INST 734 Final Exam
Examination period: May 12-18, 2018

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 May 12. **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 or photograph 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 Friday May 18, 2018.** 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 Monday May 21).

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 30 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)?

Answer <u>any three</u> of the following five questions:

1. (10 points) In lifelogging, it is possible to continuously record "egocentric" video (i.e., video shot from the person's perspective), ambient audio (i.e., audio that the person hears, which might include speech and also nonspeech events such as traffic noise or the sound of doors closing), and location (e.g., from GPS or from the signal strengths of nearby WiFi access points). One problem with recording all of this is that some of it should not be shown to anyone (e.g., recordings made while in a public bathroom) and some of it should not be shown to certain people (e.g., business phone calls should be available only to the owner and perhaps to specific business associates). Manually marking which parts can be shown to whom is not practical, so a key prerequisite to the wide adoption of lifelogging is to develop automated techniques that can identify which segments might need to be protected. **Describe the design of such a system.** For full credit, your system for recognizing content that requires protection should be easily personalized to the needs of a specific lifelogger with the minimum practical effort and it should be reasonably reliable (but it doesn't need to be perfect). The output of the system should be the start time, stop time, protection status (for simplicity, just three categories: (1) protect from everyone, (2) protect from everyone except the lifelogger, or (3) allow access by everyone), and reason (which might be either of the reasons mentioned above, or any of a number of other reasons that you can think of; no reason is needed when allowing access by everyone). Note that the requirement that your system be easily adapted to the needs of a specific lifelogger does not mean that you can not use large amounts of training data; it only means that you can not use large amounts of manually labeled training data that are specific to that individual lifelogger).

2. (10 points) One surprising result from projects like Google Books is that even noisy OCR can be much more useful (on average, across many users and many queries) than high-quality metadata (e.g., from publishers or from library catalogs). But this is not an either-or proposition – we can use metadata together with the noisy OCR to do even better. **Start out by explaining how this can be done.** For full credit, your approach should be more sophisticated (and more effective!) than simply simultaneously searching both the OCR and the metadata using query terms entered by the user. **Then explain how the approach you have described can be used to improve the search for books for which you have**

metadata but lack OCR or for which you have OCR but lack metadata. In other words, explain how having both OCR and metadata for some books can improve search for other books.

3. (10 points) It is often possible to find or create text that is associated with an image. In newspapers, we have the news story that accompanies a photograph. On the Web, we have anchor text. On photo sharing sites such as flikr, people sometimes add tags to photographs. But we can get even more creative. We can use OCR to read text that we dine in the image (such as signs or the writing on t-shirts). We can pay crowdworkers to caption photographs. We can train machine learning to generate certain kinds of captions for certain kinds of photos. And we can use EXIF metadata to get the GPS location where the photo was shot, and then we can do a Web search to see what words people use to describe those locations. Because there are so many ways we could do this, we will often – but not always -- be able to find or create some text that is associated with each image. Rather than searching based on the image itself or searching based on the associated text, it should be possible to do even better by searching using both image matching and text matching. The query to such a system would be some text, together with one or more images, and the result would be a ranked list of images that (hopefully) the user wished to see. For example, a search for "Apollo 17" with a picture of a Lunar Module on the Moon should find pictures of the Apollo 17 Lunar Module on the Moon, but not the Apollo 16 Lunar Module, and not the Apollo 17 Command Module in orbit around the Moon. **Explain how you would measure how well that system is working.** For full credit, your answer should completely specify one or more methods that can be used to measure retrieval effectiveness (e.g., you should not just say "do it like it was done in XXXX paper" or "so a XXX study" – you should present a complete design for the experiments or other type of study/studies that would be done).

4. (10 points) Consider the following four documents:

    a. Female sun god in Japanese mythology
    b. LREC conference in Japan a success!
    c. Japanese scientists study effect of sun on arctic ice flows
    d. Sun and ice wreck havoc on Japanese traffic

    And the following query: Gods worshiped in Japan

    Use BM-25 (with the default parameter values set by Robertson and Sparck-Jones) to rank the four documents in decreasing order of BM-25 score for the query. You can find the BM-25 formula in Module 3, Lecture 2 on Slide 15 or in the Robertson and Sparck Jones paper on BM-25 at http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-356.pdf (note that the two sources use different versions of the formula; you can use either one). **Show your work, including at least the (terms-by-documents) term weight matrix**. You may find it convenient to submit your answer to this question as an excel spreadsheet.

5. (10 points) To date, there has been little use of behavioral evidence to improve Cross-Language Information retrieval (CLIR) search results. One reason for this is that assembling large quantities of behavioral evidence is difficult outside of commercial settings, and CLIR has not been a substantial focus of commercial activity. Imagine that you are preparing to pitch an idea for a new startup company to potential investors. **Explain, using terms that would be easily understood by a broad audience, you would use behavioral evidence to create high quality cross-language Web search engine.** To answer this question, you need to specify how the behavioral evidence needed to perform high quality CLIR on the Web would be obtained, and how that behavioral evidence would be used to improve Web CLIR search quality. Although a practical CLIR system would also require a machine translation component (to make it available to the broadest range of potential users), you don't need to focus on improving the machine translation component in your answer to this question – it will suffice to focus on CLIR.