Your Name:

Your Andrew ID:

Homework 3

Collaboration and Originality

1. Did you receive help of any kind from anyone in obtaining your data for this assignment (Yes or No)? It is not necessary to describe discussions with the instructor or TA.

   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.

2. Did you give help of any kind to anyone in obtaining their data for this assignment (Yes or No)?

   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

3. Are you the author of every word of your report (Yes or No)?

   If you answered No:
   a. identify the text that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

Your Name:

Your Andrew ID:

Homework 3

Answer all of the following questions. Each answer should be thorough, complete, and relevant. Points will be deducted for irrelevant details (e.g., brain dumps).

Typical answers to questions 5-7 will be a page or less in length. Think about how much you would write for a final exam question during an 80-minute exam.

Do not delete the questions from your report.

Data

Download a data file from http:/boston.lti.cs.cmu.edu/classes/95-865/HW/HW3/GenData.cgi. Questions 1, 2, 3, and 4 are done with data contained in this file.

Write the data file's unique identifier here: _____

Questions

1. Use Jensen-Shannon Divergence with Jelinek-Mercer smoothing and λ=0.4 to calculate the similarity between documents d_1 and d_2. Show your work. **[14 points]**

2. Use Jensen-Shannon Divergence with Jelinek-Mercer smoothing and λ=0.4 to calculate the similarity between document d_7 and the cluster of documents d_3-d_6. Show the similarity for the Single-Link, Complete-Link, and Centroid methods of determining similarity. Show your work. **[14 points]**

3. Show how the decision tree algorithm uses training documents d_8 – d_{12} and a binary representation to build a decision tree of depth one (i.e., one decision node that splits the documents into two sets). Use the information grain splitting criterion. Show your work (i.e., show the Information Gain calculation for each feature, and show which feature is selected). **[14 points]**

4. Use training documents d_8 – d_{12}, a binary representation, and Naïve Bayes with Laplace smoothing to classify document d_{13}. Show your work (i.e., show how Naïve Bayes makes its decision, and show which class is selected). **[14 points]**

5. Describe the Semantic Association from Association (SOA) approach to generating a corpus-specific subjectivity lexicon. Describe the initial information given to the algorithm, how it generates candidate unigram and phrase sentiment terms, and how it determines the polarity of each candidate term. **[14 points]**

6. Suppose that you work for Bloomberg. You need to develop an app that will identify news stories to insert into a user's personalized news feed. You have access to all of an individual's interactions on Bloomberg's site – what they read and what they discuss in chat sessions and discussion forums. Explain how this information is used to automatically develop a model of each person's interests, and what this model looks like. Then, explain how your system uses this model of an individual and today's news (approximately 1,000 stories) to select 5 stories for the user when she logs in. Explain and justify your choices. **[14 points]**

7. Suppose that you are doing brand awareness work for a Las Vegas casino. Your boss asks you to use social media to gather information about how the public perceives each Las Vegas casino. You decide to use comments posted to TripAdvisor about each casino (i.e., your answer can ignore how you got the comments), and to represent the public's perception using brief phrase lists (e.g., 10-30 phrases) that are, and are not, used in discussions about the casino. Describe your solution. Be clear how about how you generate and select phrases for each casino, and why you consider those phrases to be a good description. **[14 points]**