CS371R: Midterm Exam

October 19, 2017

NAME: _____

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

1. (a) (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. Order words in the vectors alphabetically.

(b) (14 points) Given the following query:

"new new times"

calculate the TF-IDF weighted query vector, and compute the score of each document in C using the cosine similarity measure. Assume that term frequencies are normalized by the maximum frequency in a given query.

(c) (13 points) Assume in response to the results of the query "new new times," that the user rates the following documents as *irrelevant*:

"new york times"
"new york post"

Reformulate the query to account for relevance feedback using the Ide "Dec Hi" method, assuming $\alpha = \beta = \gamma = 1$.

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 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.

3. (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.

4. (12 points) What is the Levenstein distance between the following pairs of strings? Justify your answer.

   "thorough" and "throughout"

   "filosofy" and "philosophy"

5. (21 points) Provide short answers (1-3 sentences) for each of the following questions:

   - List **two dis**advantages of the Boolean retrieval model.

   - Briefly describe the steps involved in exact phrasal retrieval using a standard word-based inverted index. What piece of information is missing from the `ir.vsr.InvertedIndex` class, which is required for phrasal retrieval?

   - Compare query expansion using global analysis and local analysis.

   - What important aspect of relevance does the NDCG metric take into account that precision, recall, and F-measure do not?

- What is the fundamental difference in the purposes of HTML and XML?

- What is the main **dis**advantage of breadth-first search compared to depth-first search?

- What is the current best explanation for why word frequencies, wealth, city population, and many other properties of human society follow Zipf's law?

- (Extra credit) Amit Singhal, the former Head of Google Search, got his PhD from what advisor at what university?

- (Extra credit) Why did Singhal leave Google in 2016 to become senior vice president of engineering at Uber? He was actually asked to leave Uber earlier this year for not revealing this reason why he originally left Google.