**INFO 4360—COMPLEX DATA ANALYTICS**

**TEXT MINING IN SPSS**

**Lesson 10 Homework —Different Input Formats**

SPSS Modeler will accept text data feeds from a variety on inputs. So far we have only used Excel files and VAR (.txt with variable field lengths) inputs. We will now explore several other input formats.

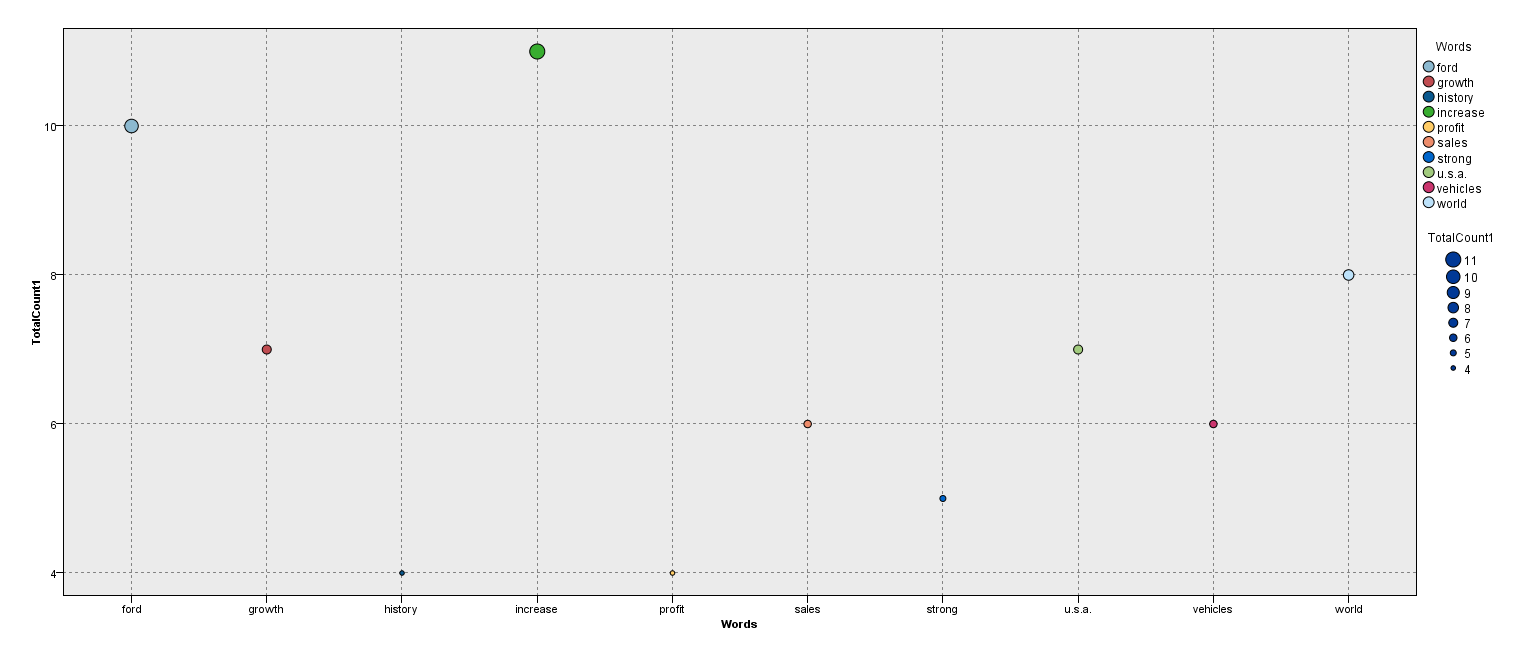
**HOMEWORK**.

Responses to these questions should be complied in a Word Doc that includes all of the graphical and text responses to the following prompts. There is no need to use anything beyond the Basic English Resource Template for all of these requirements.

1. **.TXT** **Files**. Analyze the Ford and GM annual reports (INFO4360DataAnnRptFord.txt and INFO4360DataAnnRptGM.txt) to compare and contrast the two companies. Use any output (graphic, table, counts, etc.) that you think will make an appropriate comparison. (You only need to make a single comparison.) What are the “documents” that SPSS scores within this file?

**Answer**:

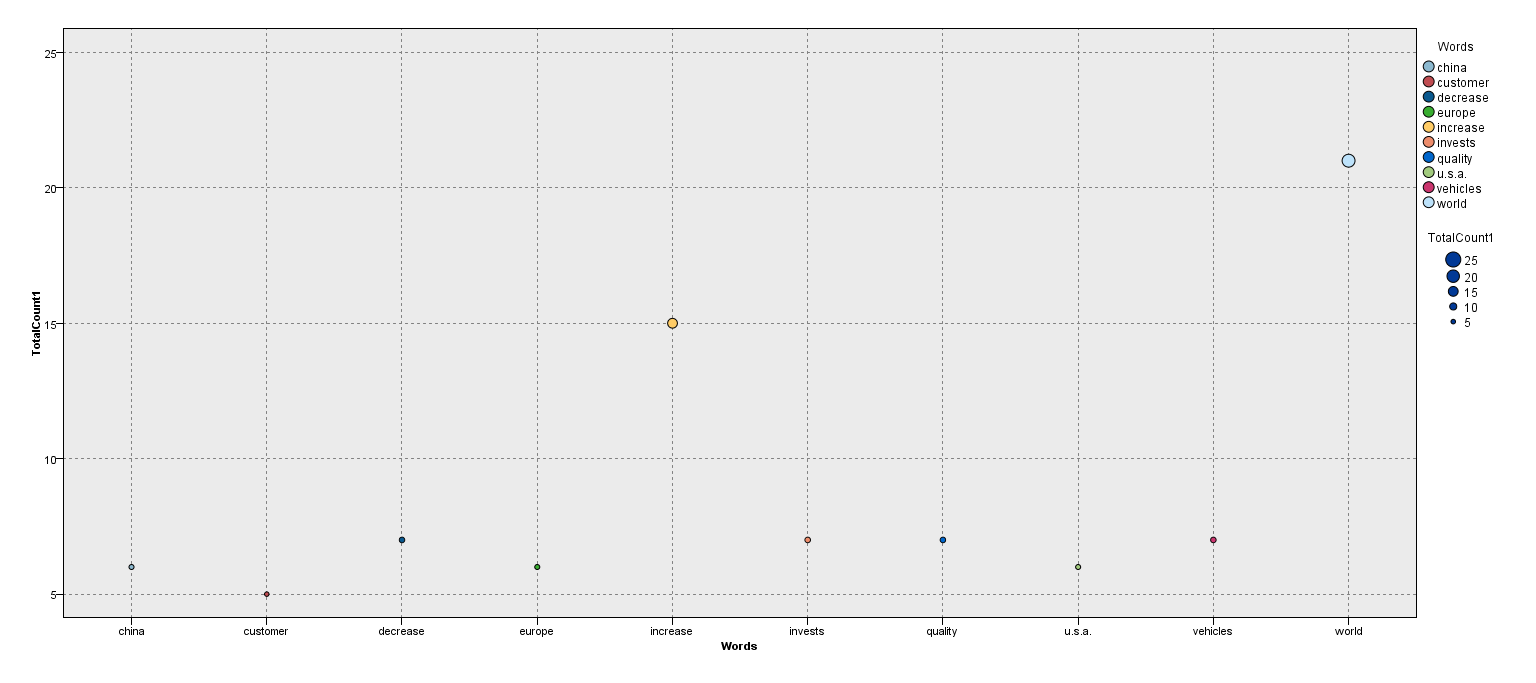
Please find below the top 10 words (based on their frequency) in Ford’s annual report:



Document Scored:



Please find below the top 10 words (based on their frequency) in GM’s annual report:



Document Scored:



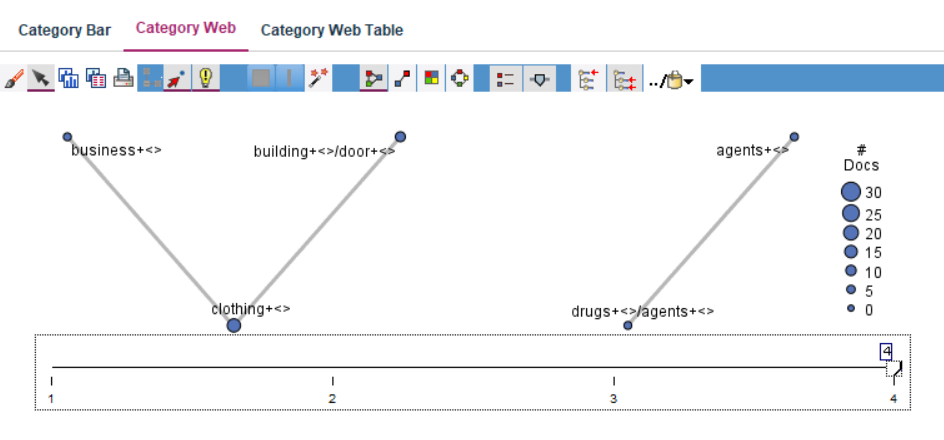
Stream File used in this exercise:



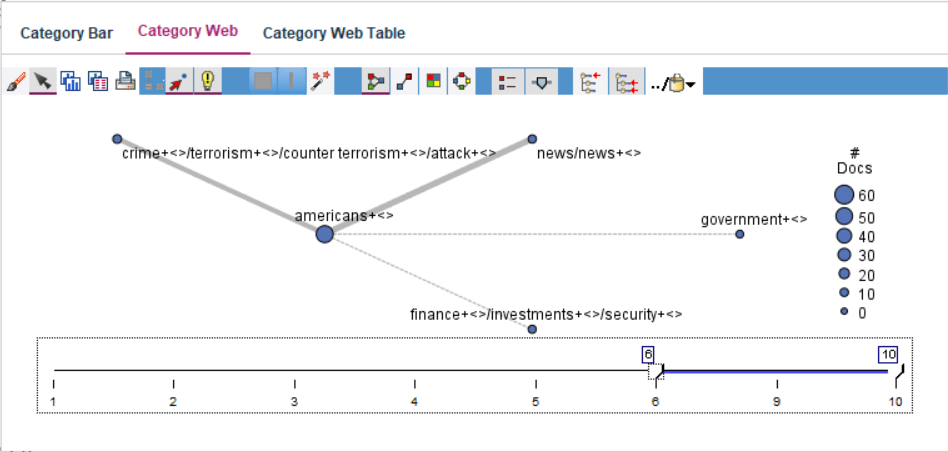
2. **WORD Docs**. Analyze the novel (INFO4360DataNovel.docx) for Concept and Category relationships. Show graphically three Category Webs from the analysis of this novel ) appropriately filtered). You will have to save the document in a different format (i.e., .txt) for SPSS to read it. What are the “documents” that SPSS scores within this file?

**Answer**:

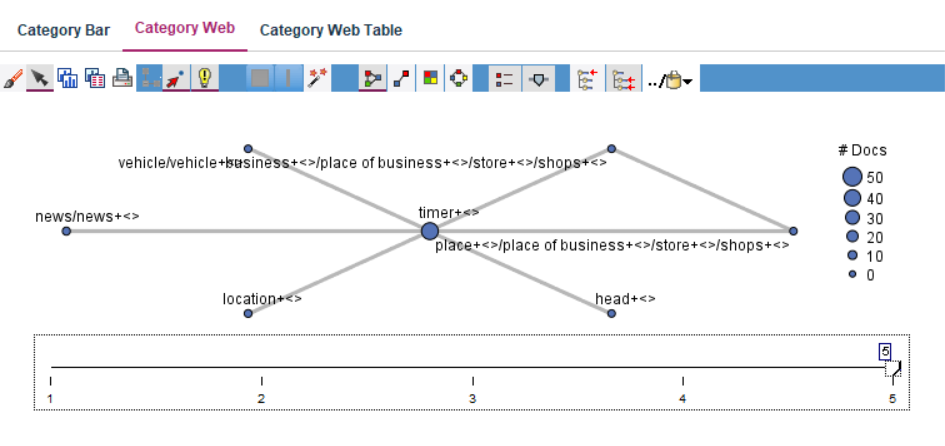
Category Web: Clothing



Category Web: americans



Category Web: timer



Document Scored:



Stream File used in this exercise:



3. **Formatted .TXT Files**. Analyze the 71 Seattle hotel ratings (INFO4360DataHotel.txt) for only the <Content> field that contains the customer comments. Create a Graph of those Concepts that are most frequently cited (i.e., 5 or more instances). Did you accomplish this entirely in SPSS Modeler (if so, how?) or did you manage the data elsewhere (if so, where?)?

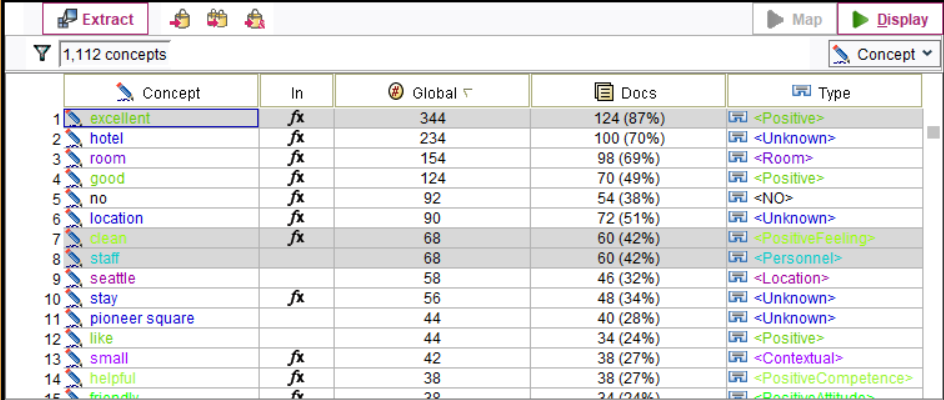
**Answer**:

Before processing the file in SPSS, I converted the file to a properly formatted XML file in Notepad++ using its “XML Tools” plugin & also used regular expressions to restore the Closing tags in this document. Please find below the modified document:

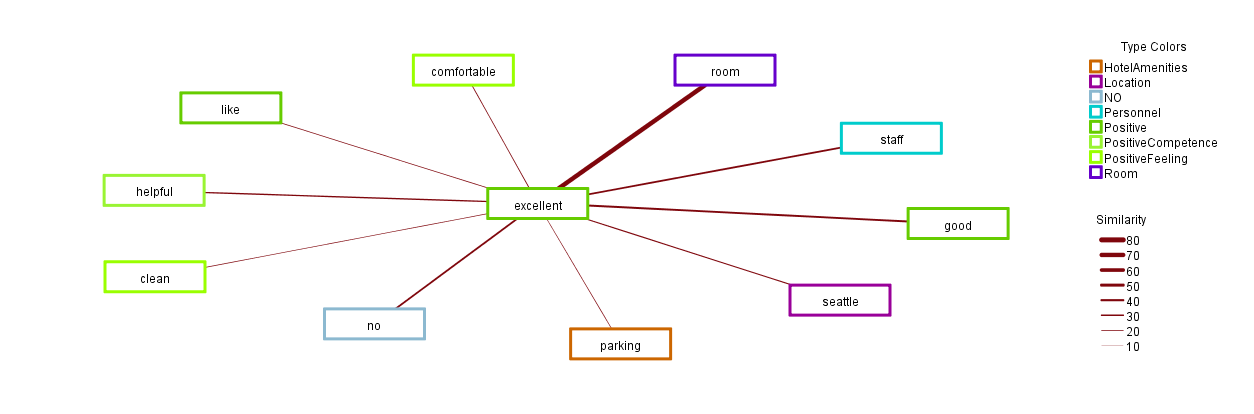


Then I used the XML node under Sources in SPSS to extract the information from this file.

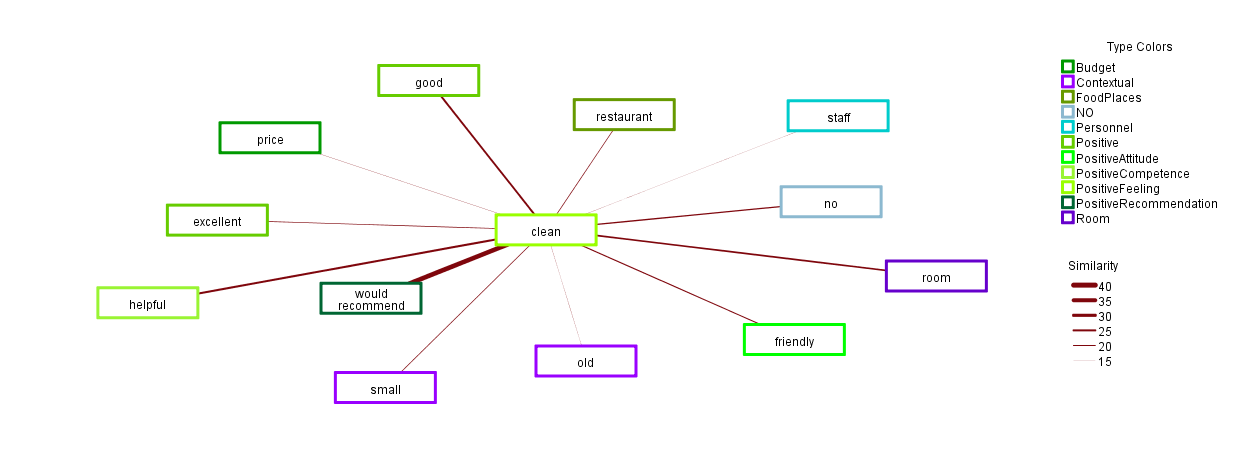
Please find below the concepts that were generated for this data:-



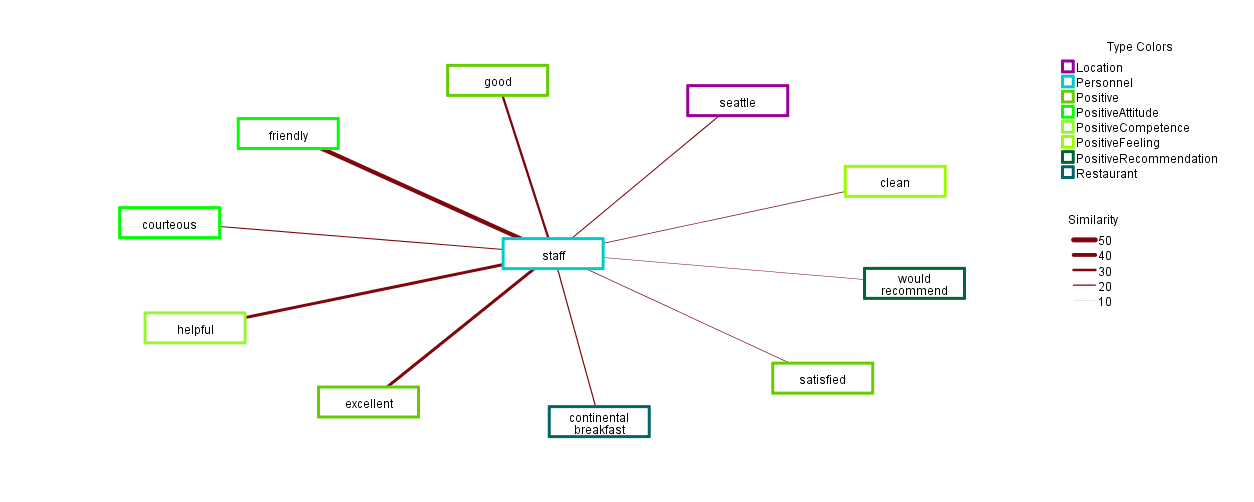
Category Map: excellent



Category Map: clean



Category Map: staff



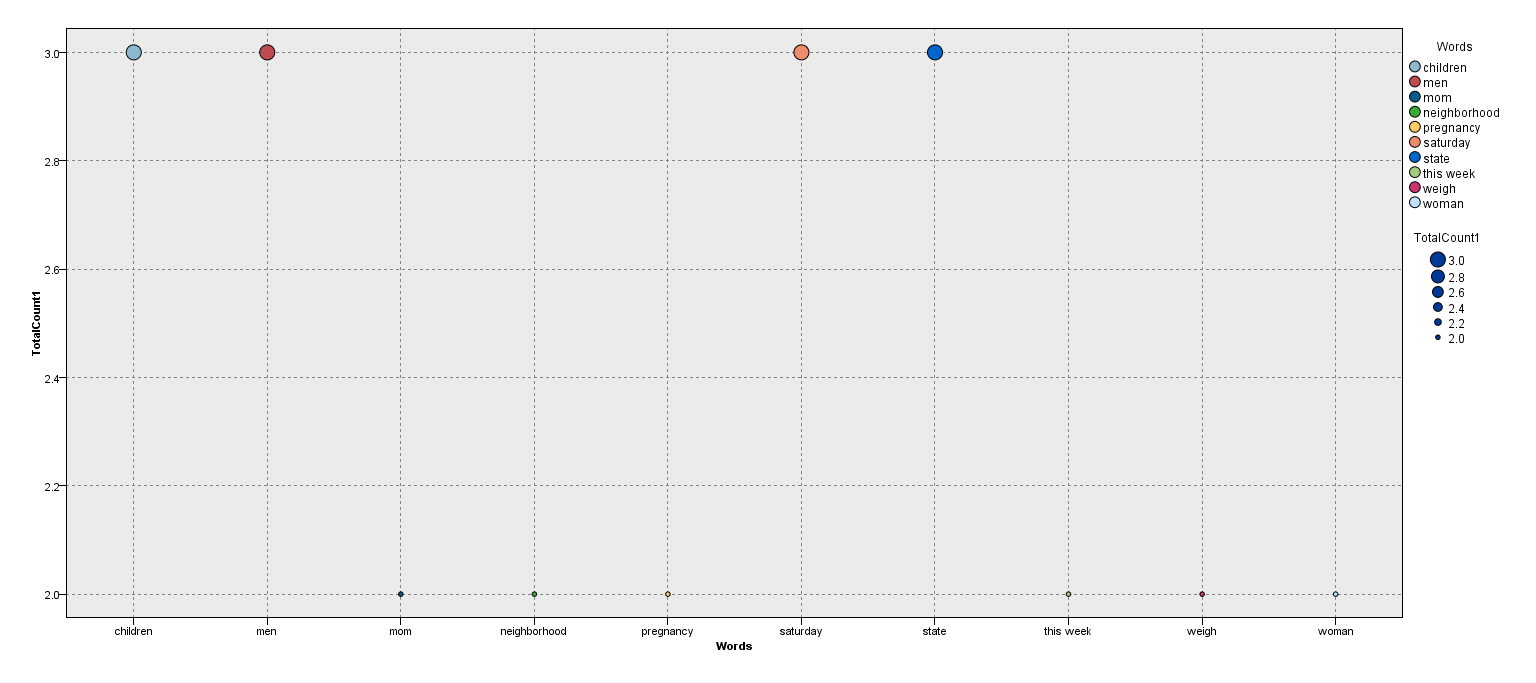
Stream File used in this exercise:



4. **RSS Feeds**. Go to the CNN site and use their RSS feed ([http://rss.cnn.com/rss/cnn\_topstories.rss (Links to an external site.)Links to an external site.](http://rss.cnn.com/rss/cnn_topstories.rss)).  Grab the latest stories and feed them into SPSS Modeler Text Analytics.  (Use the Web Feed node under the Text Analytics pane.)  Compile a list of top keywords with their frequencies, from these news feeds under the "Short Description" field.

**Answer**:

Please find below the top 10 words along with their frequencies obtained form the above link:



Stream File used in this exercise:

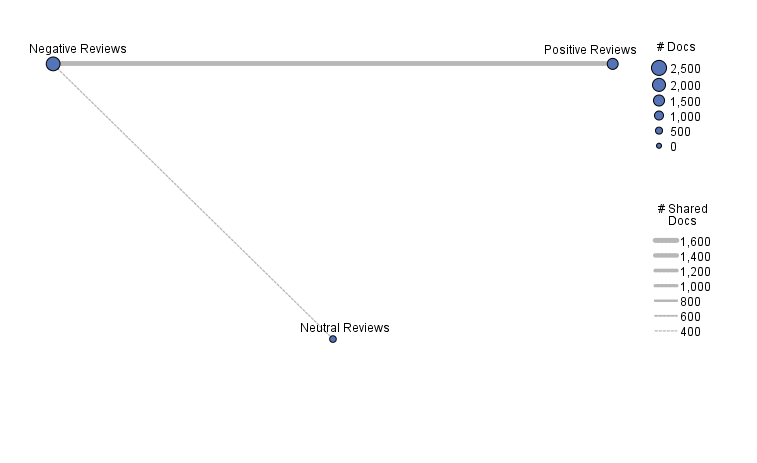


5. **FILE Feeds**. Use the File List from the Text Analytics pane to load the files in the Movies folder (on a flash drive). You can attach the Text Mining node directly to this input node, but you must specify that the inputs are path names. Ensure you include subdirectories. Create three Category Webs from the data, appropriately filtered. What text do these folders and files contain?

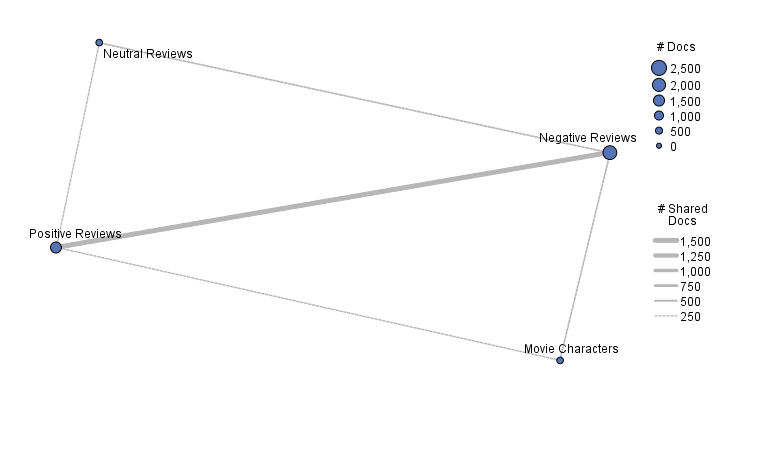
**Answer**:

These folder and files contain positive, negative and neutral comment regarding movies.

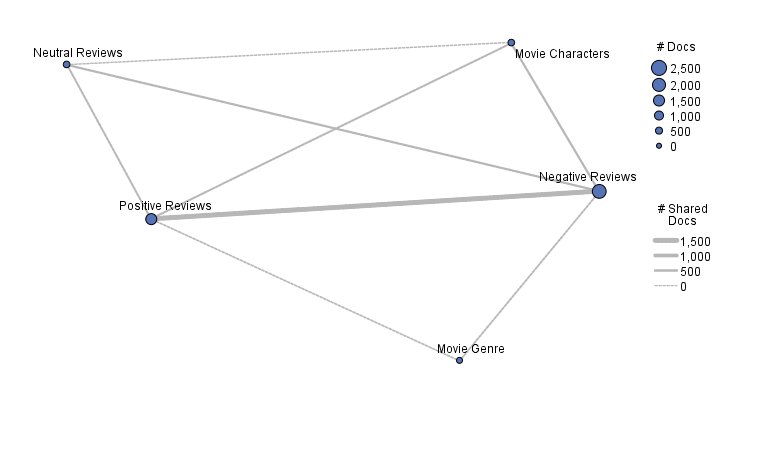
Category Web - 1



Category Web – 2



Category Web - 3



Stream File used in this exercise:



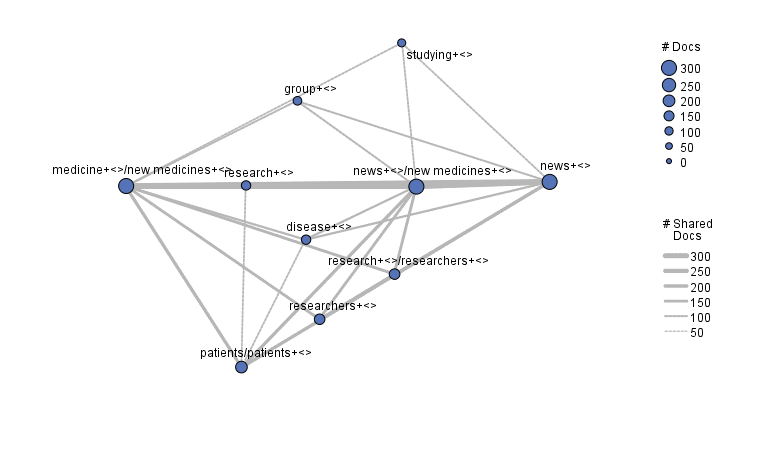
6. **FILE Feeds (Links)**. Use the File List from the Text Analytics pane to load the files in the BMS folder (on a flash drive). You can attach the Text Mining node directly to this input node, but you must specify that the inputs are path names. Create a graph of the top 20 Categories that appear in these files. What are contained in these files? (They’re links, but links to what?)

**Answer**:

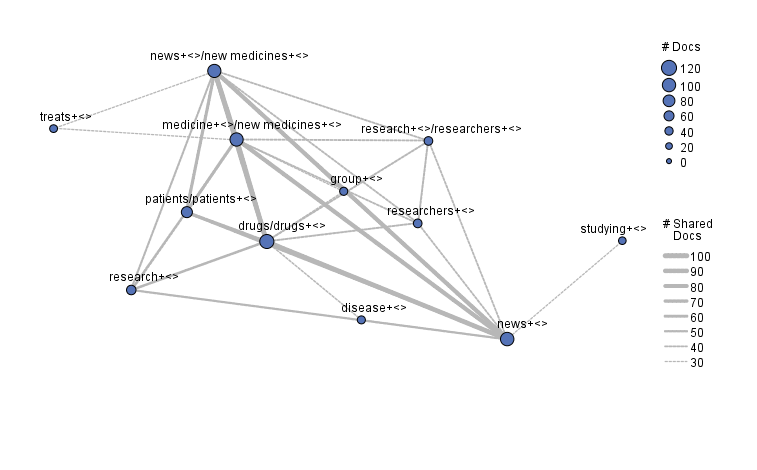
These files contain medical journals or are reports on the medical field.

Please find below category webs containing the top Categories.

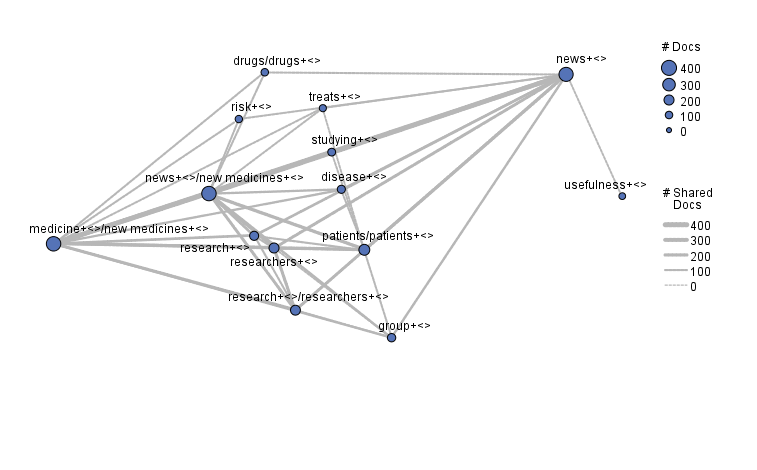
Category Web - 1



Category Web - 2



Category Web - 3



Stream File used in this exercise:

