# CS 371R Information Retrieval and Web Search: Midterm Exam

Oct. 10, 2024

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) 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. Fill in the precision-recall values corresponding to relevant documents positions in the table below and in the following tables calculate and 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.

| Document Number | Recall | Precision |
|---|---|---|
| 2 | | |
| 3 | | |
| 4 | | |
| 8 | | |

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 | |

3. (11 points) Assume in response to the results of the query "microwave dish," that using relevance feedback, the user rates the following document as *relevant*:

   - "microwave TV dish"
   - "microwave TV antenna"

   and the following document as *irrelevant*:

   - "microwave safe dish"

   Assuming simple term-frequency weights (no normalization, no IDF), show the revised query vector computed using the "Ide regular" method. Fill out the table below for the revised query assuming $\alpha = \beta = \gamma = 1$.

| antenna | dish | microwave | safe | TV |
|---|---|---|---|---|
| | | | | |

4. (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).

5. (11 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.

6. (18 points) Consider the following pages and the set of web pages that they link to:

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

   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 for each page twice for each iteration, both before and after normalization, 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:

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

What are two aspects of the web that 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) Who is the Nobel and Turing award winning founding father of AI who first explored the "rich get richer" explanation for Zipf's law?

(Extra credit) What company successfully sued another company for violating the restrictions in their robots.txt file?