Week 13 Assignment

Note: Late homework assignments are not accepted because the solutions are discussed at the start of every class on Tuesday. Your solution to this assignment must be uploaded on eCampus as a PDF file. A common approach is to put the solution into a word document and then save that into a PDF file. Please only submit PDF files.

Assignment: You can complete one of the two parts to this assignment, or both. If you do both parts, you can obtain extra points for this assignment. You are expected to complete at least one part successfully.

Data Files: Excel File <*Hotels..xlsx*> A file of customer reviews for 5 hotels in Las Vegas

doc:	An ID number for each review (1 to 1672)
hotel:	The Name of the Hotel: Bally's, Bellagio, Circus Circus, Encore, Excalibur
review:	The customer's review of their stay at the hotel

There are no outliers or missing values in these data.

The objective in this analysis is two-fold. First conduct a topic analysis and store the topic number in the data. Next conduct a sentiment analysis for each review. Merge the sentiment calculations with the topic clusters and report the average sentiment by: topic cluster, hotel and cluster x hotel.

Part 1: Create a SAS EM project names "Week 13 Homework". In that project read this data file using the file import node from the text mining tab in SAS EM.

1. Conduct a text classification analysis of these data following the analytical process described in class.
2. Use POS, stop words and stemming for building the term/doc matrix.
3. Use TF-IDF weighting for the term/doc matrix.

4. Use the text cluster node and set the SVD resolution to "high" for this analysis. Obtain the clusters and save the data file into a SAS file called "HotelClusters". The file should contain the original data plus the cluster assignments. Do not bother saving any of the other attributes produced from the text cluster node.
5. Next prepare a new diagram to compute the sentiments for each document and store those in a file called "sentiment". That file should have the document number and the sentiment for each document.
6. Merge the HotelClusters file with the sentiment file. Merge on document number.
7. From the combined file report the average sentiment by topic, the average by hotel, and then a cross table of both hotel and cluster.

REPORT:

1. Screen shot of the diagram
2. Screen shot of the text cluster terms, the 15 words that describe each cluster and the number of documents assigned to each cluster
3. The average sentiment for the entire corpus
4. The average sentiment for each text cluster
5. The average sentiment for each hotel
6. The average sentiment for each hotelxcluster combination

The average sentiment for 4-5 are easily produced using something like the following SAS code

```
PROC TABULATE  data=&EM_IMPORT_DATA;
CLASS  hotel textcluster_cluster_;
VAR  docScore;
WEIGHT ndocs;
TABLE hotel, docScore * (MIN MEAN MEDIAN MAX N PCTN);
TABLE textcluster_cluster_, docScore * (MIN MEAN MEDIAN MAX N PCTN);
TABLE hotel * textcluster_cluster, docScore * (MIN MEAN MEDIAN MAX N PCTN);
```

Part 2: Do the same assignment as Part 1 using Python:

1. Copy of your code
2. Description of seven topics extracted using TFIDF with LDA. Also set max_df=0.7 and min_df=4. Be sure to look for synonyms, there are two that I found, and also additional stop words, I found many that appear in the topic list.
3. Overall average sentiment (weighted)
4. Average Sentiment by Hotel
5. Average Sentiment by Clusters (7)
6. Average Sentiment by Hotel X Clusters

Also add the following Word Clouds:

7. Word Cloud for the words in all of the reviews.
8. Word Cloud for only the Sentiment words in all of the reviews.
9. Word Cloud for the words in the topic cluster with the lowest sentiment score.
10. Word Cloud for the Sentiment words in the topic cluster with the lowest sentiment score.
11. Word Cloud for the words in the topic cluster with the highest sentiment score.
12. Word Cloud for the Sentiment words in the topic cluster with the highest sentiment score.