DSBA/MBAD 6211: Homework Assignment 3 (50 pts)

Instructions: This is an individual assignment. The submitted solution and answers should be your own. The data file for this homework is **book_ratings.csv**, which is to be downloaded from Canvas. You are asked to use Python to perform Text Mining tasks and answer the given questions. Create a new Word document and save it as TextMining_xxxx (where xxxx is your ninernet login name). Write your full name on the first page of the Word document. Where required, write your answers or paste screenshots in this Word document. You need to submit both the Word document and Python Code file. **Your Python code should run correctly for your assignment to be graded. Code that generates error will result in loss of points (up to a maximum of 20%)**

Variables and models naming requirements:

- Include your ***name initials*** to the data frame names as well as model names in your Python coding. This is required for your work to be graded.
- For instance, my initials are **CS**, and in my coding, I would name the data frames as ***dfCS, dfCS.train***, and ***dfCS.test.*** I would also name the models as ***SVDCS, topicsCS***, etc.

Problem description and questions: The dataset boo_ratings.csv has reviews of books, along with other information, submitted on Amazon. The descriptions of the columns are given below.

- Id: Unique ID for the book
- Title: Book's title
- Price: The price of Book
- User_id: Unique ID of the user who rates the book
- profileName: Name of the user who rates the book
- review/helpfulness: Helpfulness rating (normalized to between 0 and 1)
- review/score: Rating from 0 to 5 for the book
- review/time: Time of given review
- review/summary: The summary of a text review
- review/text: The full text of a review

Tasks to complete and questions to answer (support your answers with data/analysis output):

[**Important Note:** You may find it useful to follow the code posted on Canvas for this assignment. However, the assignment context is different from the federalist papers example, and some steps have may need to be modified or dropped. Do not copy all the steps in the posted federalist papers example. In particular, don't blindly copy the comments from the posted code. Choose which steps and code are needed for the assignment. If code not necessary for solving the questions is present in your solution, it will be penalized.]

1. (10 pts) Preprocess the text data in each of the "review/summary" and "review/text" columns. Preprocessing should include tokenization, lowercasing, stop word removal, stemming and any other necessary steps. Describe each of the above step in the Word document.

2. (7 pts) Create word clouds for each of the "review/summary" and "review/text" columns, after preprocessing. Copy and paste the word clouds outputs in the Word document. Write one-two sentences for what you can interpret from each of the two word clouds.

3. (8 pts) Perform text mining for each of the "review/summary" and "review/text" columns to estimate the similarity between documents. Show the similarity output tables for the first 5 documents for the summary and text columns. Paste the appropriate screenshots in the Word document.

4. (10 pts) Perform topic modeling on "review/text" using LDA model and generate 6 topics. In the word document, show the topic model output for each of the 6 topics as the linear combination of the terms. For 2 of the topics, write a short description in 2-3 sentences for each topic.

5. (15 pts) Build predictive models for review/score, as follows:
    a. Apply SVD to extract 5 components from the "review/text" column.
    b. Combine two non-text columns with the 5 extracted components and build a decision tree model to predict the "review/score". Call it Model-1.
    c. Report the confusion matrix of Model 1.
    d. Repeat tasks 5a-c with SVD to extract 8 components and build a predictive model. Call it Model-2. Report the confusion matrix of Model-2.
    e. Which model performs better? Explain your answer.