0.001	0.015	0.006	0.003	0.001	0.007	0	0	0.001	0

Figure 1.55 Data View for the ExampleSet generated by Process02, and the relative occurence frequency of the word absolut in `hotel_73943.txt`

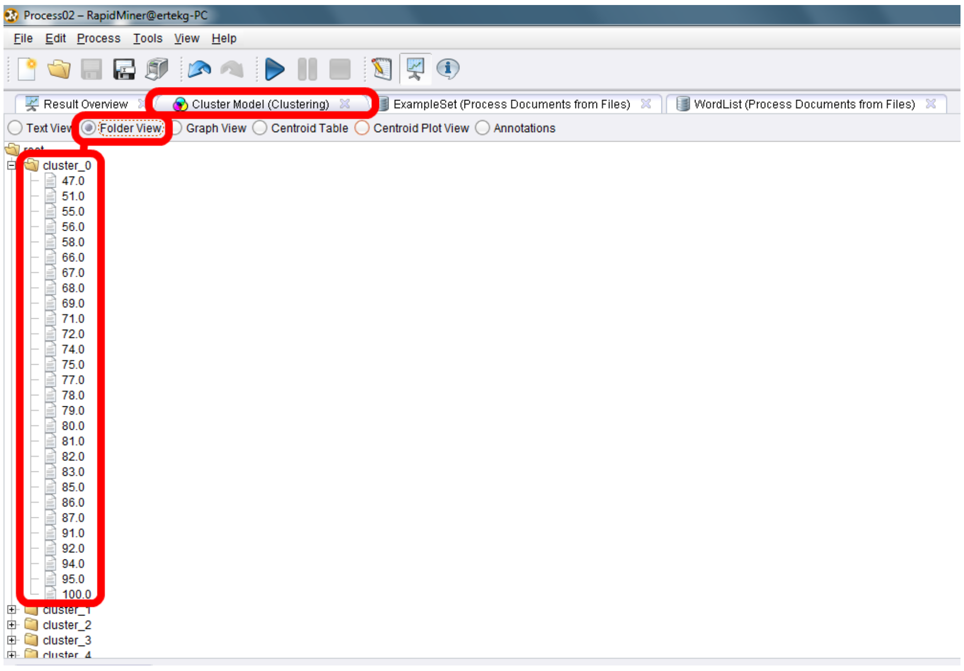

Process02 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
Cluster Model (Clustering)
ExampleSet (Process Documents from Files)
WordList (Process Documents from Files)

[✓ Text View] Folder View Graph View Centroid Table Centroid Plot View Annotations

Cluster Model

```
Cluster 0: 28 items
Cluster 1: 26 items
Cluster 2: 25 items
Cluster 3: 11 items
Cluster 4: 10 items
Total number of items: 100
```

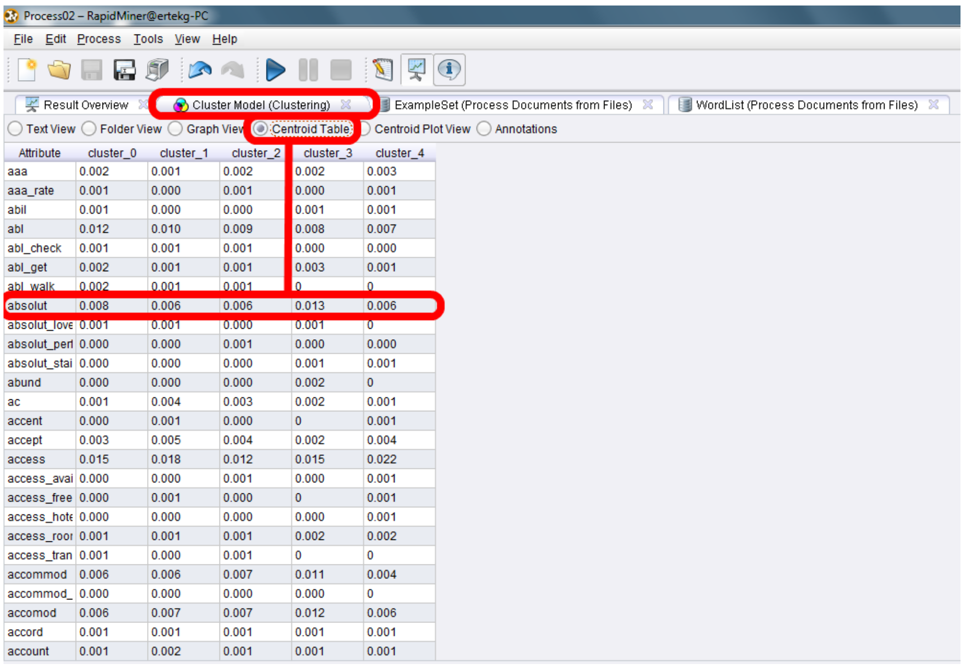
Figure 1.56 Text View for the Cluster Model generated by Process02, displaying the number of examples in each cluster


Process02 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
Cluster Model (Clustering)
ExampleSet (Process Documents from Files)
WordList (Process Documents from Files)
Text View [✓ Folder View] Graph View Centroid Table Centroid Plot View Annotations

```
root
  cluster_0
    47.0
    51.0
    55.0
    56.0
    58.0
    66.0
    67.0
    68.0
    69.0
    71.0
    72.0
    74.0
    75.0
    77.0
    78.0
    79.0
    80.0
    81.0
    82.0
    83.0
    85.0
    86.0
    87.0
    91.0
    92.0
    94.0
    95.0
    100.0
  cluster_1
  cluster_2
  cluster_3
  cluster_4
```

Figure 1.57 Folder View for the Cluster Model generated by Process02, displaying the examples in each cluster

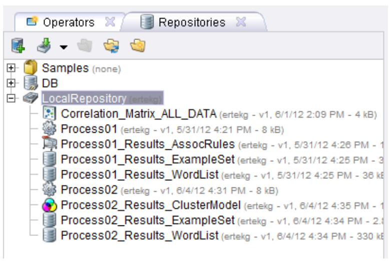
How do these clusters differ from each other? To get an answer to this question, click on Centroid Table radio button (Figure 1.58). In this Centroid table we can observe the average frequency of each word in the documents of each cluster. For example the word absolut appears much more frequently in cluster_3, compared to the other clusters: In cluster_3, the average frequency for absolut is 0.013, whereas its at most 0.008 in the others.


Process02 – RapidMiner@ertekg-PC
File Edit Process Tools View Help
Result Overview
Cluster Model (Clustering)
ExampleSet (Process Documents from Files)
WordList (Process Documents from Files)
Text View
Folder View
Graph View
Centroid Table
Centroid Plot View
Annotations

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
aaa	0.002	0.001	0.002	0.002	0.003
aaa_rate	0.001	0.000	0.001	0.000	0.001
abil	0.001	0.000	0.000	0.001	0.001
abl	0.012	0.010	0.009	0.008	0.007
abl_check	0.001	0.001	0.001	0.000	0.000
abl_get	0.002	0.001	0.001	0.003	0.001
abl_walk	0.002	0.001	0.001	0	0
absolut	0.008	0.006	0.006	0.013	0.006
absolut_love	0.001	0.001	0.000	0.001	0
absolut_perf	0.000	0.000	0.001	0.000	0.000
absolut_stai	0.000	0.000	0.000	0.001	0.001
abund	0.000	0.000	0.000	0.002	0
ac	0.001	0.004	0.003	0.002	0.001
accent	0.000	0.001	0.000	0	0.001
accept	0.003	0.005	0.004	0.002	0.004
access	0.015	0.018	0.012	0.015	0.022
access_avai	0.000	0.000	0.001	0.000	0.001
access_free	0.000	0.001	0.000	0	0.001
access_hote	0.000	0.000	0.000	0.000	0.001
access_roor	0.001	0.001	0.001	0.002	0.002
access_tran	0.001	0.000	0.001	0	0
accommod	0.006	0.006	0.007	0.011	0.004
accommod_	0.000	0.000	0.000	0.000	0
accomod	0.006	0.007	0.007	0.012	0.006
accord	0.001	0.001	0.001	0.001	0.001
account	0.001	0.002	0.001	0.001	0.001

Figure 1.58 Centroid Table for the Cluster Model generated by Process02, displaying the average frequency of each word in each cluster

Finally, save the results of Process02, just as you did for Process01. The final picture in the Repositories view will be as in Figure 1.59.



```
⊞ Samples (none)
⊞ DB
⊟ LocalRepository (ertekg)
   Correlation_Matrix_ALL_DATA (ertekg - v1, 6/1/12 2:09 PM - 4 kB)
   Process01 (ertekg - v1, 5/31/12 4:21 PM - 8 kB)
   Process01_Results_AssocRules (ertekg - v1, 5/31/12 4:26 PM - 1
   Process01_Results_ExampleSet (ertekg - v1, 5/31/12 4:25 PM - 3
   Process01_Results_WordList (ertekg - v1, 5/31/12 4:25 PM - 36 kB
   Process02 (ertekg - v1, 6/4/12 4:31 PM - 8 kB)
   Process02_Results_ClusterModel (ertekg - v1, 6/4/12 4:35 PM - 1
   Process02_Results_ExampleSet (ertekg - v1, 6/4/12 4:34 PM - 2.
   Process02_Results_WordList (ertekg - v1, 6/4/12 4:34 PM - 330 kB
```

Figure 1.59 Final view of LocalRepository with the processes and their results saved

1.6 CONCLUSIONS

In this chapter, we have shown the basic RAPIDMINER operators for text mining, as well as how they can be used in conjunction with other operators that implement popular conventional data mining techniques. Throughout the chapter, we have discussed two processes, Process01 and Process02, which complement text processing with association mining and cluster modeling, respectively. We have shown how these text mining processes, which are formed as combinations of text processing and data mining techniques, are modeled and run in RAPIDMINER, and discussed how their results are analyzed. RAPIDMINER has an extensive set of operators available for text processing and text mining, which can also be used for extracting data from the web, and perform a multitude of other types of analysis. While most of these operators were not discussed in this chapter due to space limitations, the reader is strongly encouraged not to suffice with the contents of this chapter, and explore and experiment other text processing and mining operators and capabilities within RAPIDMINER, its official extensions, and the processes posted by the RAPIDMINER user community.

Acknowledgement

This work is financially supported by the UbiPOL Project within the EU Seventh Framework Programme, under grant FP7-ICT 248010. The authors thank Assoc. Prof. Yücel Saygın for supporting and helping the writing of this chapter. The authors also acknowledge the earlier help of Sabancı University alumni Yücel Balıklılı and Emir Balıkçı with the design of the first process in this chapter.

GLOSSARY

Decision Tree Decision Trees are....

Data Exploration Data Exploration is of particular importance when...