INST 734 Final Exam
Examination period: December 14-19, 2015

You have 3 hours (180 minutes) to complete this exam, which begins when you first look at anything other than the first page of this file. You may use any materials, whether physical or online, that (you reasonably believe) existed before December 14. **You may not communicate with any other person between the time you first open this file (other than just to save it without looking past this first page) and the time you turn in your exam.** To be clear, this means you can't send or receive email, you can't answer the phone, and you can't talk to members of your family or anyone else, in person or in any other way.

You may write your answers on paper and scan them or you may create a file on your computer using any software whose results I can read (Word, PDF, …). If you elect to create a file, you can edit this document to add your answers, or you can create a new document and indicate your name, the start and end times, and which questions you are answering. **The exam must be submitted within 3 hours after you first open this file (other than just to save it without looking past this first page), and no later than 11:59 PM on Saturday, December 19, 2014.** The exam must be submitted using ELMS, and you must also email it to oard@umd.edu (as a backup). Both must be done immediately after completing your exam.

Additional time will not be allowed in the event of system failures, so save your file often.

Once you have opened this file (other than to save it without looking past this first page), it would be a violation of academic integrity to discuss any aspect of this exam with any person during the examination period or for two days thereafter (i.e., before Tuesday December 22).

Please answer only as many questions as the exam asks for. If you answer more questions, I will grade your answers in the order you write them, up to the number requested. For each question, I have indicated the expected length of an answer, but shorter or longer answers that are complete and correct will receive full credit. If you find part of a question to be confusing, explain your confusion and state any assumptions that you make that allow you to write an answer. When grading, I will consider any reasonable confusion (that I caused and that you clearly explain).

A total of 100 points are available on this exam. Each question is worth the indicated number of points. You do not need to answer the questions in order. You should plan your time to maximize the number of points that you receive!

After completing the examination, please hand type (or hand write) the following statement into the file you submit: "I pledge on my honor that I have not given or received any unauthorized assistance on this examination." The use of cut-and-paste to enter this statement is not allowed.

Course Goals (from the Web page

- Appreciate the capabilities and limitations of information retrieval systems.
- Understand the design and implementation of retrieval systems for text and other media.
- Evaluate the performance of an information retrieval system.
- Identify current research problems in information retrieval.

Everyone must answer all three parts of question 0!

0. (0 points)
    a. What is your name?
    b. What date and time did you start this exam (by opening the file for the first time)?
    c. What date and time did you end this exam (by submitting it using ELMS and by email)?

Everyone must answer all parts of question 1:

1. (20 points) Batch evaluations, such as those we did for Exercise E5, are necessarily somewhat artificial. Among the critiques that have been leveled against them are that real users can refine their query (whereas in a batch evaluation the query is typically fixed), the usefulness of a document to a user may depend on what documents they have seen previously (whereas in a batch evaluation the relevance judgments are fixed in advance), and users might not actually read down a ranked list in order (but virtually all of our batch evaluation measures assume that they do). Many more critiques could be added to that list. We even had an additional reading (summarized by one of your classmates) that suggested that improvements seen in batch evaluations may not translate to actual improvements in subsequent user studies. But every year hundreds of people participate in batch evaluations of information retrieval systems in evaluations such as TREC, CLEF, NTCIR and FIRE, and many more people publish batch evaluation results as conference or journal papers. What is it about batch evaluations that explains their popularity? Good answers to this question should focus on the actual value of batch evaluations (don't, for example, just say that people do them because that's what they learned in school – that may be a reason, but it is not a good reason!). There are several good answers that you might give – you should try to give as many as you can because no one factor could fully explain the popularity of this approach among information retrieval researchers. You should be able to answer this question in between half a page and a page.

Answer all parts of any <u>two</u> of the following four numbered questions:

2. (40 points) "Word embeddings" are a relatively new technique for mapping terms to rather short (e.g., 200-element) vectors in which words that have very similar usage (e.g., happy and glad) have nearly identical vectors, words with similar usage (e.g., doctor and nurse) have similar vectors (e.g., vectors for which the cosine similarity measure would be near the maximum possible value of 1.0), and words with completely different meaning (e.g., rock and chapter) have very different vectors (i.e., vectors for which the cosine similarity measure would be near the minimum possible value of 0.0). Describe the design of an information retrieval system in which word embeddings are used to improve recall (i.e., the

comprehensiveness of a search). There are many quite different designs that are possible, and you only need to describe one. But you need to describe it in your own words – don't just find a paper on this topic and copy the words from there (although of course you are welcome to search for papers on this, as you can for any of the questions). You should be able to answer this question in between half a page and a page.

3. (40 points) Answer all three parts of this question. You may turn in a spreadsheet with your calculations if you wish, but you should also explain your computations (at least briefly) for parts (b) and (c) using text. This will probably only take you a few lines of text in total (plus, of course, the tables and the final ranked list that show the results of your computations).

   a. (5 points) Create the term frequency matrix for the four document titles following each number and period below (the number is the document number):

      1. Solar power shines in Paris talks
      2. Solar plexus called source of inner truth
      3. The shine is off the rose; orchids power future sales growth
      4. Shiners convention in Toledo ends in disarray

   In building this matrix, use spaces and punctuation to "tokenize" text into terms in the usual way, stem terms by removing only the three most common English endings (s, ed, ing), convert everything to lower case, and do not remove stopwords. Your term-by-document matrix should have one row for each unique term and one column for each document (where in this case, the only thing in a document is its title). The entry in each cell should be the number of occurrences of the term that is associated with the element's row in the document that is associated with the element's column. You may leave elements for which the value is zero blank if you like, or you may place a zero there. Turn in this matrix.

   b. (25 points) Build a corresponding matrix of TF*IDF term weights for each term in each document. Show your work. Again, you can leave 0 values blank if you wish. Use the second of the two formulas for TF*IDF on module 3, lecture 2, slide 13 (which is the same one we used in exercise E3). This is the same formula that you used in Exercise E3. Turn in this matrix.

   c. (10 points) Use the matrix from part (b) to rank order the documents with respect to the following query:

      solar power from sunshine

   Show your work, and turn in the resulting ranked list (as a rank ordered list of document numbers, with the highest-scoring document listed first).

4. (40 points) In the guest cameo for Module 8, Jaap Kamps gave us a tour of cultural heritage institutions in Amsterdam. Much of the holdings of the Dutch National Archives are, as you might expect, in Dutch, and much of those holdings are handwritten manuscripts. Describe the design of an automated system for searching an enormous collection of scanned handwritten Dutch manuscripts, some of which have been transcribed (i.e., the contents have been manually retyped), many of which have been described (e.g., using Encoded Archival Description), and some of which (because of the size of the collection) have neither been

transcribed nor described. Your system should be capable of finding any scanned manuscript, regardless of whether it has been transcribed, described, both, or neither. Note that it need not be equally good at finding each manuscript, but you should design a system that would do about as well as possible at finding each of the four types of documents (transcribed, described, both or neither). Note also that you do not need to be able to find every document using a single query (although it is possible to design such a system!). For example, you could use one type of query to search only for described documents, another type of query to search only for transcribed documents, and a third type of query to search for documents that are neither described nor transcribed. You can assume that the users of your system will know how to read and write in Dutch. You should be able to answer this question in about a page.

5. (40 points) Describe the design of a user study in which we want to compare two systems for recommending products that people may be interested in buying, one of which uses a recommender system based on past purchase behavior. That recommender system first finds other people with similar shopping interests and then recommends products that those users have purchased (this is a so-called "user-item" recommender system). The other system works by simply suggesting the (same) most popular items to everyone. For example, if I buy a lot of pirate movies, the first system might recommend a pirate movie that many other pirate movie fans have bought (and that I haven't seen), but the second system might recommend the new Star Wars movie to me (and to everyone else who uses that second system) just because it is popular this week. The goal of your user study should be to determine which system provides the more useful recommendations when used by fairly new users who have (up until now) purchased only a few (3 to 5) products. Your description should include who the participants in your study will be, how many participants you will need in order to be able to draw reliable conclusions, what you will ask them to do, what data you will collect, and how you will analyze the results. You should be able to answer this question in about a page.

Don't forget to hand type (or hand write) the honor pledge <u>after you finish the exam</u>.

---------- End ----------