**LBSC 796/INFM 718R: Information Retrieval Systems**
**Spring, 2005**

## Midterm Exam

Name: Jimmy Lin

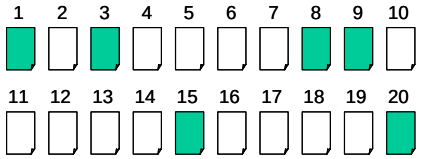
- Please show sufficient work to demonstrate your understanding of the material. This is mostly for your benefit, because it will allow partial credit to be awarded.
- This exam has seven questions, six of which are divided into multiple parts.
- You will have until 8:45pm to complete the exam.
- Good luck!

| Question | Points | Total |
|---|---|---|
| 1 | 16 | 16 |
| 2 | 10 | 10 |
| 3 | 48 | 48 |
| 4 | 20 | 20 |
| 5 | 17 | 17 |
| 6 | 12 | 12 |
| 7 | 2 | 2 |
| **Total** | 125 | 125 |



## Question 1. Evaluation (16 points)

An information retrieval system returns the following ranked list for a particular query:




Colored blocks represent relevant documents; white blocks represent irrelevant documents. From the known relevance judgments, you know that there are eight relevant documents in total.

**A.** (4 points) What is the Mean Average Precision (MAP)?

$$\frac{1+\frac{2}{3}+\frac{3}{8}+\frac{4}{9}+\frac{5}{15}+\frac{6}{20}+0+0}{8} \approx .39$$

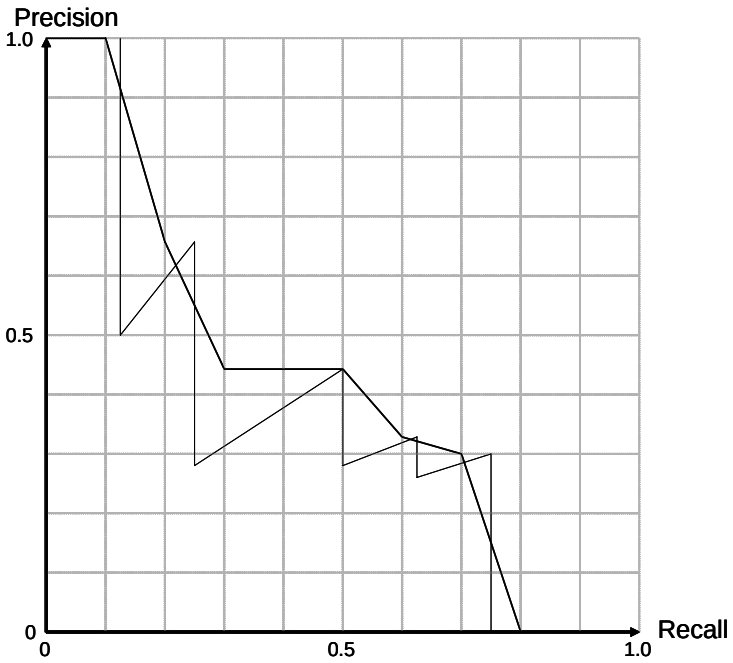
**B.** (2 points) What is the R-precision?

$3/8 = 0.375$

**C.** (2 points) What is Precision at 10?

$4/10 = 0.4$

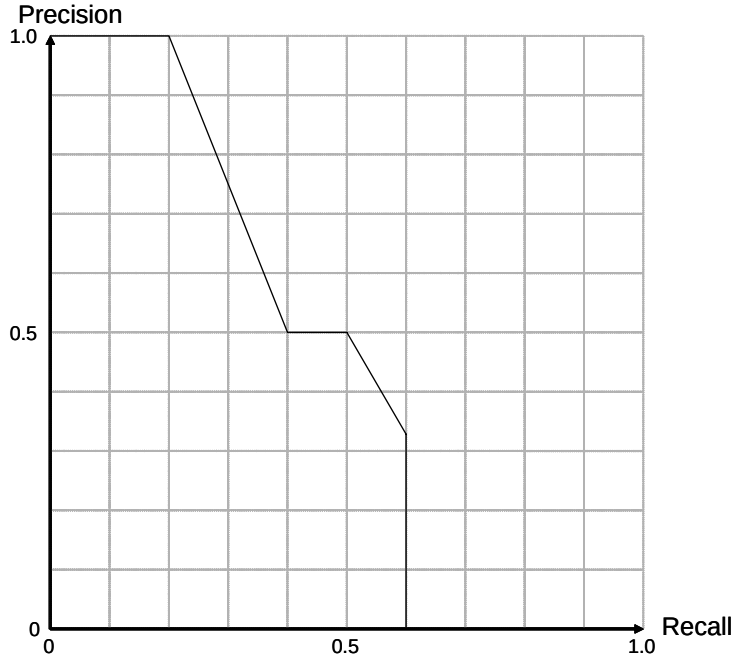
**D.** (8 points) Plot the ROC curve (precision-recall curve), **both** uninterpolated and interpolated versions:






## Question 2. More Evaluation (10 points)

Assume a document retrieval system produced the following interpolated ROC curve (precision-recall curve) on a particular query (based on 20 hits):




You know that there are ten relevant documents.

**A.** (2 points) What is the precision after the system has retrieved three relevant documents?

Having retrieved 3 documents = .3 recall. Precision is .75 at that point.

**B.** (2 points) Going down the hit list, I discover that I've retrieved *n* documents, and all of them are relevant. What's the maximum possible value of *n*?

2. Beyond .2 recall, precision drops; .2 recall translates to 2 documents.

C. (6 points) Where are the relevant documents in the hit list? Mark a relevant document with an **R** in the corresponding box. Leave irrelevant documents unmarked.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| R | R | | R | | | | R | | R |

| 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | R | | |



## Question 3. Boolean and Vector Space Retrieval (48 points)

Assume the following fragments comprise your document collection:

Doc 1: Interest in real estate speculation
Doc 2: Interest rates and rising home costs
Doc 3: Kids do not have an interest in banking
Doc 4: Lower interest rates, hotter real estate market
Doc 5: Feds' interest in raising interest rates rising

Assume the following are stopwords: an, and, do, in, not

**A.** (10 points) Construct the term-document matrix for the above documents that can be used in Boolean retrieval. The index terms have already been arranged for you alphabetically in the following table:

| Term | Doc 1 | Doc 2 | Doc 3 | Doc 4 | Doc 5 |
|---|---|---|---|---|---|
| banking | 0 | 0 | 1 | 0 | 0 |
| costs | 0 | 1 | 0 | 0 | 0 |
| estate | 1 | 0 | 0 | 1 | 0 |
| feds | 0 | 0 | 0 | 0 | 1 |
| have | 0 | 0 | 1 | 0 | 0 |
| home | 0 | 1 | 0 | 0 | 0 |
| hotter | 0 | 0 | 0 | 1 | 0 |
| interest | 1 | 1 | 1 | 1 | 1 |
| kids | 0 | 0 | 1 | 0 | 0 |
| lower | 0 | 0 | 0 | 1 | 0 |
| market | 0 | 0 | 0 | 1 | 0 |
| raising | 0 | 0 | 0 | 0 | 1 |
| rates | 0 | 1 | 0 | 1 | 1 |
| real | 1 | 0 | 0 | 1 | 0 |
| rising | 0 | 1 | 0 | 0 | 1 |
| speculation | 1 | 0 | 0 | 0 | 0 |

**B.** (2 points each) What documents would be returned in response to the following queries?

*interest NOT rates*

Docs 1 and 3

*( interest AND rates ) NOT ( rising OR kids )*

( interest AND rates ) → Docs 2, 4, 5
( rising OR kids ) → Docs 2, 3, 5
( interest AND rates ) NOT ( rising OR kids ) → Doc 4

*( ( real AND estate ) OR home ) AND ( interest AND rates )*

( ( real AND estate ) OR home ) → Docs 1, 2, 4
( interest AND rates ) → Docs 2, 4, 5
( ( real AND estate ) OR home ) AND ( interest AND rates ) → Docs 2, 4

*( kids AND home )*

None



Doc 1: Interest in real estate speculation
Doc 2: Interest rates and rising home costs
Doc 3: Kids do not have an interest in banking
Doc 4: Lower interest rates, hotter real estate market
Doc 5: Feds' interest in raising interest rates rising

stopwords: an, and, do, in, not

**C.** (20 points) Construct the vector space term-document matrix for the above documents (repeated from before) using *tf.idf* term weighting. Normalize your vectors. The following blank tables are provided for your convenience. You can use as many or as few of them as you wish. Clearly indicate your final answer.

| | | **TF** | | | | |
|---|---|---|---|---|---|---|
| **Term** | **IDF** | **Doc 1** | **Doc 2** | **Doc 3** | **Doc 4** | **Doc 5** |
| banking | .699 | | | 1 | | |
| costs | .699 | | 1 | | | |
| estate | .398 | 1 | | | 1 | |
| feds | .699 | | | | | 1 |
| have | .699 | | | 1 | | |
| home | .699 | | 1 | | | |
| hotter | .699 | | | | 1 | |
| interest | 0 | 1 | 1 | 1 | 1 | 2 |
| kids | .699 | | | 1 | | |
| lower | .699 | | | | 1 | |
| market | .699 | | | | 1 | |
| raising | .699 | | | | | 1 |
| rates | .222 | | 1 | | 1 | 1 |
| real | .398 | 1 | | | 1 | |
| rising | .398 | | 1 | | | 1 |
| speculation | .699 | 1 | | | | |

**TF.IDF**

| Term | Doc 1 | Doc 2 | Doc 3 | Doc 4 | Doc 5 |
|---|---|---|---|---|---|
| banking | | | .699 | | |
| costs | | .699 | | | |
| estate | .398 | | | .398 | |
| feds | | | | | .699 |
| have | | | .699 | | |
| home | | .699 | | | |
| hotter | | | | .699 | |
| interest | | | | | |
| kids | | | .699 | | |
| lower | | | | .699 | |
| market | | | | .699 | |
| raising | | | | | .699 |
| rates | | .222 | | .222 | .222 |
| real | .398 | | | .398 | |
| rising | | .398 | | | .398 |
| speculation | .699 | | | | |
| *length* | .897 | 1.09 | 1.21 | 1.35 | 1.09 |

**Normalized TF.IDF**

| Term | Doc 1 | Doc 2 | Doc 3 | Doc 4 | Doc 5 |
|---|---|---|---|---|---|
| banking | | | .578 | | |
| costs | | .641 | | | |
| estate | .444 | | | .295 | |
| feds | | | | | .641 |
| have | | | .578 | | |
| home | | .641 | | | |
| hotter | | | | .518 | |
| interest | | | | | |
| kids | | | .578 | | |
| lower | | | | .518 | |
| market | | | | .518 | |
| raising | | | | | .641 |
| rates | | .204 | | .164 | .204 |
| real | .444 | | | .295 | |
| rising | | .365 | | | .365 |
| speculation | .779 | | | | |



**D.** (4 points each) Simulate the retrieval of documents in response to the following queries. Indicate the order in which documents will be retrieved, and the similarity score between the query and each document.

*interest rising*

Doc 2: .365
Doc 5: .365
Doc 1: 0
Doc 3: 0
Doc 4: 0

*real estate interest*

Doc 1: .888
Doc 4: .59
Doc 2: 0
Doc 3: 0
Doc 5: 0

**E.** (2 points) Consider Doc 5: "Feds' interest in raising interest rates rising." Do the two instances of the term "interest" have the same meaning? What problem is this an example of?

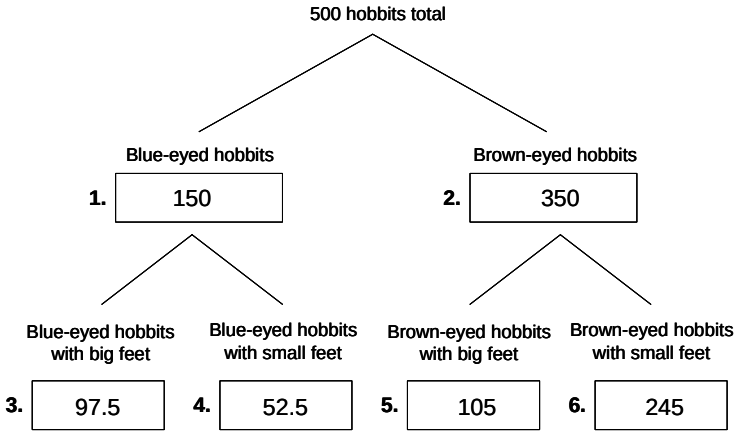
Polysemy.

## Question 4. On hobbits (20 points)

This question is about hobbits. Let's say that hobbits come in two different eye colors: blue and brown. Let's also say that some hobbits have small feet, and some hobbits have large feet (by hobbit standards, of course).

Now, suppose we know the following things:

- Brown-eyed hobbits are more common than blue-eyed hobbits. The probability that a random hobbit has brown eyes is 70%.
- Blue-eyed hobbits tend to have larger feet. 65% of blue-eyed hobbits have "large feet".
- Brown-eyed hobbits tend to have smaller feet. 70% of brown-eyed hobbits have "small feet".




**A.** (1 point) Are eye color and feet size independent for hobbits?
No.

(2 points each) For the following questions, write your answer in the box indicated. Out of 500 hobbits:

**B.** How many blue-eyed hobbits would you expect? (Box 1)
**C.** How many brown-eyed hobbits would you expect? (Box 2)
**D.** How many blue-eyed hobbits with big feet would you expect? (Box 3)
**E.** How many blue-eyed hobbits with small feet would you expect? (Box 4)
**F.** How many brown-eyed hobbits with big feet would you expect? (Box 5)
**G.** How many brown-eyed hobbits with small feet would you expect? (Box 6)



**H.** (3 points) If you saw a hobbit with large feet wearing sunglasses, what is the probability that this hobbit has **brown** eyes?

$$\frac{105}{97.5+105} \approx .519$$

**I.** (3 points) If you saw a hobbit with large feet wearing sunglasses, what is the probability that this hobbit has **blue** eyes?

$$\frac{97.5}{97.5+105} \approx .481$$

**J.** (1 points) If you saw a hobbit with large feet wearing sunglasses, what eye color would you guess this hobbit had?

Brown.

## Question 5. True or False (17 points)

For each of the following statements, write **T** in front of the question if it is true; write **F** in front of the question if it is false. Note, a true assertion with a false justification is considered false.

**The following statements are about evaluation:**

**False** **A.** In a general information seeking environment involving real users, precision and recall are inversely correlated.

**False** **B.** Tests of statistical significance tell us which of the parameters is causing the difference in performance

**False** **C.** Tests of statistical significance are *not* important in quantitative user studies

**False** **D.** In Web search, R-precision is important.

**The following statements are about *pure* Boolean retrieval systems:**

**False** **E.** The AND operator sometimes discovers non-existent relationships between terms in the same document and across different documents. (What do I mean by relationships between terms? For example, if you wanted to find fast cars, you might use the query "fast AND cars".)

**False** **F.** The OR operator assigns a higher score to documents that contain both terms.

**False** **G.** The NOT operator can be used to discover all unwanted terms that should be excluded from a search.

**True** **H.** Proximity operators are less efficient computationally because the index needs to store positional information.

**The following statements are about vector space retrieval systems that use the *tf.idf* weighing scheme with vector length normalization:**

**True** **I.** We can use any arbitrary weighting scheme in building the vector representation of the query.

**False** **J.** Suppose your collection contained the document "We hold these truths truths to be self-evident…" You later discovered that some idiot had written "truths" twice, which you corrected. To fix the index, you need to recalculate all the vectors of every document in the collection.



**True** **K.** Document vectors vary in length because different terms receive different weights in different documents.

**The following statements are about retrieval systems based on language models:**

**False** **L.** P(A|B) never equals P(B|A).

**True** **M.** Smoothing is necessary because otherwise the model would assign a zero probability to queries that contain terms not present in the original document (from which the model was built).

**The following statements pertain to issues regarding text representation:**

**True** **N.** Sense-based indexing (or conceptual indexing) *attempts* to improve both precision and recall. (Note the word *attempts*; disregard that fact that it doesn't actually work.)

**True** **O.** Stemming is primarily recall-enhancing.

**False** **P.** In single word queries, polysemy is a serious problem, but word sense disambiguation can be used to find the correct sense.

**False** **Q.** If I search for term *X*, and term *X* has many synonyms, precision is more likely to be a problem than recall.

## Question 6. Question the Assumptions (12 points)

In class, we discussed four main simplifying assumptions that are made by modern document retrieval systems. They are listed below. For each assumption, ***briefly*** provide an example why it is not true in real-world information seeking environments. Each example should not take more than one sentence.

*Binary relevance*

Some documents are partially relevant.

*Document independence*

Reading one document can influence your relevance judgment on another document.

*Term independence*

Seeing "White House" in a document is not the same as see the term "white" and the word "house".

*Uniform priors*

Some documents in a collection are inherently of high quality than others.



## Question 7. Project (2 points)

Did you email Philip and me a brief description of your plans for the final project (or otherwise had discussions with me)?

**YES** NO

Have a good spring break!