Week 9 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:
TextFiles -> A directory of 8 text files, each a different book. SAS and Python will use the same data files.

Part 1: Create a SAS EM project names "Week 9 Homework". In that project read this data file using the file import node from the text mining tab in SAS EM. Construct the term/document matrix for the following four scenarios using the Parse node in SAS EM.

for scenario 2
stemming doesnt make sense
if u you dont use pos
coz it uses noun verb tag
for stemming

Scenario	Remove Stop Words	POS	Stem
1	Yes	Yes	Yes
2	Yes	No	Yes
3	Yes	No	No
4	No	No	No

Tokenizing is done for each case

POS is parts of speech tagging and not tagging noun verb tagging\

stemming means
Past tense verbs are converted to
present tense: slept ! sleep

Report a screen shot of the diagram and the file import property window.

For **each of the four scenarios**, report the following.

a. Screen shot of the parse node properties.
b. The total number of terms extracted.
c. A table showing the top 20 terms along with the document counts for each term. The top 20 terms are the 20 terms with the highest frequencies (term counts).

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

Use pandas to read the 8 text documents, and NLTK to prepare the term/document matrices described in Part 1.

Report the following:

a. Your python code
b. Run the 4 scenarios described in Part 1 For each scenario report:
    1. The total number of terms extracted for that scenario.
    2. The top twenty terms sorted by the number of times each term occurs among the 20 documents.