CS 371R Information Retrieval and Web Search: Midterm Exam

Oct. 18, 2018

NAME: _____

Be sure to show your work on all problems in order to allow for partial credit.

1. (14 points) Assume that simple term frequency weights are used (no IDF factor), and the only stopwords are: "is", "am" and "are". Compute the cosine similarity of the following two simple documents:
   (a) "precision is very very high"
   (b) "high precision is very very very important"

2. (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 5 relevant documents for this query, and that the only relevant documents in the ranked list are in the 2nd, 3rd, 4th, and 8th 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.

3. (14 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	0.6		
2	D78	1.0		
3	D90	0.0		
4	D17	0.5		
5	D78	0.9		

n	–doc–	relevance (gain)	——IDCG——
1			
2			
3			
4			
5			

4. (13 points) Write a Perl regular expression (regex) for matching the final line in a US Postal address in Texas or California. Assume that it consists of a city name of one or two alphanumeric words followed by a comma and then any amount of optional whitespace, followed by one of the two-letter state abbreviations (TX or CA) followed by some whitespace and then a 5 digit zip-code with an optional "plus four" digits introduced by a hyphen.

5. (12 points) Assuming Zipf's law with a corpus independent constant $A = 0.1$, what is the fewest number of most common words that together account for more than 18% of word occurrences (i.e. the minimum value of m such that at least 18% of word occurrences are one of the m most common words).

6. (12 points) Consider the following web pages and the set of web pages that they link to:

   Page A points to pages B and D.
   Page B points to pages C, F, and G.
   Page C points to page D.
   Page D points to page H.
   Page G points to pages E and H.
   Page H points to page C.

   Show the order in which the pages are indexed when starting at page A and using a breadth-first spider with duplicate page detection. Assume links on a page are examined in the orders given above.

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

What are two aspects of the web that make make web search fundamentally different from earlier, traditional IR?

In the vector-space model, why is it **not** necessary to normalize term frequencies when the resulting document vectors are only used for computing cosine similarity?

But why **is** it necessary to normalize term frequencies when the resulting document vectors are used for standard vector-space relevance-feedback methods?

What is the functional role (i.e. purpose) of the IDF factor in standard term weighting?

How does stemming typically affect recall? Why?

Why does thesaurus-based query expansion typically not work very well?

On what type of plot does a power law result in a straight line? What is the slope of the line (in terms of the parameters of the power law, $y = kx^c$)?

(Extra credit) Before moving to the US, both the founding father of IR, Gerald Salton, and the inventor of the hash table, Hans P. Luhn, were born in what country?

(Extra credit) Google settled a lawsuit with what early web company after that company acquired the company that had patented the pay-per-click model?