Student Name: ________________________________

Final Exam
Text Analytics (95-865)
December 15, 2011

Answer all of the following questions. Each answer should be thorough, complete, and relevant. Points will be deducted for irrelevant details. Use the back of the pages if you need more room for your answer.

In most cases, an answer should include an explanation. For example, "Stemming" is a weaker answer than 'Stemming, because …."

The exam should take you about 70-80 minutes to complete. The points are a clue about how much time I think each question should take to answer. Plan your time accordingly.

Good luck.

1. **Precision and Recall:** What are the Precision and Recall of class A for the following set of text categorization results? (You can use fractions, if you wish.) “Predicted” indicates that the classifier predicted that the item is a member of the category. “Actual” indicates what the classifier should have predicted. **[6 points]**

Predicted	**Actual**
A	A
B	A
A	A
A	B
B	A
B	B
A	A
A	A
B	A
A	A

2. Write the formula for Zipf’s Law. Why is it important? What practical use does it have? **[8 points]**

3. **Text representation:** In this class we have talked about several different types of text, for example, news articles, biomedical science papers, Epinions.com posts, and Twitter messages. The following questions examine the text processing operations required to make these different types of text useful for text analytics tasks.

   a. Give a general overview of the kinds of text processing operations typically used to make text more useful for text analytics tasks. Describes common text processing operations, and why each is used. **[10 points]**

   b. Consider the four types of text described above: News articles, biomedical science papers, Epinions.com posts, and Twitter messages. For each type of text, state how you would modify the general architecture that you described in question 1.a to accommodate its unique characteristics. It is fine for no change to be required for some type of text, but you must explain why. [**8 points]**

4. Suppose you have been hired by Microsoft to work on a new ‘brand awareness’ product. Your manager believes that accurate recognition of concepts and phrases in blogs and tweets is important to accurate sentiment detection and text classification. What technique would you recommend to provide very accurate (e.g., high Precision) recognition of product-related concepts and phrases? What strengths and weaknesses would your method have? **[10 points]**

5. Clustering algorithms use a similarity metric to organize a set of objects into groups (clusters). We discussed several similarity metrics in class (e.g., lnc.ltc, Jensen-Shannon). Pick one. Write its formula, and explain the different terms. **[10 points]**

6. Suppose you use software to find a list of people associated with Barack Obama in the press. However, Barack Obama is mentioned with many people, so the result is a long list of names that is difficult to understand. How could you cluster the names, so that similar types of people are grouped together? Be specific about the representation you would use for names (i.e., what type of vector would represent each person), and why it is an appropriate choice. **[10 points]**

7. Suppose that you work for T-Mobile, which runs customer discussion groups on its website. There are many active discussions happening simultaneously – too many for the company to monitor them all.

    a. How can the company get a general understanding of what is being discussed, and how it changes from week to week? Be specific about your choices of text representation, algorithm, etc. **[8 points]**

b. Each discussion page has slots for two ads. The company would like to select ads that are a good match to the page and also to the individual viewing the page. Assume that there are many ads and many users.

i. How does the company use a person's behavior (posts, comments, reading behavior) to create a model of an individual's interests? **[7 points]**

ii. How are the contents of the web page modeled? **[4 points]**

iii. How is an ad modeled? **[4 points]**

iv. How are the best two ads for this web page and this individual selected? **[7 points]**

c. After observing the effectiveness of your solution for awhile, the company realizes that advertising revenue is improved if ad selection is tuned differently for people based on their primary interest in using the website. There are five types of primary interest: “phone hardware”, “phone GUI”, “phone apps”, “coverage”, and “price”. For a particular user, how do you know which type of user he or she is? **[8 points]**