CS371R: Midterm Exam

October 17, 2019

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. (13 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. (14 points) Given the following document vectors:

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

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

3. (13 points) A user makes the query "cheap austin flights" and gets the ranked results in the table below. The document vector for each document is next to the corresponding document. The stop word "in" has been removed.

	austin	cheap	flights	kayak	rental
1. "kayak cheap flights"	0	1	1	1	0
2. "cheap kayak rental"	0	1	0	1	1
3. "kayak in austin"	1	0	0	1	0

The query vector for "cheap austin flights" is:

austin	cheap	flights	kayak	rental
1	1	1	0	0

Assume in response to the results of the query "cheap austin flights " that the user rates the following documents as *irrelevant*:

"kayak cheap flights"
"cheap kayak rental"

Recalculate the query vector for "cheap austin flights" to account for relevance feedback using the Ide "Dec Hi" method, assuming $\alpha = \beta = \gamma = 1$.

Fill in the new query vector in the table below:

austin	cheap	flights	kayak	rental

4. (14 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	

5. (13 points) Given a corpus that consists of the following two documents:

   "new orleans"
   "new hampshire"

   Compute a normalized association matrix that quantifies term correlations in terms of how frequently they co-occur. Order terms in the matrix alphabetically.

6. (12 points) What is the Levenshtein distance between the following pairs of strings? List the edit operations you used to transform the first string into the second string to find the Levenshtein distance.

   "beauracracy" and "bureaucracy"

   "Levenshtein" and "Levanstine"

7. (21 points) Provide short answers (1-3 sentences) for each of the following questions:

- 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?

- What are the two major limitations of the Porter Stemmer?

- 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?

- List three other items that exhibit a Zipfian distribution other than the frequency of words in text and number of in-links on the web?

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

- How does one construct a random directed graph that, like the web, is "scale free" where the number of edges into a node exhibits a Zipfian power-law distribution?

- In Matt Lease's automatic fact checking system, given a claim, relevant article headlines, and the trustworthiness of the source of each article, what are the two values that the system predicts?

- (Extra credit) From what conference was Tim Berner Lee's first paper about the world-wide-web rejected?

- (Extra credit) Herbert Simon, one of the founders of AI and investigators of the cause of Zipfian distributions, is the only recipient of both of what two major scientific awards (be specific)?