Week 10 Assignment

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 <*CaliforniaCabernet.xlsx*> -> contain over 13K reviews of California Cabernet Sauvignon. The reviews are in the column labeled 'description'. The full data dictionary is:

review: A number unique for each review (an ID)
description: The actual review (text)
year: Year the wine was bottled. This is missing for some wines.
points: The points assigned by the reviewer to the wine. These range from 80 to 100. Better reviews have higher points.
price: The retail price for a bottle of the wine ($0-$3000).
winery: The winery where the wine was bottled. (a text label)
Region: Region of California (text) where wine was produced.

There are no outliers in these data, but many of the years are missing.

Part 1: Create a SAS EM project names "Week 10 Homework". In that project read this data file using the file import node from the Sample tab in SAS EM, like reading any other Excel file. Check to ensure the "Role" of the column "description" is set to "Text" with "Level" as "Nominal".

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 text-cluster node to develop exactly 9 clusters for the wine reviews.
5. Develop a table showing the average points and average price for wines in each cluster. Include the 15 words that describe the clusters.

REPORT:

1. Report a screen shot of the diagram and the property windows for all nodes in the diagram.
2. Table of average points and price for each topic group.

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

Follow the process described in the week 9 notes. Use pandas to read the file. Use LDA to identify the top 9 topics.

Analysis:

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. These are the default condition in *TextAnalytics.my_analyzer*.
3. Use TF-IDF weighting for the term-doc matrix.
4. Use LDA to develop exactly 9 clusters for the wine reviews.
5. Develop a table showing the average points and average price for wines in each cluster. Include the 15 words that describe the clusters.

REPORT:

1. The Python code (.py file).
2. Table of average points and price for each topic group.