

ONLINE TUTORIAL T3: TEXT MINING PROJECT

The sample data file 4Cars.sta, available at this Web site contains car reviews written by automobile owners. Car reviews related to four popular brands were extracted from the Web site (carreview.com) and one of the brands was renamed to conceal certain findings. The attributes of this file are:

Unstructured Data	Structured Data
1. Summary	4. Overall Rating
2. Strengths	5. Price paid
3. Weaknesses	6. Car type

Each row (or case or instance) contains opinions (Summary, Strength or Weaknesses) filed by car owners about the cars they own along with other information, the rating for the car, the overall price they paid, and the car type. The purpose of this analysis is to see whether we can extract some information from textual corpus via the nascent concept of text mining.

To get started, follow these steps:

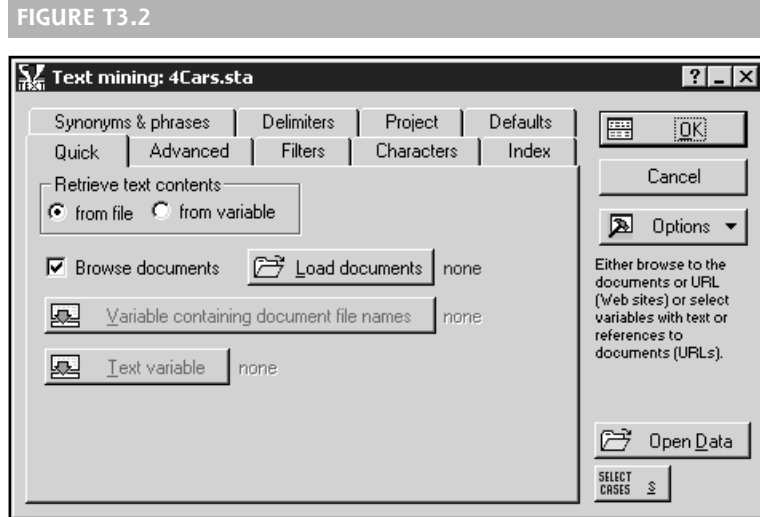
1. Open STATISTICA by selecting **Start, Programs, STATISTICA**.
2. Close the **Welcome to STATISTICA** dialog, the **Data Miner** workspace, and the spreadsheet.
3. From the **Files** menu, select **Open**. Open the **4Cars.sta** file from the **Examples/Datasets** folder. (Note: In most default installations of STATISTICA, you will find the sample data files in the **Examples/Datasets** folder of STATISTICA.) The file looks as shown in Figure T3.1.
4. Next, select **Statistics, Text & Document Mining, Web Crawling** and then select **Text Mining & Document Retrieval** to display the **Text mining** startup panel (shown in Figure T3.2).

This dialog contains nine tabs: **Quick, Advanced, Filters, Characters, Index, Synonyms & phrases, Delimiters, Project, and Defaults**. Use the options on this dialog to specify the documents to be analyzed; the words, terms, and phrases that are to be included or ignored; and the database (internal, and used only for this module) where the indexed terms are to be stored. You can also select an existing database and thus “score” new documents, using the terms selected (and saved) in that database.

Next, you need to extract the word frequencies from the text contents from the first variable, **Summary**. To do so, follow these steps:

FIGURE T3.1

	1 Summary	2 Strengths	3 Weaknesses	4 Overall Rating	5 Price Paid	6 Car Type
1	This car is unbelievable,	Acceleration! Speed, L	fuel injection delay whil	5	36000	Lexus
2	I'm a high school senior,	Interior -Woodgrain, Le	I have my vehicle on le	5	20000	CarZZ
3	I think ES400 is better, j	Very good engine. Dep	Looks like Camry. Heig	3	7650	Lexus
4	I can't even imagine buyi	Engine, Handling, Styl	I guess I would have m	5	33000	CarZZ
5	Yes, I would recommend	Couldn't believe how qu	Handling- feels like a b	5	34000	Lexus
6	I love this car. I've been t	Looks. Mine is in billiar	Some parts are expens	4	9200	BMW
7	I recommend this car to	This is a great car. Pov	In dash lights and heac	5	22000	Lexus



1. From the **Quick** tab select the **from variable** option from the **Retrieve text contents** section.
2. Click the **Text variable** button to display **Select a variable containing texts** dialog (see Figure T3.3) and then select the variable **Summary** (which is the variable that contains the text body).
3. Click **OK** to return to the **Quick** tab.
4. Click the **Index** tab and then click the **Edit stop-word file** button. Drag down to the bottom of the **stop word list** and add the word **car** to the list, as shown in Figure T3.4.
5. Click the **OK [Save]** button to add the word *car* to the existing stop word file. The terms contained in this stop list will be excluded from the indexing that occurs during the processing of the documents.
6. Click the **OK** button on the **Text mining: 4Cars.sta** dialog to begin the processing of the documents. After a few seconds (or minutes, depending on the speed of your computer hardware), the **TM results: 4Cars.sta** dialog is displayed, as shown in Figure T3.5.

The TM results dialog gives a brief summary (i.e., number of documents, selected words, and unselected words) and displays the words extracted by the Text Mining module. (The options available at this point are described in some detail in the next step.)

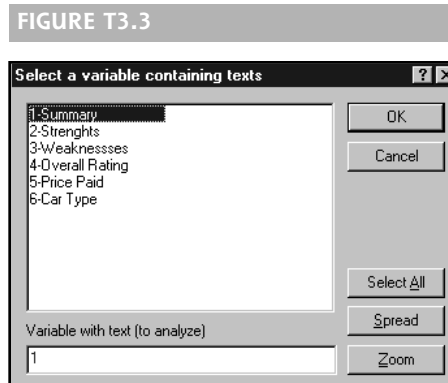
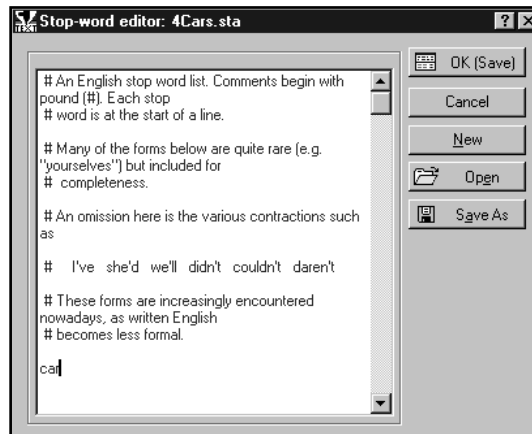


FIGURE T3.4

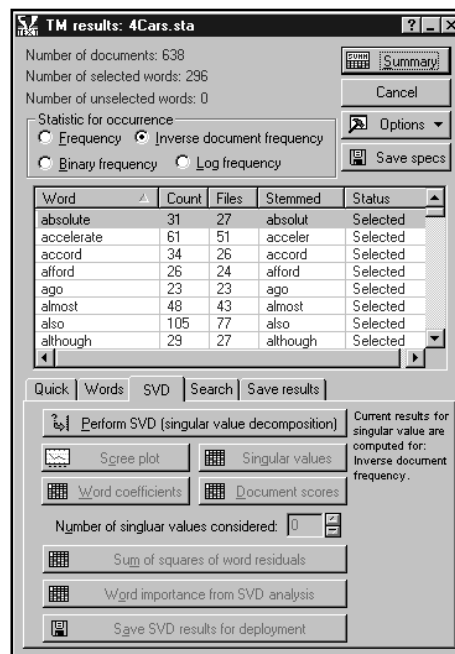


7. Click the **Count** header to sort the words by frequency. You will see that the word *drive* appears with the highest frequency, followed by *great*. You can select the following options on this window, as required:

Frequency. Select this option button to analyze and report the simple word frequencies.

Binary frequency. Select this option button to analyze and report binary indicators instead of word frequencies. Specifically, this option simply enumerates whether a term is used in a document. The resulting documents-by-words matrix will contain only 1s and 0s, to indicate the presence or absence of the respective word. Like the other transformations of simple word frequencies,

FIGURE T3.5



this transformation will dampen the effect of the raw frequency counts on subsequent computations and analyses.

Inverse document frequency. Select this option button to analyze and report inverse document frequencies. One issue that you might want to consider more carefully, and reflect in the indices used in further analyses, is the relative *document frequencies (df)* of different words. For example, a term such as *guess* might occur frequently in all documents, whereas another term, such as *software*, might occur only in a few. The reason is that one might make guesses in various contexts, regardless of the specific topic, whereas software is a more semantically focused term that is likely to occur only in documents that deal with computer software. A common and very useful transformation that reflects both the specificity of words (i.e., document frequencies) as well as the overall frequency of their occurrences (i.e., word frequencies) is the *inverse document frequency (idf)*. This option includes both the dampening of the simple word frequencies via the log function and a weighting factor that evaluates to 0 if the word occurs in all documents, up to the maximum value when a word occurs in only a single document. You can easily see how this transformation will create indices that reflect both the relative frequencies-of-occurrences of words and their semantic specificities over the documents included in the analysis.

Log frequency. Select this option button to analyze and report logs of the raw word frequencies. A common transformation of the raw word frequency counts is to “dampen” the raw frequencies and see how they will affect the results of subsequent computations.

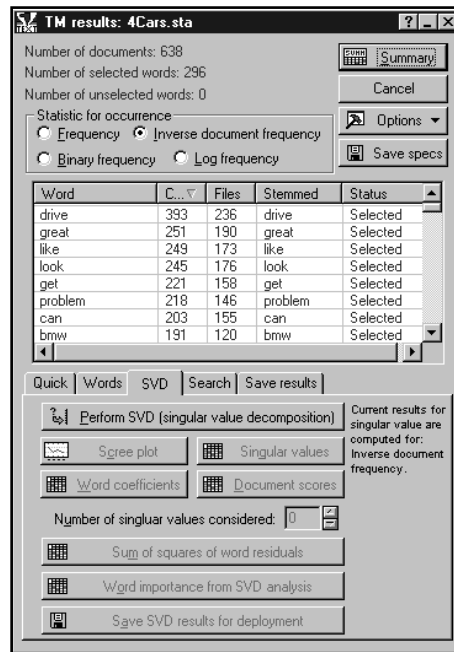
Summary. Click the **Summary** button to compute the summary of word occurrence in document; you get the same results here as with the option by the longer name on the **Quick** tab. Specifically, the results spreadsheet will contain a row for each input document and a column for each word or term. The entries in the cells of the results dialog depend on the option selection in the **Statistic for occurrence** group box on this dialog. The summary spreadsheet can quickly be turned into an input spreadsheet for subsequent analyses; to do this, you use the options on the **Save results** tab to write the respective word statistics to another file or database.

In this case, select the option **Inverse document frequency** (it’s the most efficient and frequently used option to represent the word counts) from the **Statistic for occurrence** section. Also, select the **SVD** tab (see Figure T3.6).

At this point, there are different features/techniques available from different tabs (**Quick**, **Words**, **SVD**, **Search**, and **Save results**) from which the “numericized” words from the **TM results** dialog can be saved into a standalone spreadsheet for further analysis. (Refer to the electronic manual to learn more about the features available in these tabs.)

Next, you need to perform the singular value decomposition. The SVD tab of the **TM results** dialog provides options to perform singular value decomposition on the document-by-words matrix, based on the selected words only, and with the word frequencies or transformed word frequencies as currently selected in the **Statistic for occurrence** group box on the **TM results** dialog. Note that the results will be available only for one particular type of matrix (i.e., transformation of the word frequencies), and when you change your selection on the **TM results** dialog, any previously computed results for singular value decomposition will be discarded. In addition, when you save SVD results for deployment, only the singular value decomposition results for the specified word frequencies or their transformations will be saved.

FIGURE T3.6



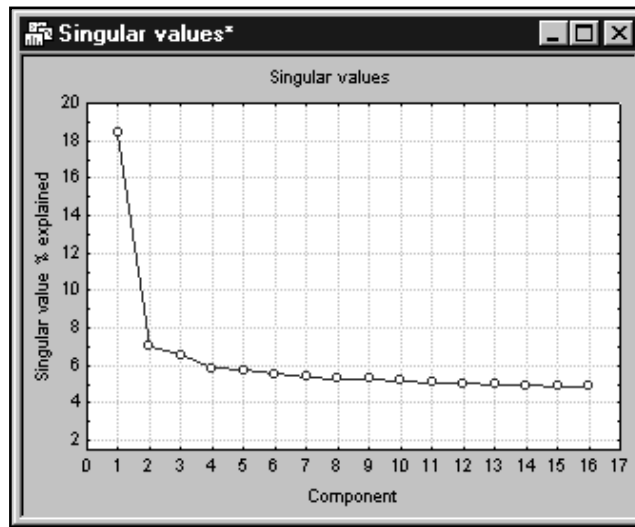
Singular value decomposition is an analytic tool for feature extraction that can be used to determine a few underlying dimensions that account for most of the common contents, or meaning, of the documents and words that were extracted.

Save SVD results for deployment. Click this button to save the SVD results for deployment (i.e., to “score” new documents). Specifically, this option saves the current SVD results for the currently selected words; this information is saved in the current database, which can then be used in subsequent analyses to automatically index and score new documents. Thus, this option is essential for many applications of text mining (as, for example, discussed in the introductory overview at statsoft.com/textbook/stathome.html), for example, to implement automatic mail filters or text routing systems.

Now you are ready to perform the SVD. To do so, follow these steps:

1. Click the **Perform SVD** button on the **SVD** tab. When this computation is complete, all the other options available in this tab are enabled. At this point, you can extract different details of the statistics and results that can be used for further analysis.
2. Click the **Scree plot** button to view the graph shown in Figure T3.7. This plot is useful for determining the number of singular values that are useful and informative and that should be retained for subsequent analyses. It helps to visually determine the number of components that explain the variance among the inputs. You can tell by looking at the graph that the first component explains slightly more than 18 percent of the total variance for 295 words that were used as inputs, followed by the second component, which explains 7 percent, and so on. So 25 percent of the variance present within the inputs is explained by the top two

FIGURE T3.7



components. Usually, the number of informative dimensions to retain for subsequent analysis is determined by locating the elbow in the scree plot. The scree test involves finding the place where the smooth decrease of singular values appears to level off to the right of the plot. To the right of that point, presumably, you find only SVD scree; *scree* is a geological term that refers to the debris that collects on the lower part of a rocky slope. Thus, no more than the number of components to the left of this point will be useful for analysis.

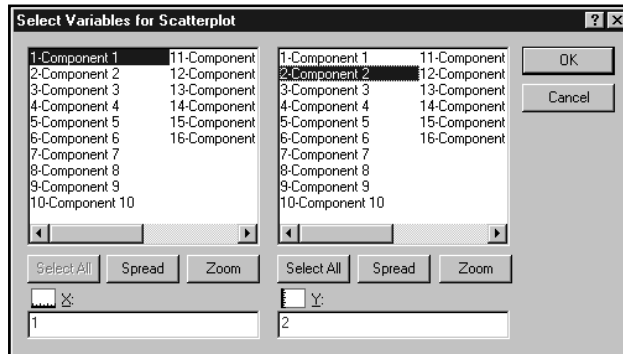
3. Click the **Word coefficient** button to view the results spreadsheet, which holds the word coefficient based on the results from the SVD (see Figure T3.8).

As explained earlier, you can now draw scatterplots to explore the meaning of the dimensions to which the words and documents are mapped (i.e., to understand the semantic space for the extracted words or terms or documents). You should now

FIGURE T3.8

	Component 1	Component 2	Component 3	Component 4	Component 5
absolute	0.000297	-0.000702	-0.000479	0.000259	-0.001227
accelerate	0.000573	-0.001923	0.002586	0.002156	0.001176
accord	0.000372	0.000001	-0.000712	0.000701	0.002021
afford	0.000329	0.000563	0.000296	-0.001284	-0.000527
ago	0.000260	0.000342	-0.000219	-0.000618	-0.000312
almost	0.000465	-0.000459	0.000417	0.000691	-0.000258
also	0.000733	-0.000067	0.001840	0.001052	0.000761
although	0.000359	-0.000449	0.000089	-0.000182	0.000688
always	0.000365	-0.000718	-0.000617	0.000969	-0.001980
amaze	0.000250	-0.000526	-0.000374	0.001766	-0.000374
another	0.000456	0.000522	-0.001231	0.000500	0.000360
anyone	0.000273	0.000271	-0.000707	-0.000578	-0.000567
around	0.000394	-0.000548	0.000007	0.000996	-0.000327
ask	0.000373	0.002229	0.001270	-0.000088	0.000745
audi	0.000412	-0.000725	-0.001599	0.001119	0.000982

FIGURE T3.9



use the top two components to draw a scatterplot to plot the important words picked by the components. Here’s what you do:

1. Right-click the **SVD Word Coefficients** spreadsheet header (which appears in the left pane of the workbook) and select the option **Use as Active Input**.
2. Select **Scatterplots** from the **Graphs** menu option. Then click the **Variables** button in the **2D Scatterplot** dialog.
3. Select **Component 1** as the **X:** variable and **Component 2** as **Y:** variable, as shown in Figure T3.9.
4. Click **OK** in the **Select Variables for Scatterplot** dialog and then the **2D Scatterplot** dialog to view the two-dimensional scatterplot graph (see Figure T3.10).
5. Click the **Brushing** toolbar button on the **Graphs** toolbar (which is by default displayed as the third layer of menu options within the STATISTICA application) to display the **Brushing 2D** dialog, shown in Figure T3.11.

FIGURE T3.11

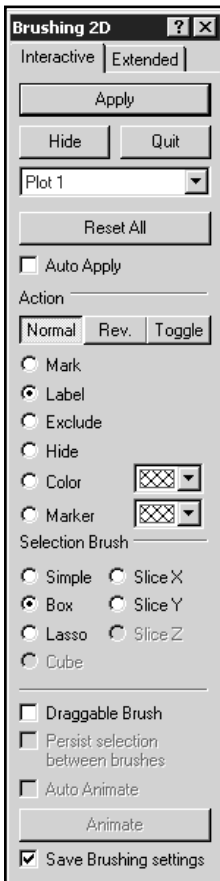


FIGURE T3.10

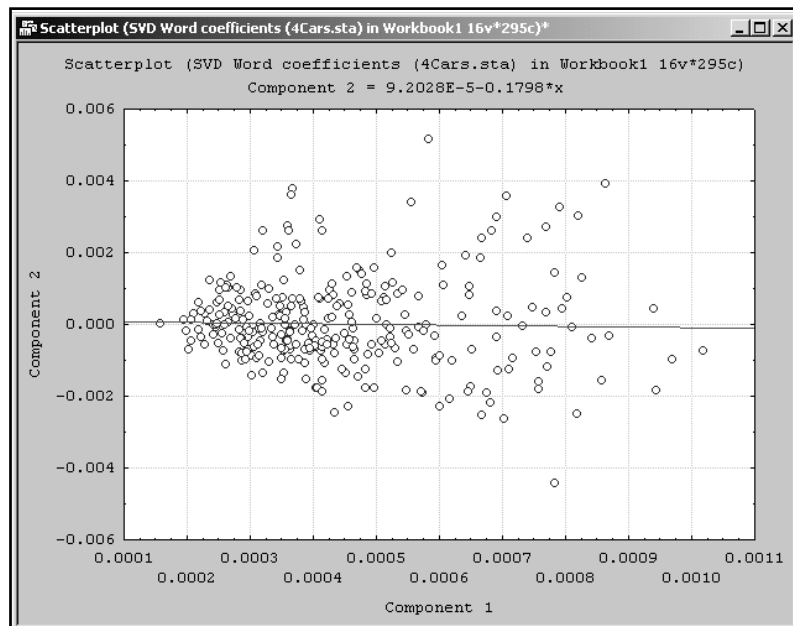
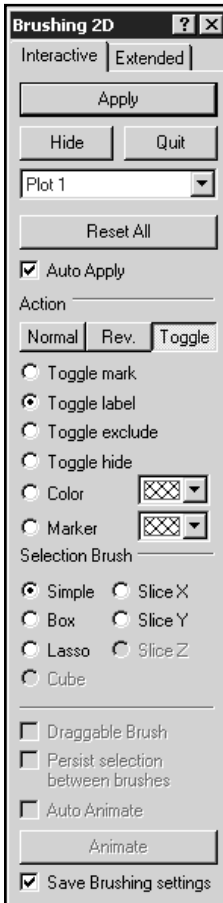


FIGURE T3.12



The **Brushing 2D** dialog contains tools for identifying points or groups of points on both 2D and 3D graphs to be marked, labeled, or temporarily turned off (i.e., removed from the graph and from considerations for fit lines applied and so on) When brushing is activated, the mouse pointer turns by default into a gun-sight-style cross-hair.

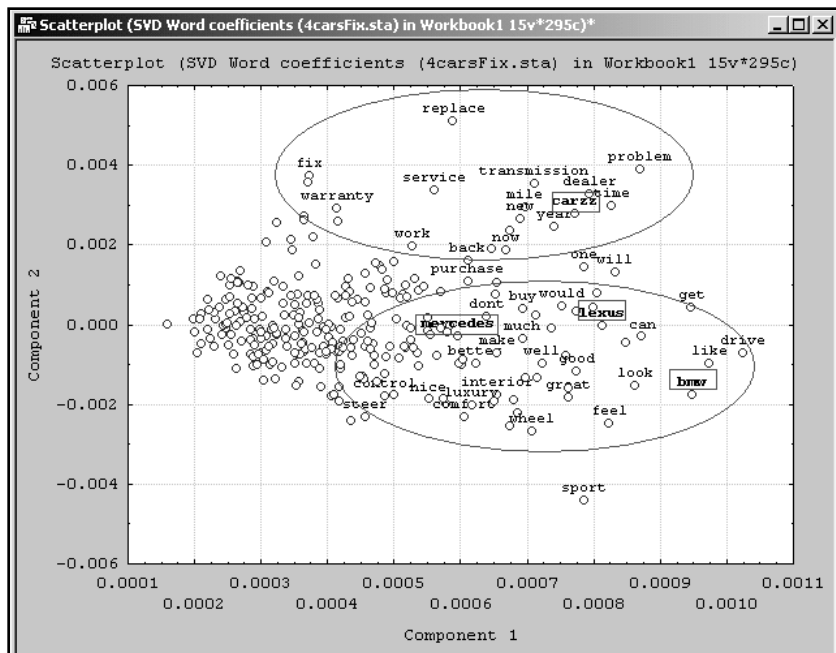
The pointer can be used to select/highlight either individual points (by selecting the **Point** option button under **Selection Brush** on the **Brushing 2D** dialog) or groups of points (via the **Lasso** or **Box** tool). Other options, such as **Slice X, Y, and Z** and **Cube**, can be used to define areas on a two- or three-dimensional plot or volumes on a three-dimensional plot. The areas or volumes defined by the **Lasso, Box, Slice, and Cube** options can be animated to move over the extent of the plot (or a matrix of plots, in some cases) to explore the spatial distribution of values.

With the points highlighted, clicking the **Update** button on the **Brushing 2D** dialog causes the action specified (e.g., labeling, marking, turning off) to be executed. You can reverse actions taken by clicking the **Reset All** button. The **Quit** button closes the dialog, leaving the actions already applied intact.

6. Check the **Auto Apply** option (displayed below the **Reset All** button).
7. Select the **Toggle** tab and then select the **Toggle Label** option and select the **Simple** option from the **Selection Brush** section. Figure T3.12 shows how the **Brushing 2D** dialog should now look.
8. Use the gun-sight-style mouse pointer to click each point to view the actual words that are represented by points. As you start clicking on the points, you see the words being displayed. Figure T3.13 shows some of the words that were used in the reviews, along with the brands that were picked for this analysis.

You can look at such graphs to visualize the semantic (i.e., of or related to meaning in language) spaces of related words. Words that appear close to one another are related to one another. You can see from this graph that the words

FIGURE T3.13



within the second ellipse contain positive words (e.g., comfort, better, quality, good, great), and the first ellipse contains negative words (e.g., problem, replace, fix). You can see that the reviewers used positive words to describe the brands BMW, Lexus, and Mercedes, and they used more negative words to describe CarZZ. This gives you a clear picture of how the reviewers described their experiences with the brands they were using. You can also use other pairs of components to draw scatterplots and further understand or drill-down into the other dimensions.

9. Click the **Document Scores** option to view the spreadsheet shown in Figure T3.14, which holds document scores.

You can now use the document score results (displayed as components) in the spreadsheet to draw scatterplots, as illustrated previously. This time, you see the documents IDs instead of the words, and the resulting scatterplot can be used to visualize the documents that are related (i.e., document IDs falling close are related to each other). You can also try clustering techniques by using the top components (which explains a good percentage of variance) as inputs to identify clusters or groups of related reviews or documents. You can then drill down into the cluster results by using other tools (e.g., feature selection, classification trees) to further explore/understand the differentiating factors of these clusters.

1. Click **Num of vars to add to input data** and increase the value from **1** to **311** to insert word frequencies of 295 words and 16 components from the SVD analysis.
2. Click **Add variables to input spreadsheet** to create **311** new variables within the input spreadsheet. Now you will see 311 new variables created to the right of the existing variables, as shown in Figure T3.15.
3. Click **Save statistics values to input data** on the **TM results: 4Cars.sta** dialog to view the **Assign statistics to variables, to save them to the input data** dialog (see Figure T3.16).
4. Select all the words that appear in the **Statistics** pane and then select new variables from 7 to 318. (Use the **Shift** key to perform this operation.)
5. Click the **Assign** button to assign the word frequencies and the SVD scores to the new variables. Figure T3.17 shows how the **Assign statistics to variables, to save them to input data** dialog should look now.

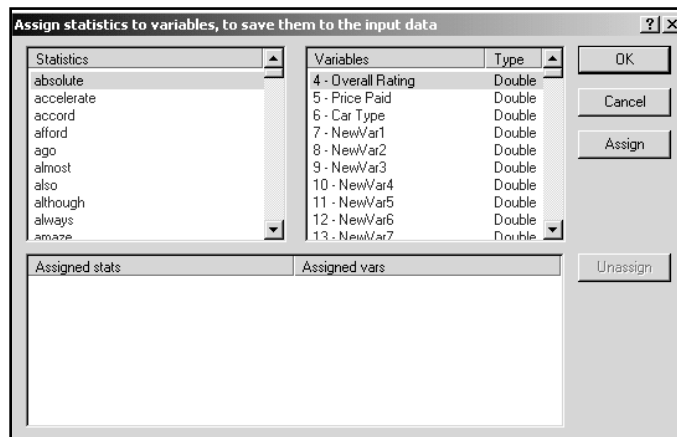
FIGURE T3.14

	Component 1	Component 2	Component 3	Component 4	Component 5
1	0.032035	0.021082	0.020308	-0.045252	0.003652
2	0.010481	0.012563	-0.004221	0.007756	-0.019043
3	0.016792	-0.001061	-0.028411	-0.010442	-0.026448
4	0.009531	0.002627	-0.014248	0.003438	0.003097
5	0.001963	0.001793	-0.002437	0.004146	-0.001726
6	0.031449	0.005077	-0.048610	-0.046402	0.016795
7	0.004019	0.005368	-0.009038	0.004062	-0.006938
8	0.002617	0.006918	0.000517	0.003557	-0.006245
9	0.008535	0.001635	0.004116	0.001007	0.002633
10	0.004506	0.004009	-0.008493	0.010399	-0.001952
11	0.008864	0.003790	-0.010586	0.012938	0.011547
12	0.031431	0.022804	-0.007904	0.015523	0.031048
13	0.017047	-0.010857	-0.011562	-0.019477	0.002036
14	0.051825	0.034704	0.062652	-0.034832	-0.038698
15	0.004047	-0.002673	-0.007051	0.000384	-0.002311
16	0.000548	0.001811	0.000540	0.000444	-0.001205

FIGURE T3.15

	1 Summary	2 Strenghts	3 Weaknesses	4 Overall Rating	5 Price Paid	6 Car Type	7 NewVar1	8 NewVar2
1	This car is unbelievably	Acceleration! Speed, L	fuel injection delay	5	36000	Lexus		
2	I'm a high school senic	Interior -Woodgrain, l	I have my vehicle or	5	20000	CarZZ		
3	I think ES400 is better,	Very good engine. De	Looks like Camry. H	3	7650	Lexus		
4	I can't even imagine bu	Engine, Handling, Styl	I guess I would have	5	33000	CarZZ		
5	Yes, I would recommen	Couldn't believe how	Handling- feels like	5	34000	Lexus		
6	I love this car. I've bee	Looks. Mine is in billi	Some parts are expe	4	9200	BMW		
7	I recommend this car t	This is a great car. Po	In dash lights and h	5	22000	Lexus		
8	Overall great car. Coml		Rear window regulat	2	21000	CarZZ		
9	FirsZy, I am not stricly	I am thinking.... Still t	I only have 1200 ch	2	30000	M8enz		
10	I WOULD RECOMMEND	THIS IS THE BEST VEH	EXPENSIVE TO MAIN	4	32000	M8enz		

FIGURE T3.16



6. Click **OK** to write the results extracted from the Text Mining module to the input spreadsheet. When you click **OK**, you see the extracted word frequencies and the SVD components being inserted into the spreadsheet. The spreadsheet that holds the results is shown in Figure T3.18.

FIGURE T3.17

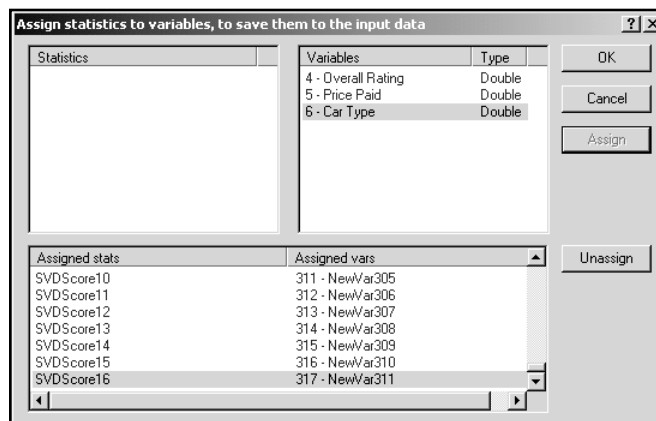


FIGURE T3.18

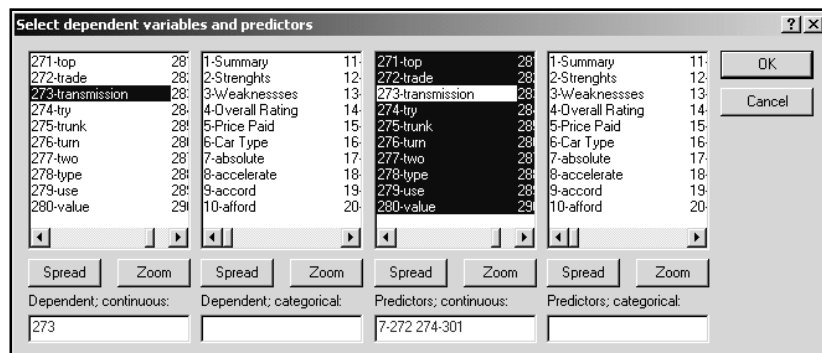
	4 Overall Rating	5 Price Paid	6 Car Type	7 absolute	8 accelerate	9 accord	10 afford	11 ago	12 almost	13 also	14 although	15 always
1	5	36000	Lexus	0	2.52651	0	0	0	0	0	0	0
2	5	20000	CarZZ	0	0	0	0	0	0	0	0	0
3	3	7650	Lexus	0	0	0	0	0	0	0	0	0
4	5	33000	CarZZ	0	0	0	0	0	0	0	0	0
5	5	34000	Lexus	0	0	0	0	0	0	0	0	0
6	4	9200	BMW	0	0	0	0	3.32284	0	0	0	0
7	5	22000	Lexus	0	0	0	0	0	0	0	0	0
8	2	21000	CarZZ	0	0	0	0	0	0	0	0	0
9	2	30000	MBenz	0	0	0	0	0	0	0	0	0
10	4	32000	MBenz	0	0	0	0	0	0	0	0	0

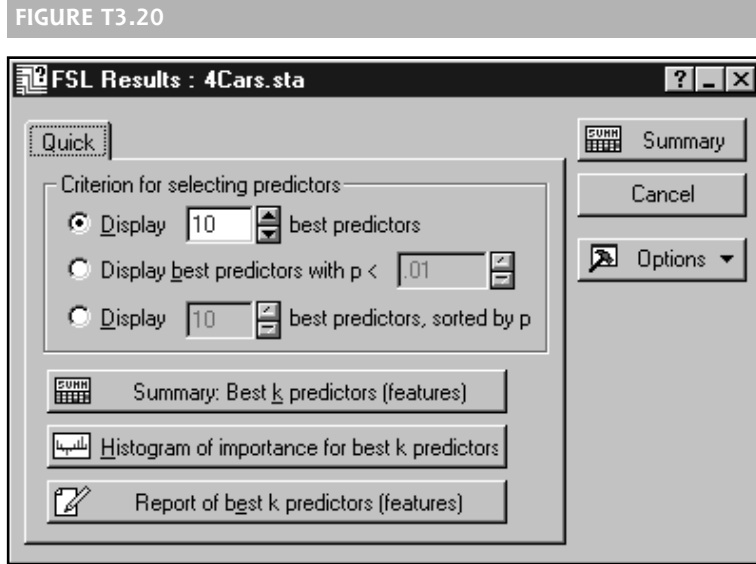
You now have the file at the point where you can try numerous analytic techniques available in the STATISTICA tool set. Let's next look at how to drill down further into specific words of interest. Recall that the first ellipse in the scatterplot contained negative words, and *transmission* was one word that appeared in that ellipse. Let's say you need to find the words that are related to *transmission*. There are several ways you could proceed (e.g., use the feature selection tool, classification trees, correlations matrices). In this case, you'll use the feature selection tool to identify the best predictors for the word *transmission*:

1. Select **Feature Selection and Variable Screening** from **Statistics-Data Mining** menu.
2. Click the **Variables** button and select **transmission** as **Dependent; Continuous**; and all the other extracted words as **Predictor; Continuous**, as shown in Figure T3.19.
3. Click **OK** on the **Select dependent variables and predictors** dialog and the **Feature Selection and Variable Screening** dialog to view the **FSL Results** dialog, shown in Figure T3.20. This dialog contains the following options:

Display k best predictors. At this point, you can request the best *k* predictors; for regression-type problems (for continuous dependent variables), the *k* predictors with the largest *F* values will be chosen; for classification-type problems, the *k* predictors with the largest Chi-square values will be chosen.

FIGURE T3.19



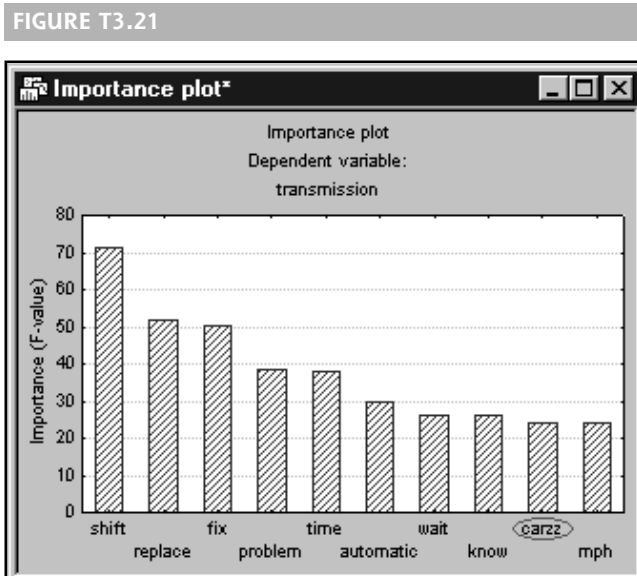


Display best predictors with $p <$. Select this option button to display the list of best predictors for which the p value is less than the value specified in the adjacent edit field. The list of predictors will be sorted in ascending order, by p .

Display k best predictors sorted by p . Select this option button in order to select the k best predictors, based on the probability (p) criterion.

4. In this case, leave the option at the default setting and click the **Histogram of importance for best k predictors** button to view the graph shown in Figure T3.21.

The graph displays the top 10 important words that are related to the word *transmission*. From this histogram, you can tell that when the word *transmission* was mentioned, reviewers also used words such as *shift*, *replace*, *fix*, *problem*, and *time*. The feature selection tool also identified *carzz* as one of the important predictors.



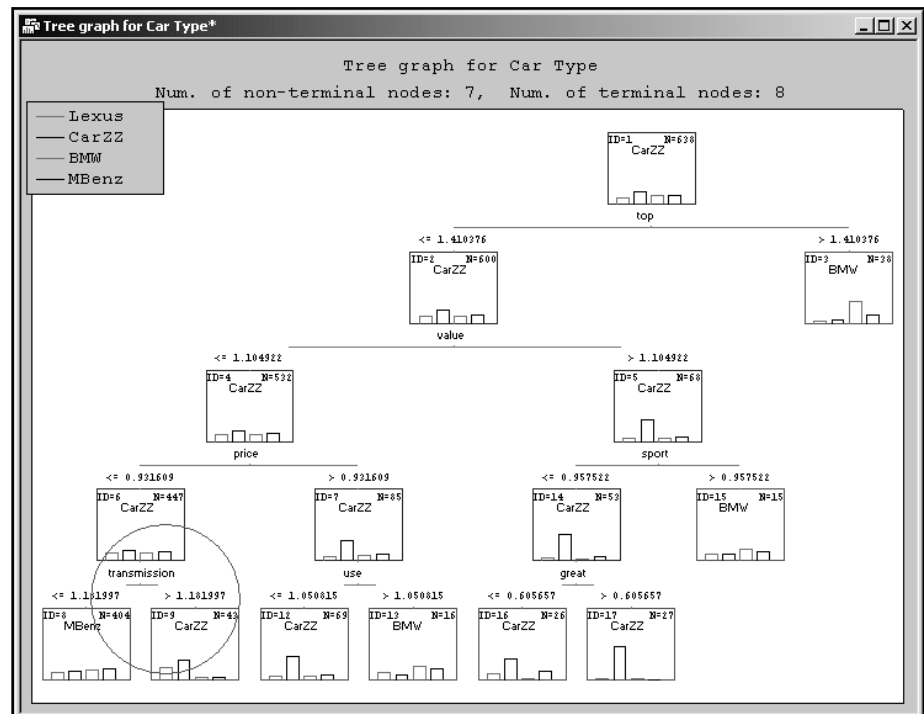
You can also use other techniques, such as classification trees, to see whether you can extract similar useful information, using the prepared file that contains inverse document frequencies.

5. Click the **Grow tree** button to build the tree and then click on button **Remove 1 level** to reduce the tree by one level.
6. Click the **Tree graph** button under the **Review tree** section. You should now see the **Tree graph for Car Type** window, as shown in Figure T3.22.

Interpretation of the tree solutions is relatively simple and straightforward. As you can see from the graph, the C&RT algorithm distinguished eight decision outcomes (contained in eight terminal nodes highlighted in red) built on seven if-then conditions to classify the car type. Terminal nodes (or terminal leaves, as they are sometimes called) are points on the tree beyond which no further decisions are made. The tree starts with the top decision node (also called the root node), with all the 638 cases (reviews, in this case) predominated by the car type **CarZZ** category. **CarZZ** had the highest frequency of reviews among the four car types, as indicated in the histogram. The legend identifying which bars in the node histograms at the nodes correspond to the four categories is located in the top-left corner of the graph.

Recall that the purpose of this analysis is to learn how to discriminate between the different car types, based on the extracted inverse document frequency of words that are used as predictors. The interpretation of this tree is straightforward. The root node is split on the inverse document frequency of the word *top*, forming two new nodes, one with car type **BMW** as the predominant category among 38 cases falling into a terminal node. This simply means that whenever the importance of word *top* is high (i.e., > or higher inverse document frequencies describing the importance of its occurrence),

FIGURE T3.22

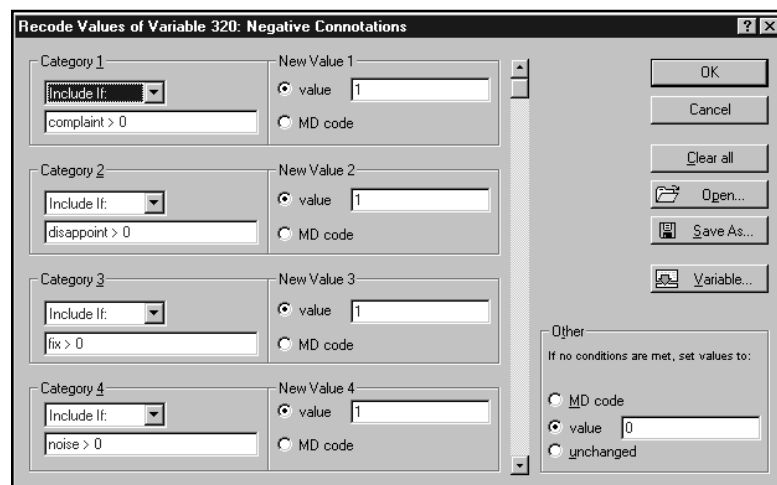


reviewers were mostly mentioning BMWs. Similarly, when the interactions of *word importance* (represented by inverse document frequencies) for *value* and *sport* and for *price* and *use* were high, they were referring to BMWs. You can also tell by looking at this tree that when the words *value*, *price*, *great*, *transmission*, and so on were used, the reviewers were also mentioning *CarZZ*. Therefore, we can tell that *CarZZ* was the brand that had the transmission problem. This gives us further corroborating evidence for the results identified by the feature selection tool.

Another possible approach to make use of this rich textual information would be to recode a new indicator variable for comparative study. For instance, you could create a new indicator variable derived from the negative connotation words (e.g., *complaint*, *disappoint*, *fix*, *noise*, *problem*, *repair*, *replace*) and use crosstabs or interaction plots to compare which brand accumulated the highest number of negative words. To do this, you need to first create a new variable named **Negative Connotations** for this purpose:

1. Make the spreadsheet **4Cars.sta** active (i.e., minimize the result workbook that is open). Then select the **Add variables** option from **Insert** menu.
2. Type **SVDScore16** in the **After:** text box. Name the new variable **Negative Connotations**.
3. Click **OK** to add a new variable named **Negative Connotations** to the spreadsheet. Next, you will use the **Recode** function to create a binary indicator to determine whether the review contain any negative connotations (e.g., *complaint*, *disappoint*, *fix*, *noise*, *problem*, *repair*, *replace*).
4. Select column 320, named **Negative Connotation**. Then select the **Recode** option from the **Vars** menu list.
5. Type the *if conditions* to derive the new variable from the variables holding the negative word. (i.e., if the inverse document frequency of the variables *complaint*, *disappoint*, *fix*, *noise*, *problem*, *replace*, and *repair* is greater than 0, flag the case or review as 1, else 0). Note that you can enter up to 256 conditions in this dialog. Figure T3.23 shows how the **Recode Values of Variable** dialog looks after you enter all the conditions.
6. Click **OK** to perform the recode operation.

FIGURE T3.23



7. Save the file **4Cars.sta** for future requirements. You now see the binary values within the **Negative Connotation** variable indicating whether the reviews contained a negative connotation word. Before you perform the comparative study, you have to make sure that there is an equal number of cases for each car type so that the study is not biased for a particular car type. You first need to decide the number of cases that has to be extracted for each car type. You will next find the frequency of the four categories of car type so that you can identify the car type or brand with the lowest frequency. That number of cases for each car type will then be extracted.
8. Select column 6, **Car Type**, in the **4Cars.sta** spreadsheet. Then select the **Basic Statistics/Tables** option from the **Statistics** menu.
9. Select **Frequency tables** option in the **Basic Statistics and Tables:** dialog and click **OK**.
10. **Car Type** should be already selected within the **Variables** selection dialog. Click the **Summary** button to view the **Frequency table: Car type** box, which is shown in Figure T3.24.

The results table shows that Lexus had the least number of reviews compared to the other car types. Therefore, you need to extract 119 cases for each car type for comparative study.

11. Minimize the active workbook that holds the frequency table results.
12. Select the **Subset/Random Sampling** option from the **Data** menu. Click the **Options** tab and then select **Calculate based on approximate N**.
13. Click the **Stratified Sampling** tab. Click the **Strata Variables** button and select the variable **Car Type**. Click **OK** to make the selection.
14. Click the **Codes** button. Click the **All** button and then click **OK**. You now see all the categories of car types in the **Stratification Groups** column.
15. Enter **119** against each strata under the **Approximate N** column to extract a sample set with 119 brands for each category. You should now see the dialog shown in Figure T3.25.
16. Click **OK** to view the new stratified sample spreadsheet that holds an equal number of cases for each car type.

If you now check the frequency for variable **Car Types**, you will have a balanced spreadsheet with equal number of categories. You can use this spreadsheet to draw an interaction plot to perform a comparative study:

1. Select **Basic Statistics/Tables** from the **Statistics** menu. The dialog shown in Figure T3.26 appears.
2. Select **Tables and banners** option from the **Basic Statistics and Tables dialog** and then click **OK**.

FIGURE T3.24

Category	Count	Cumulative Count	Percent	Cumulative Percent
Lexus	119	119	18.65204	18.6520
CarZZ	218	337	34.16928	52.8213
BMW	148	485	23.19749	76.0188
MBenz	153	638	23.98119	100.0000
Missing	0	638	0.00000	100.0000

FIGURE T3.25

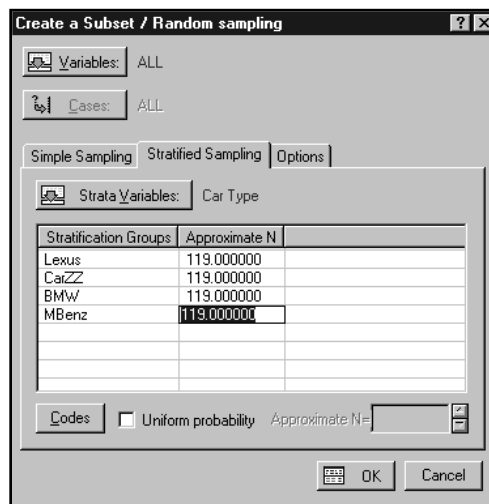
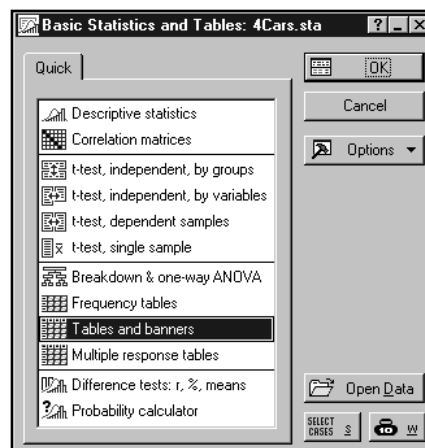


FIGURE T3.26



3. Click the **Specify tables (select variables)** button on the **Crosstabulation Tables** dialog.
4. Select the variable **Car Type** in the **List1:** section and **Negative Connotations** in the **List2:** section, as shown in the **Select up to six lists of grouping variables:** dialog (see Figure T3.27).
5. Click **OK** on the Variable Selection (above) and **Crosstabulation Tables** dialog to view the **Crosstabulation Tables Results:** dialog. Select the **Advanced** tab to view the options shown in Figure T3.28.
6. Click the **Interaction plots of frequencies** button to view the **Interaction Plot: Car Type X Negative Connotations** plot, as shown in Figure T3.29.

From this graph, you can tell that **MBenz** accumulated the largest number of reviews containing negative connotations (around 50 or so), followed by **CarZZ**, **BMW**, and **Lexus** (according to category 1, which represents the reviews that have neg-

FIGURE T3.27

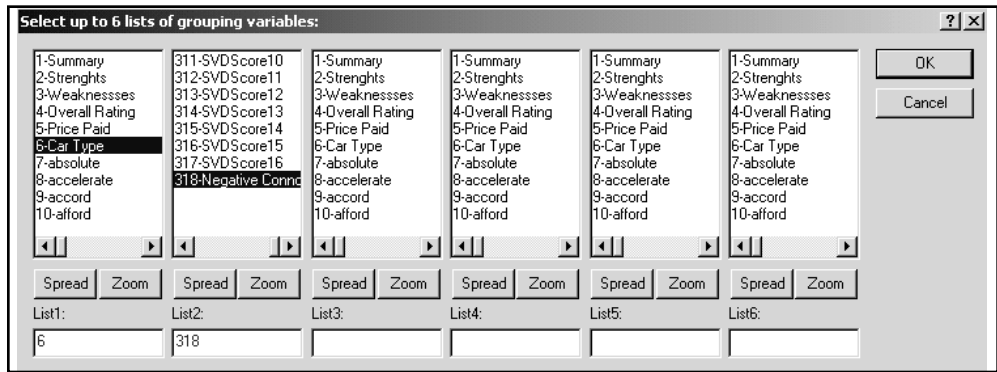


FIGURE T3.28

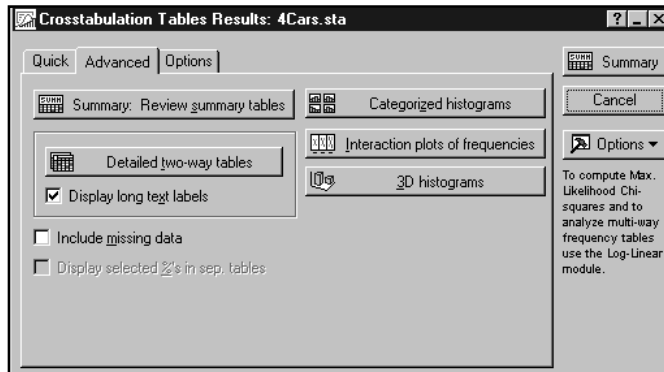
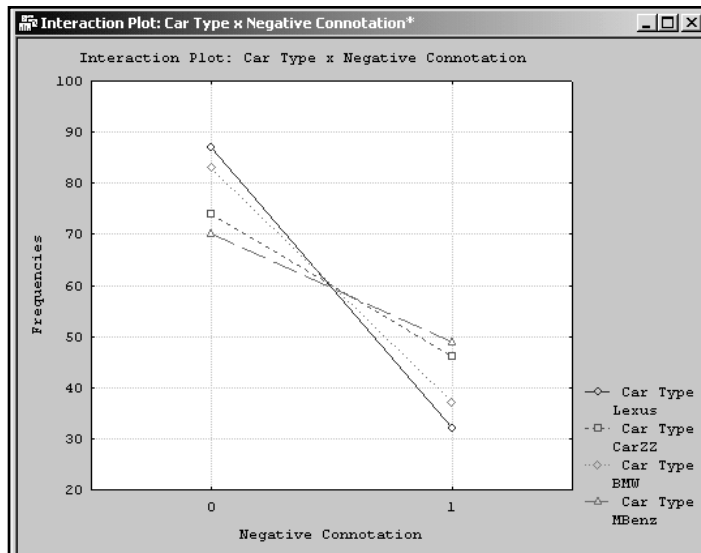


FIGURE T3.29



ative connotations). You have identified, by using this simple approach, that Lexus had the fewest number of negative connotation words compared to the other car types. If you had more information about the state, the city, the manufacturing unit for each car/brand, and so on, you extract useful information to identify the places/units that elicited the largest numbers of complaints.

CONCLUSION

This simple illustration helps you understand how the STATISTICA Text Miner module, along with numerous STATISTICA Data Miner tools/techniques, can be used to find solutions for problems that require knowledge of language and computing technology. More importantly, you have seen how extraction of useful insights/information from unstructured data can be used as inputs for decision-making purposes. The STATISTICA system is particularly well suited for these purposes because of the seamless integration of all components of its data and text-mining facilities.