University of Virginia
Department of Computer Science

CS 4780/6501: Information Retrieval
Spring 2020

Tuesday 3:30pm, March 31st to Thursday 3:30pm, April 2nd

Name: Yunlu Li
ComputingID: yl4df

- This is a **closed book** and **closed notes** exam. No electronic aids or cheat sheets or discussing the questions with anyone else are allowed.
- You are expected to finish this exam within 75 minutes.
- There are 7 pages, 4 parts of questions (the last part is for bonus questions), and 115 total points in this exam.
- Please carefully read the instructions and questions before you answer them.
- Please directly fill in your answers to the editable area below each question. And make sure to ***save*** your answers before you close the PDF file. (***NO*** argument about this after the grading is done).
- If you need any clarification of the exam questions, please directly send an email to the instructor within the exam period.
- Try to keep your answers as concise as possible; our grading is *NOT* by keyword matching.

Total	/100+15

Academic Integrity Agreement

I, the undersigned, have neither witnessed nor received any external help while taking this exam. I understand that doing so (and not reporting) is a violation of the University's academic integrity policies, and may result in academic sanctions.

Signature: Yunlu Li Digitally signed by Yunlu Li Date: 2020.03.31 19:16:34 -04'00'

Your exam will not be graded unless the above agreement is signed.

1 True/False Questions (12×3 pts)

Please choose either True or False for each of the following statements. For the statement you believe it is False, please give your brief explanation of it (you do not need to explain when you believe it is True). Three point for each question. *Note: the credit can only be granted if your explanation for the false statement is correct.*

1. Cosine similarity is proved to be equivalent to Euclidean distance in a vector space.
   True/False, and Explain: False

   Cosine similarity is preferred because document length is normalized.

2. The time complexity of searching for query q in an inverted index is $O(|q| \times |L|)$, where $|q|$ is the length of query q and $|L|$ is the average length of posting lists in the index.
   True/False, and Explain: True

3. Browsing mode is preferred for document retrieval when the user has a clear mind of what he/she is looking for.
   True/False, and Explain: False

   Browsing is preferred when users does not know which keywords to search for.

4. Vector Space Model is equivalent to Bag-of-Word model.
   True/False, and Explain: False

   BoW is a special case of Vector Space Model, which uses n-grams as the bases.

5. We can easily get the number of unique terms in a particular document from an inverted index.
   True/False, and Explain: False

   We should use BoW to easily get number of unique terms in a particular doc.

6. kappa statistic is lower bounded by -1.
   True/False, and Explain: True

7. Stemming helps improve recall of a Boolean retrieval model.
   True/False, and Explain: True

8. $P(D, Q|R) = P(Q|R)P(D|Q, R)$ is the key step used in deriving the query generation model, a.k.a, the language model, where D denotes a candidate document, Q denotes a user query, and R is the relevance label.
   True/False, and Explain: False

   The equation given above is Document Generation Model.

9. Relevance quality of a returned document is judged against all the query terms in the given query.
   True/False, and Explain: False

   Relevance quality is judged with respect to information need.

10. Mean Average Precision prefers a system to return as many relevant documents as possible.
True/False, and Explain: True

11. We seldom use a database system to solve web search problems mainly because of the efficiency concerns.
True/False, and Explain: False

The concern is that DB system cannot deal with unstructured data/text.

12. Maximum likelihood estimator is problematic when one does not have good knowledge about how to set up the prior.
True/False, and Explain: False

Bayesian estimation is problematic when one does not know prior.

2 Multiple Choice Questions (6×4 pts pts)

Please choose all the answers that you believe are correct under each question.

1. Good "*basic concepts*" in a vector space model should be: a, c
(a) orthogonal to each other;
(b) based on linguistic study;
(c) able to automatically compute the weights in each document;
(d) understandable by human.

2. Which of the following answer choices describes the correct order of the steps necessary to build an inverted index? Assume each document is represented by the tuple.
(a) sort by document ID, divide documents among different machines, sort by term ID, merge information about each term; a
(b) sort by term ID, divide documents among different machines, sort by document ID, merge information about each term;
(c) divide documents among different machines, sort by document ID, sort by term ID, merge information about each term;
(d) sort by term ID, merge information about each term, divide documents among different machines, sort by document ID.

3. Zipf's law tells us: b
(a) head words take major portion in English vocabulary;
(b) in a given French corpus, if the most frequent word's frequency is 1, then the second frequent word's frequency is around 0.5;
(c) comparing to tail words, removing head words helps more to reduce the storage of documents represented by a vector space model when using a dense matrix data structure;
(d) sublinear TF scaling is necessary.

4. What are the key assumptions behind the derivation of the RSJ ranking model:
   (a) binary relevance;
   (b) uniform document prior;
   (c) available relevance annotations;
   (d) attributes not in the query are equally likely to occur in relevant and non-relevant documents.

   a, b, d

5. Classical retrieval evaluation typically makes the following assumptions:
   (a) independent result relevance evaluation;
   (b) top down sequential browsing;
   (c) query is a proxy of a user's information need;
   (d) precision is preferred over recall.

   a, b, c

6. In which of the following situations a ranking system's MRR performance would be equal to its MAP performance:
   (a) the first returned document is relevant;
   (b) all relevant documents are returned;
   (c) there is only one relevant document;
   (d) there is no relevant document.

   c, d

3 Short Answer Questions (40 pts)

Most of the following questions can be answered by one or two sentences. Please make your answer concise and to the point.

1. Name five essential components of a search engine. (5pts)

   - Web Crawler
   - Document Analyzer & Document Indexer
   - Ranking Model
   - Query Parser
   - Retrieval Evaluation

2. Describe the three stages of sorting-based inverted index construction (aka map-reduce). (3pts)

   - Sort by document ID
   - Local sort by term ID
   - Global merge sort

3. Why are N-grams not an ideal way to account for phrases in a query? (3pts)

   N-grams is constructed before query time. If you want to cover all queries, you have to cover all n-grams, which is impossible.

4. In class, we analyzed how Yahoo utilized a browsing method for its home page, while Google currently utilizes a querying method for its home page/search engine design. Provide two key differences between these two methods. (4pts)

   - With or without use of keywords to express information need.
   - With or without organized paths and structures of information

5. How does Zipf's law ensure effective inverted index compression? (4pts)

   We represent the posting list by gaps between DocIDs. Based on Zipf's law, we know that for less frequent words, we have shorter list but large gaps and for more frequent words, we have lots of smaller gaps. The heavily biased distribution makes compression effective.

6. What are the three key elements of classical IR evaluation? (6pts)

   - A document collection
   - A set of information needs expressed as queries
   - A set of relevance judgement on query document pairs

7. What is NDCG? How is it better than P@k, MAP and MRR? (3pts)

   NDCG is Normalized Discounted Culmulative Gain. It discounts the gain of relevant documents as the rank goes lower, sums them up, and gets normalized against ideal rank. It is better than the other three because it provides a way to evaluate multiple grades of relevance.

8. Suppose that we know there are 4 relevant files C, D, E, F. The result is returned in the following position: K, D, A, F, G, C. When calculating the mean average precision, after adding up the precision at each relevant position, what should we divide the sum by? (2pts)

   4

9. Describe a circumstance in which one would want to use P@K over another ranking approximation like Mean Reciprocal Rank? And how about vice-versa (i.e. use MRR over P@K)? (4pts)

   - You care the number of relevant documents in top K retrieved documents.
   - You care the returned position of first relevant document.

10. What are the three key heuristics commonly shared by vector space models and BM25? (6pts)

    - Term Frequency
    - Inverse Document Frequency
    - Document Length Normalization

4 Bonus Questions (15 pts)

The first question is supposed to be open ended. Your answers have to be very specific to convince the instructor that you deserve the bonus (generally mention some broad concepts will not count).

1. Microsoft Bing would like to become a *real-time personalized* web search engine, i.e., its search results can be optimized with respect every action the user has taken in the system. Where do you think in the classical search engine architecture that Bing needs to renovate, and how exactly? (12pts)

   The first component that can be renovated is the user's query log. This component is essential to make search engine personalized. I think some machine learning algorithms, like Rancom Forest or ANN, can be incorporated with query log to predict user's interests and preference.

   The secod component is Crawler. Based on user's interests and preference, we can make focused crawling more personalized by prioritizing crawling on the fields that the user is interested.

   The third one is the Ranking Model. There are several ranking factors that should be taken into consideration, like user's preference and search location derieved from query logs. The geographical and topical proximity should be measured to increase the degree of personalization.

   Another one that could be renovated is Collaborative Ranking. Clustering algorithms, like k-means, could be utilized to group similiar users together based on query logs. Search results can be ranked on such a neighbor-based method.

2. List your favorite (labeled as +) and disliked (labeled as -) aspects of this class. (3pts)

   +: Well designed Machine Problems. Really helps to understand core concepts.
   -: Receive little instruction on projects.