

**Student Name:** \_\_\_\_\_

**Midterm Exam**  
Information Retrieval (11-741)  
March 9, 2006

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.

The exam should take you about 70 minutes to complete. The points are a clue about how much time we think each question should take to answer. We assume about 1.5 points per minute, so a 10-minute question is worth 15 points. Plan your time accordingly.

Good luck.



2. Suppose that you have a large collection of text documents that were each obtained by first scanning a magazine or journal page, and then performing optical character recognition (OCR), so that the resulting text has errors in about 10% of the words. Typical errors would be mis-recognized letters, spaces inserted incorrectly into the middle of a word, extra noise words and missing letters or punctuation.

Describe 3 specific effects this noise could have on the components of a retrieval system, such as document scoring functions, query processing, index representation, and so on. For each effect, explain how you might modify that part of the system to deal with the noise. **(15 points)**

3. Provide the formula for the Rocchio relevance feedback algorithm. Include a brief definition of each symbol. **(8 points)**

4. The simple probabilistic model of information retrieval is described by the following equation.

$$g(d) \approx \sum_{i=1}^n d_i \log \left( \frac{p_i(1-q_i)}{q_i(1-p_i)} \right)$$

What does  $p_i$  represent? Give an example of how  $p_i$  has been estimated in prior research. **(10 points)**

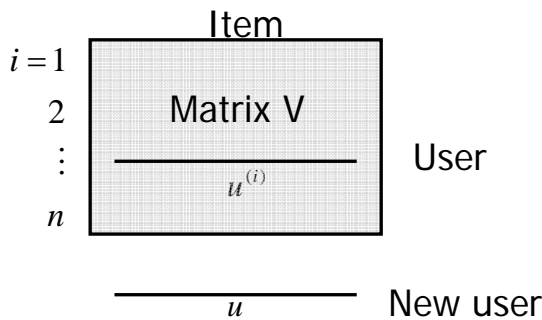
5. The statistical language modeling approach to information retrieval is generally associated with unstructured queries, but it is possible to incorporate probabilistic versions of many common query operators. A proximity operator such as “X NEAR/3 Y” matches “X Y”, “X a Y”, and “X a b Y”, but not “X a b c Y”. How would you implement this operator in a statistical language modeling framework? **(15 points)**

6. In federated search, what is the centralized sample database, how is it created, and why is it important? **(12 points)**

7. Collaborative Filtering (CF)

- a. Let  $x = (x_1, x_2, \dots, x_j)$ ,  $y = (y_1, y_2, \dots, y_j)$  be two user profiles. Provide the definitions of dot-product, cosine similarity and Pearson correlation coefficient. Discuss when the differences among these metrics would make the results of CF better or worse, comparing to each other. **(8 points)**

- b. The two forms “a” and “b” of CF shown below may be equivalent under certain conditions. Show some case(s) where they are equivalent and explain (prove) why. Show some case(s) where they are not equivalent and discuss why (either in words or in formula). Denote the user-item matrix as  $V_{I \times J}$ , a user profile (row) in  $V$  as  $u^{(i)}$ , a new user profile as  $u$ , an expanded user profile as  $u'$ , an empirical parameter as  $\beta$ , and an item-item matrix as  $A$ . **(16 points)**



$$a) \quad u' = u + \beta \frac{1}{I} \sum_{i=1}^I \text{sim}(u, u^{(i)}) \times u^{(i)}$$

$$b) \quad u' = u + \beta u A$$