# CS 494: INFORMATION RETRIEVAL EXAM 1

FALL 2021

**Name:**

This test consists of 5 questions. The number of points for each question is shown below.

- Read all questions carefully before starting to answer them.
- Write all your answers clearly using no more than one page per exercise (ideally type your answers). Solutions exceeding one page per exercise will only be graded for the first page (with the rest being ignored).
- Show your work. Correct answers without justification will not receive full credit. However, also be concise. Excessively verbose answers will be penalized.
- Clearly state any simplifying assumptions you may make when answering a question.
- **Be sure to write your name on the test paper.**
- When you finish your exam, please submit in blackboard. If you type it, just submit the PDF. If you write it on a paper, please scan it (possibly with your phone) and then submit it as a PDF.
- The exam is available from 9am to 11:59pm Monday 10/11/2021.

| Question | 1 | 2 | 3 | 4 | 5 | total |
|---|---|---|---|---|---|---|
| Points | 20 | 20 | 20 | 20 | 20 | 100 |
| Your Points | | | | | | |

**Exercise 1 - 20 points. (Boolean Retrieval)**
Assume the following collection of short documents:

```
Doc 1:  John gives a book to Mary.
Doc 2:  John who reads a book loves Mary.
Doc 3:  Who does John think Mary loves?
Doc 4:  John thinks a book is a good gift.
```

i. Construct a term-document matrix that can be used to perform Boolean retrieval. The index terms have already been listed for you in the following table (note that terms have been stemmed and stopwords have been removed):

| Term | Doc1 | Doc2 | Doc3 | Doc4 |
|---|---|---|---|---|
| john | | | | |
| mary | | | | |
| give | | | | |
| book | | | | |
| read | | | | |
| love | | | | |
| think | | | | |
| gift | | | | |
| good | | | | |

ii. What documents would be returned in response to the following queries?

(john AND ¬ mary) OR think

john AND mary AND ¬ think

book AND (read OR john)

**Exercise 2 - 20 points. (Inverted Index and Cosine Similarity)**
Assume the following collection of short documents:

```
Doc 1:  John gives a book to Mary.
Doc 2:  John who reads a book loves Mary.
Doc 3:  Who does John think Mary loves?
Doc 4:  John thinks a book is a good gift.
```

After performing stemming and removing stop words, the vocabulary is:

$$\{john, mary, give, book, read, love, think, gift, good\}$$

i. How does stemming affect precision in a general Vector Space Retrieval model? Explain.

ii. Construct an inverted index. Show the index graphically with linked lists.

iii. Consider the query "love Mary" and simulate the retrieval of documents in response to this query. Show how the inverted index is used to identify relevant documents and how the cosine similarity between the query and the relevant documents is calculated incrementally using a hashtable. Rank the documents based on their cosine similarity.

**Exercise 3 - 20 points. (Query likelihood language model)**
Suppose we have a collection that consists of three documents given below.

```
Doc 1:  Language and Vision, Language, Language Human Robot Interaction
Doc 2:  Speech and Language
Doc 3:  Natural Language Speech and Vision Speech
```

Assume that we also have the following query: *Vision and Language*.

Build the following language model for this collection. Compute the model probabilities for the query, and show the final ranking of the documents. Ignore stop words and punctuation.

Estimate unigram models of documents using a mixture model between the documents and the collection with $\lambda = 0.2$. Remember the probability of $Q$ given $d$ is:

$$P(Q|d) = \prod_{w \in Q} [(1 - \lambda)P(w|M_c) + \lambda P(w|M_d)]$$

**Exercise 4 - 20 points. (Query Expansion)**
Assume again the following collection of short documents (as in Exercise 2):

```
Doc 1:  John gives a book to Mary.
Doc 2:  John who reads a book loves Mary.
Doc 3:  Who does John think Mary loves?
Doc 4:  John thinks a book is a good gift.
```

After performing stemming and removing stop words, the vocabulary is:

$$\{john, mary, give, book, read, love, think, gift, good\}$$

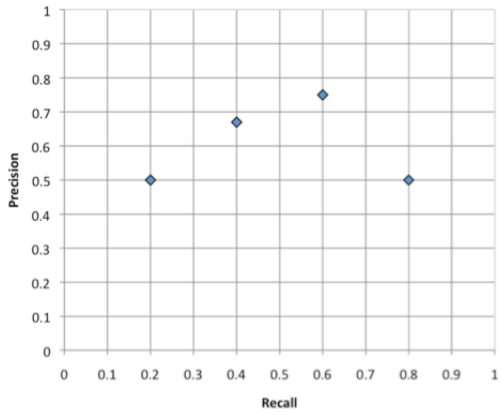
Consider running the standard Rocchio algorithm on the above collection for the query "love Mary" to perform query expansion and calculate the updated query $\vec{q_m}$. Remember that:

$$\vec{q_m} = \alpha\vec{q_0} + \beta\frac{1}{|D_r|}\sum_{\vec{d_j}\in D_r}\vec{d_j} - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d_j}\in D_{nr}}\vec{d_j}$$

For calculation, use $\alpha = 1$, $\beta = 1$, and $\gamma = 1$. Also, assume $D_r = \{Doc2\}$ and $D_{nr} = \{Doc1, Doc3, Doc4\}$.

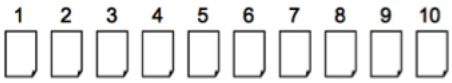
**Exercise 5 - 20 points. (IR Evaluation)**

The following is a plot of uninterpolated precision-recall values for an IR system that retrieves 10 ranked documents when queried on a particular topic. You know that there are 5 relevant documents for this topic (but only 4 of them are retrieved by the IR system).




i. Draw the interpolated precision-recall curve in the plot above.

ii. In the diagram below, each box represents a retrieved document. Based on the above precision-recall plot, which retrieved documents are relevant? Write an "R" on the relevant document. Leave the non-relevant documents empty. Show your work.




iii. What is the R-precision value for this set of retrieved ranked documents?