Demo with Android App Reviews: Text Cluster and Text Topics

Text Cluster and Text Topic Using Android App Reviews

Case Study: Android App Reviews (This case study has been published by the author as a paper in the Proceedings of the 2013 SAS Global Forum.)

Data

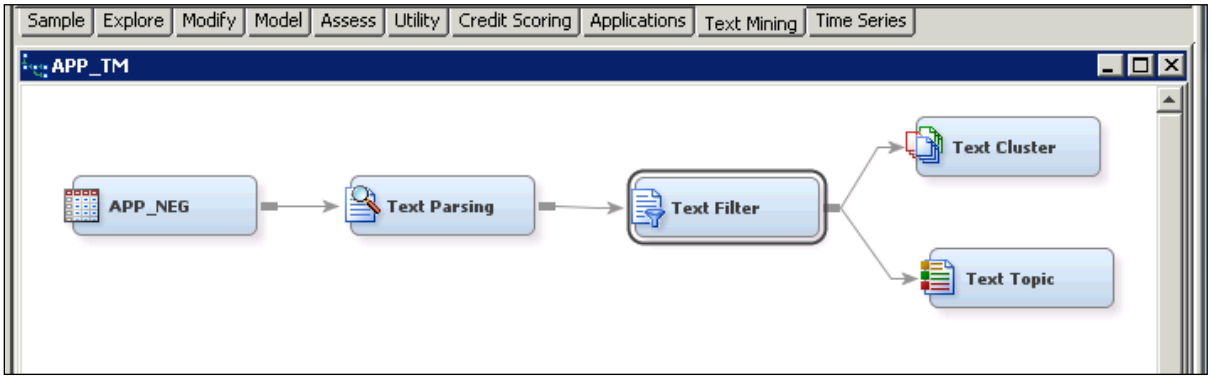
We have created an artificial data set of 600 reviews by modifying and anonymizing actual customer reviews posted online. Of these reviews, 500 are used for building models, and the remaining 100 are used for testing models. Raw textual data has been categorized into positive and negative groups based on 5-star numerical ratings given by a consumer on the review site at the time the review was written by the same consumer. Comments greater than or equal to 4 stars are considered as positive, and those less than or equal to 2 stars are considered as negative for the purpose of this case study. For modeling, we have two data sets for text analytics.

- Negative reviews: APP_neg.sas7bdat
- Positive reviews: APP_pos.sas7bdat

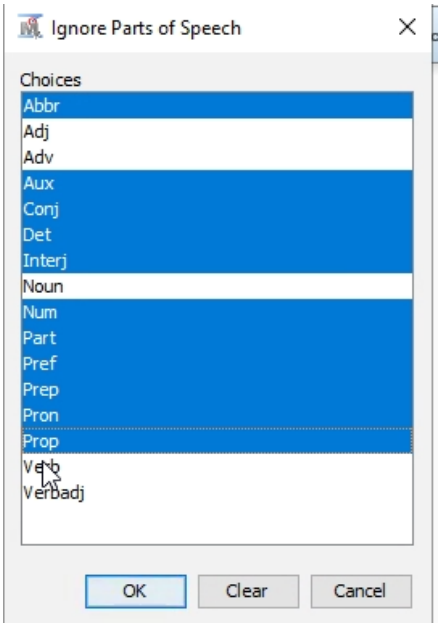
1. Start a new project in SAS Enterprise Miner or, work with an existing project.
2. Create a new library (**File** ⇨ **New** ⇨ **Library**) and give it an appropriate name (**COURSE**). Point to the folder that stores data files.
3. Right-click **Data Sources** in the project panel to create two data sources with the two data files (positive and negative reviews) from the **Course** library. Use all default steps when creating the data source. Make sure that data roles and levels are as shown below. Both data files have two columns, **ID** and **Text**. Column **ID** is a unique nominal variable to identify each textual comment. The column **Text** contains the actual reviews.

Name	Role	Level
id	**ID**	**Nominal**
text	**Text**	**Nominal**

4. Create a diagram by right-clicking **Diagrams** in the project panel and give it an appropriate name. Drag the **APP_NEG** (negative reviews) data source into the workspace. Then add Text Parsing, Text Filter, Text Topic, and Text Cluster nodes and attach to each other as shown below. **Retain the default options for all nodes unless specifically mentioned below.**



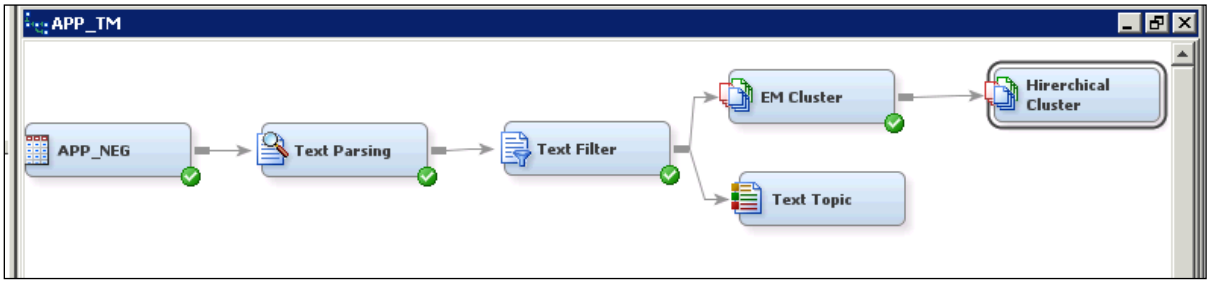
5. In the Text Parsing node, find the **Ignore Parts of Speech** option on the properties panel and click the ellipsis button. Press the Ctrl key and select **abbr**, **aux**, **conj**, **det**, **interj**, **num**, **part**, **pref**, **prep**, **pron**, and **prop** (as shown below). Click **OK** to save and exit.



6. In the properties panel of the Text Cluster node, change the value of Max SVD Dimensions to **40**. Singular Value Decomposition (SVD) is used to reduce dimensionality by converting the term frequency matrix into a lower dimensional form. Smaller values of k (2 to 50) are thought to generate better results for text clustering using short textual comments such as the ones used in this case study. Also, change the value of Descriptive Terms to **8**. This will display eight terms to describe each cluster in the results.
7. Right-click the **Text Cluster** node and select **Run**. Examine the results.

   Many clusters are generated. But, some of the clusters have very few documents suggesting we may need to force the algorithm to a smaller number of clusters. The descriptive terms from each cluster give us a sense of the characteristics of the grouping of comments.

   In the next few steps, we will run hierarchical clustering on the same data to get a sense of how the groupings change as we apply different techniques on the same data.
8. Right-click the **Text Cluster** node and select **Rename**. Type **EM Cluster** in the rename box.
9. Right-click **EM Cluster** and select **Copy** and **Paste**. In the pasted node, **change** the cluster algorithm to **Hierarchical**, rename the pasted node as **Hierarchical Cluster**, and connect it to the EM Cluster node.



10. Right-click **Hierarchical Cluster** and select **Run**. Examine the results.

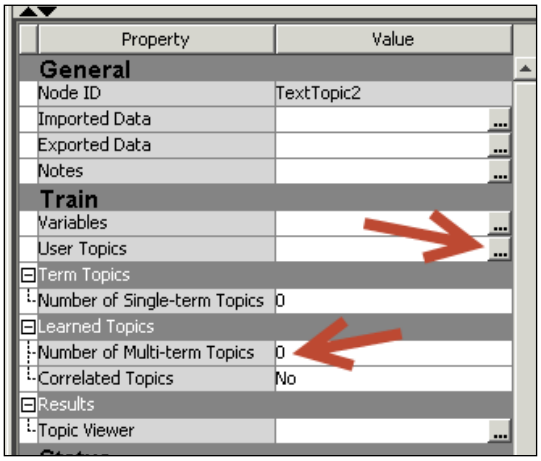
    There is a lot of information in the panels of hierarchical clustering results. In particular, look at the *Hierarchy Data table* and the Cluster *Hierarchy diagram* to understand how clusters are formed. For example, Cluster ID 2 and 5are split from Cluster ID 1 that has all 250 observations.

Then, Cluster ID 2 is split into Cluster ID 3 and 10, and so on. Again a large number of clusters are produced some with few observations, but the descriptive terms for these clusters differ from the descriptive terms from the EM clusters. This happens because which review is assigned to which cluster depends on the algorithm.

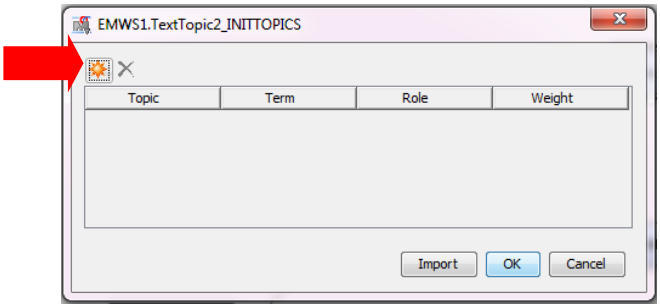
At the end of the day, as an analyst you will have to decide which is most useful solution (recall the purpose of doing clustering is to generate insights).

11. Close **Hierarchical Cluster** results.
12. Right-click the **Text Topic** node, and select **Run**. This runs the node with default settings. Examine the results.

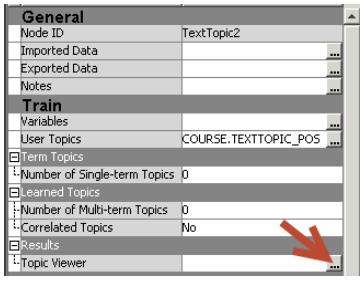
    You will find that features such as Weather, Update, Time/ Clock, or Money/ Price appear several times in the 25 topics generated. Combining what we have seen earlier from text clustering and what we see now in the topics, we will modify and customize the topics to narrow these down to a handful of relevant topics.
13. Create customized user topics. Right-click and rename the current Text Topic node as **Text Topic (Default).**
14. Add another Text Topic node and attach it to the Text Filter node. Right-click and rename this as **Text Topic (Custom)**. In the properties panel of the Text Topic (Custom) node, change the value of **Number of Multi-term topics** to **0**, because we want to generate customized user topics and we do not need SAS Enterprise Miner to do this for us automatically. Click the ellipsis button (highlighted) next to **User Topics**.



15. In the pop-up window, you can create customized topics that you are interested in. To add a new topic, click the orange star button. To delete an existing and unwanted item, select the row that you want to remove and then click the cross button.



16. In this case study, for the negative reviews, we will ***create eight customized topics***. We can manually add topics by typing in topic name and term as shown below. Here, each Topic represents a unique topic; each Term is used as a value of each Topic. For example, battery and battery usage are both treated as values for the topic battery life. Weight indicates the importance of each term within its topic. In this case study, we give the same importance to each term. Hence, all weights will be equal to **1**. In practice, you should experiment with different weights based on your domain expertise and then test to see how those work. But, instead of typing all of the topic terms, click **Import** and select the file **texttopic_1** in your library.
17. Click **OK** to save and exit. Right-click **Text Topic node** and select **Run**. Select **Results** to open and see user-defined topics along with the number of documents. You will find that the newly created topics generally are in many documents with the exception of Privacy, which has a frequency of 5 in #Docs.
18. To get a better understanding of these topics and how they are related to the terms, you can use the interactive topic viewer by selecting the ellipsis button in the properties panel next to **Topic viewer** in the Results section.



The interactive topic viewer enables users to select each topic and find the terms and documents that relate to those topics.

You should explore these results on your own to get a better sense of how the customized topics work. Start by clicking on a topic in the top panel (such as GPS or Privacy) and see how the terms and documents change in the middle and the lower panel.