CS371R: Midterm Exam

October 10, 2023

NAME: _____

UT EID: _____

INSTRUCTIONS:

- You have 1 hour and 15 minutes to complete the exam.
- The exam is closed book, closed notes, and closed computer, except for a scientific calculator and the provided equation sheets.
- Mark your answers **on the exam itself**. We will not grade answers on scratch paper or the back pages of the exam that are unnumbered.
- Make sure that your answers are legible and your handwriting is dark. We will be scanning the exams and grading them using Gradescope.
- Be sure to show your work on all problems in order to allow for partial credit.

1. (15 points) Corpus C consists of the following three documents:

   "new york times"
   "new york post"
   "los angeles times"

   Assuming that term frequencies are normalized by the maximum frequency in a given document, calculate the TF-IDF weighted term vectors for all documents in C by completing the tables below.

   Fill in the term frequencies in the table below:

	angeles	los	new	post	times	york
"new york times"						
"new york post"						
"los angeles times"						

   Fill in the inverse document frequencies in the table below:

angeles	los	new	post	times	york

   Fill in the TF-IDF weighted term vectors in the table below:

	angeles	los	new	post	times	york
"new york times"						
"new york post"						
"los angeles times"						

2. (16 points) Given the following document vectors:

	chai	latte	muffin	pumpkin	spice
"pumpkin spice latte"	0	1	0	1	1
"chai latte muffin"	1	1	1	0	0

and the following query:

"chai pumpkin spice pumpkin muffin",

calculate the TF weighted query vector (no IDF factor) by filling out the table below. Assume that term frequencies are normalized by the maximum frequency in a given query.

chai	latte	muffin	pumpkin	spice

Compute the score of both of the documents using the cosine similarity measure.

3. (16 points) Assume that an IR system returns a ranked list of 10 total documents for a given query. Assume that according to a gold-standard labelling there are 4 relevant documents for this query, and that the relevant documents in the ranked list are in the 1st, 3rd, 5th, and 7th positions in the ranked results. Calculate and clearly show the interpolated precision value for each of the following standard recall levels: {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0} for this individual query.

   Fill in the precision-recall values corresponding to relevant documents positions in the table below:

Document Number	Recall	Precision
1		
3		
5		
7		

Fill in the interpolated precision-recall values in the table below:

Recall	Precision
0.0	
0.1	
0.2	
0.3	
0.4	
0.5	

Recall	Precision
0.6	
0.7	
0.8	
0.9	
1.0	

4. (16 points) The table below shows the final ranked list of results for an IR search together with their continuous human-rated relevance values. Assume the table contains all documents with non-zero relevance. Compute the values of the DCG and NDCG evaluation metrics for each value of n and add them to the table. Complete the second table to show the idealized DCG (IDCG) values.

n	doc	relevance (gain)	——DCG——	—-NDCG—-
1	D23	1.0		
2	D78	0.4		
3	D90	0.8		
4	D17	0.5		

n	–doc–	relevance (gain)	——IDCG——
1			
2			
3			
4			

5. (16 points) Consider the following web pages and the set of web pages that they link to:

   Page A points to pages B, C, and E.
   Page D points to pages B, and C.

   All other pages have no outgoing links.

   Consider running the HITS (Hubs and Authorities) algorithm on this subgraph of pages. Simulate the algorithm for three iterations. Show the authority and hub scores before and after normalization for each page for each iteration. Order the elements in the vectors in the sequence: A, B, C, D, E.

   (a) Show work for iteration 1 below:

(b) Show work for iteration 2 below:

(c) Show work for iteration 3 below:

6. (21 points) Provide short answers (1-3 sentences) for each of the following questions:

- What are two problems with the Boolean model of document retrieval?

- Why is vector inner product (dot product) alone not a good similarity metric for vector space retrieval, i.e. why is normalizing by the length of the vectors important?

- When evaluating relevance feedback, or any machine learning method that uses training data, what must be true about the data used to test the system?

- What aspect of the relevancy of search results does the NDCG metric account for that is not accounted for by other metrics?

- How should one order the items in the queue in a web spider in order to implement a "directed/focused" crawler?

- Why is Zipf's law important to the significant performance advantage achieved by using an inverted index?

- What is the **disadvantage** of using breadth-first search for spidering compared to using depth-first search?

- (Extra credit) Who is the famous mathematician who co-authored papers with many other mathematicians and is used to assign a number to scientists based on their distance from him in the small-world co-authorship graph.

- (Extra credit) What early web company settled a lawsuit with Google over violating the patent for the PPC (Pay Per Click) business model for web search after buying the original company that invented it.