Text Mining SAS Course Descriptions (Combination Exercise and Self-Guided Demonstration)

This exercise is intended to show how useful information retrieval and text mining can be for activities such as call routing. In this example, there is no target variable and the emphasis is on information retrieval and document categorization.

SAS Education supports more than 300 courses. Prospective students often have questions about curriculum and specific course content. For example, a prospect might ask for information about courses that discuss neural networks. Text mining provides a solution for automating queries based on keywords.

Descriptions of the SAS courses can be found at [http://support.sas.com/training](http://support.sas.com/training).

The SAS course descriptions data set `DMTX51.SASCOURSES` contains descriptions of courses supported in 2011. The data set has 735 rows (documents) and four columns. Some courses have multiple versions that are associated with different releases of the software. The metadata is shown below.

NEW: The necessary SAS data tables for this problem are on the SAS server, and they have to be accessed through a library (because that is how this functionality is implemented in the software). So you must first create a new library when you create a project. Steps are as follows.

1. From the top menu, select File -> New -> Library.
2. Select ‘Create New Library’ in the pop-up window, and hit Next.
3. In the next window, enter WorkBook for the Name, and for Library Information, navigate to oda… -> tutorials -> data -> studio, and click on the folder dmtx51, as shown below. Then click Open.
4. Accept the rest of the steps to finish.

Now in the project, you can access SAS data tables in this library by selecting **WorkBook** as the library name. In particular, wherever a reference is made to **DMTX51**..., you can browse/select the library WorkBook and find the tables there.

The variable **CourseOutline** contains the course outline text.

The following flow diagram implements the text mining analysis:
Frequency filtering is a methodology to create or add to a stop list. You can run the Text Parsing node with the default stop list and then use frequency filtering to add terms to this list. Frequency filtering specifies a cutoff frequency. Terms with a frequency below the cutoff are added to the stop list. You can also specify a cutoff frequency at the high end so that terms with a frequency above the cutoff are added to the stop list. For creating a start list, keep terms with frequencies between the high and low cutoff values. The start list DMTX51.SASCOURSESESTART contains a start list that was obtained using domain knowledge and frequency filtering.

a. Create a diagram named SAS Course Outlines. If you need to, create a data source for the DMTX51.SASCOURSES data set. (The metadata is presented above.) Drag this data source into the diagram.

b. Attach a Text Parsing node to the Input Data Source node. Change the Synonyms property so that there is no synonyms table and add DMTX51.SASCOURSESESTART as a start list.

c. Attach a Text Filter node to the Text Parsing node. Change frequency weighting from Default to Log. Change term weighting from Default to Inverse Document Frequency. Log is the default frequency weight, but Entropy is the default term weight. Inverse Document Frequency is recommended for documents larger than a paragraph. Run the Text Filter node.

d. Open the Filter Viewer (also called the Interactive Filter Viewer). Determine how many documents contain the term neural network by doing a search on this term. How many documents did your search return? Why is this number not 23?

e. Select the document corresponding to the course with the code BASEL52. Select Edit ⇪ Toggle Show Full Text. You can read the course outline for BASEL52.

f. Select Clear ⇪ Apply to return all of the documents in the collection. Navigate back to the neural network row in the Terms table. Right-click on the neural network TERM cell and select View Concept Links. The concept link plot appears. What are some of the terms strongly associated with neural network?

g. Close the Filter Viewer. Attach a Text Topic node to the Text Filter node. For User Topics in the Property Sheet, select the data set DMTX51.SASTOPICS. Keep all of the other default settings. Run the Text Topic node.

h. Access the Results window. Which topic contains the most documents?

i. Close the Results window for the Text Topic node. Look at the exported data and determine what variables were created by this node.

j. Open the Topic Viewer from the Property Sheet and explore the results.

A custom topic is similar to a predefined query. The topic weight shown in the documents window determines whether the topic is present. (That is, the query is satisfied.) If the topic weight exceeds the document cutoff, then the document is classified as having the topic.
k. Close the Topic Viewer. Attach a **Text Cluster** node to the **Text Filter** node. Most users attach the Text Topic node and Text Cluster node directly to each other, but they work independently. Neither requires any results from the other.

l. Use the default setting and run the **Text Cluster** node. Open the Results window. How many clusters were created? Can you interpret some of the clusters from the displayed descriptive terms? How many SDV variables were created?