Student Name: _____

Andrew ID: _____

Final Exam
Text Analytics (95-865)
December 11, 2014

Answer all of the following questions. Each answer should be thorough, complete, and relevant. Think about your answer before you start writing. 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.

Calculators, phones, and other computational devices are not permitted. If you must do math, write fractions, and show your work so that it is clear that you know how to do the calculation.

Advice about exam answers....

- Answers sometimes treat machine learning like a magic wand that solves any problem, for example, "I will train a classifier to do <x>". That answer does not show that you understand the problem, thus it receives little credit. Show that you understand the problem by saying where the training data comes from, what features you would use, and what algorithm you would use.

- Answers sometimes say "I would use <technique> to do <x>". That answer shows that you remember a name, but it does not show that you remember how the technique works, or why it is the right tool for this problem. Give a brief description of how the technique works and why it is the right tool for this job. If the technique needs other information, explain where the information comes from.

1. Show how idf is calculated. Describe its purpose. Describe whether it is useful for clustering and categorization, and why. **[5 points]**

2. Given a choice between Kullback-Leibler divergence and Jensen-Shannon divergence as the similarity metric for document-to-document similarity comparisons (e.g., in clustering), which is the better choice, and why? **[5 points]**

3. Write the formula for cosine correlation. Define each of your terms. Identify and briefly describe the three main components of cosine correlation (treatment of frequency within a document and the corpus; treatment of document length). **[10 points]**

4. Most search engines provide access to document term list data structures. Draw a document term vector data structure. Explain briefly what it contains and how it is organized. Explain briefly a type of text analytics task that it supports. **[10 points]**

5. Describe how to generate a word (or phrase) cloud for an arbitrary set of documents. Be clear about what kinds of terms are candidates for the cloud, how terms are selected for inclusion in the cloud, and how their weights are determined. **[10 points]**

6. How many classifiers does k-fold classification train? Describe how much training data is used to train each classifier, and what each classifier is used for. **[10 points]**

7. Suppose that you work for T-Mobile, which operates a large discussion forum where its customers have discussions about T-Mobile products and services. The following two questions relate to this environment. You may find it helpful to read both questions before answering part a.

    a. Your boss wants to know what issues people discuss, and how those issues have changed during the last 12 months. She wants the results organized into 10 high-level categories that she has given you (phones, service plans, coverage, …). Within each category, she wants to see a detailed list of issues that were discussed, and their frequency during each quarter (3-month period). **[12 points]**

b. Your boss noticed a list of phones being discussed in the ‘phones’ categories. T-Mobile has good sales data about each phone, so she knows which phones are generally popular or unpopular. However, she is wondering if the sentiment about a phone (e.g., the Galaxy S6) expressed on the site in one month is a predictor of its sales volume in the next month. Describe the system that you would you design to answer this question. Be clear about how it will determine sentiment about specific products and be robust to unexpected events (e.g., new iPhones bending). **[12 points]**

8. Suppose that you work for a startup company that must classify a high-volume stream of texts into 100 categories. The classification must be very accurate. You have training data, a categorization algorithm (perhaps SVM), and an initial set of classifiers. Now you need to know the accuracy of the classifiers. You have the estimate provided by k-fold cross validation, but this is a high-stakes product launch – you need additional confirmation. Ideally you would examine the labels that are assigned to each document, but … you have 100 classifiers, and many documents. Random sampling would be expensive (e.g., 100 documents per category × 100 categories). How can you organize the large number of labeled documents so that you can be more strategic about your manual review**?** **[13 points]**

9. Suppose that you edit a scientific journal that uses peer-reviewing. Submissions to the journal are accepted only if a panel of three reviewers believes that the paper is acceptable. Reviewers are drawn from a population of people who have previously published scientific papers at top conferences or in top journals. You have access to their papers (about 10,000 papers over the last ten years), but it is a large community (about 1,000 authors), so you don't know the reviewers personally. Your task is to develop a system that will help you select good reviewers. Given a new paper, it must produce a list of 10 potential reviewers, ranked by their knowledge of the topics covered by the paper. Describe the system you would build, and how it would work. Be very clear about what information it uses, and how it ranks reviewers. **[13 points]**