בס"ד

איחזור מידע תשס"ז – 372.1.4406
סמסטר אביב מועד ב' 09.08.07
ד"ר ברכה שפירא, ארז שלום

משך הבחינה: שעתיים וחצי
חומר עזר מותר – לא מחשב נייד!

יש לענות על כל 6 השאלות:

1. 16% מנוע איחזור החזיר את הרשימה המדורגת הבאה לשאילתא מסויימת (√ מסמל מסמך רלוונטי ו – X מסמך לא רלוונטי). נתון שישנם סה"כ שמונה מסמכים רלוונטים לשאילתא במאגר.

1. √
2. X
3. √
4. X
5. X
6. X
7. X
8. √
9. √
10. X
11. X
12. X
13. X
14. X
15. √
16. X
17. X
18. X
19. X
20. √

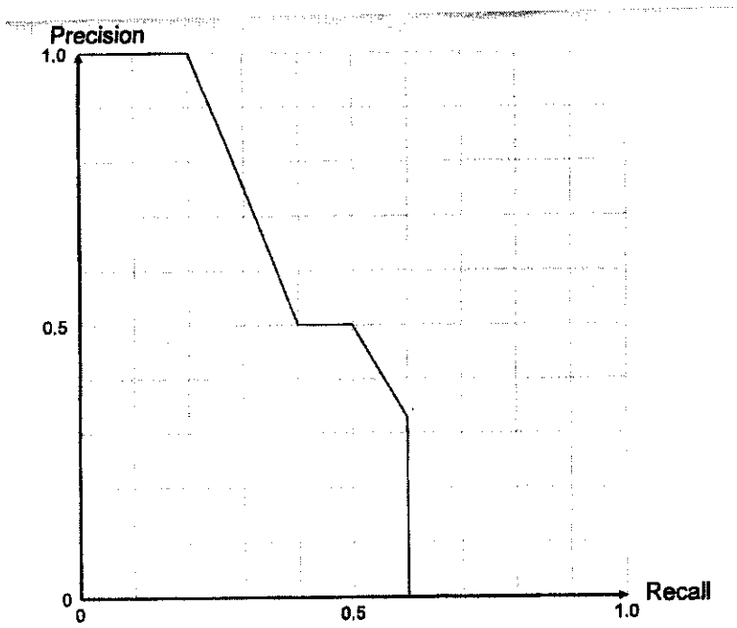
א. 4% - חשב את ה Mean Average Precision על פי שאילתא זו.
ב. 2% - מהו ה R-precision לשאילתא הזו?
ג. 2% חשב precision ב-10 מסמכים.
ד. 8% צייר 2 גרפים של -precision-recall של השאילתא. אחד על פי התוצאות המדוייקות של השאילתא, והשני על פי אינטרפולציה לנקודות recall סטנדרטיות.

2. 12% נתון גרף Precision-recall על תוצאות של הרצת מנוע מסויים על שאילתא מסויימת. המנוע החזיר 20 תוצאות. ידוע שבמאגר 10 מסמכים רלוונטים לשאילתא.

א. 3% מהו ה precision לאחר שהמנוע החזיר 3 מסמכים רלוונטים.

ב. 3% לאחר עיון ברשימת המסמכים שהמנוע החזיר אפשר לזהות סדרה של n מסמכים רצופים שכולם רלוונטים. מה הגודל המקסימלי של n?

ג. 6% הראה את הרשימה שהמנוע החזיר (הצג רשימה של 20 מסמכים מדורגים וסמן מי מהם רלוונטי ומי לא- כדוגמת הרשימה בשאלה 1)



3. 30% נתון המאגר הבא:

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

נתונה רשימת stopwords: an, and,do,in, not

א. 5% צור אינדקס למאגר עבור מנוע בוליאני טהור.

ב. 4% אילו מסמכים יוחזרו לשאילתות הבאות :

Interests NOT rates

(interest and rates) NOT (rising OR kids)

ג. 10% צור ווקטורים למסמכים על פי המודל הווקטורי – נרמל את ה Term לתדירות מקסימלית.

ד. 11% אילו מסמכים יחזרו ובאיזה סדר לשאילתא הבאה (חישוב הדמיון על פי קוסינוס) :

Interest rising

4. 20%

א. 15% יש לבצע clustering על המסמכים הבאים בשיטת k-means כאשר k=2 וחישוב הדמיון מתבצע על פי inner-product. המסמכים 2 ו- 4 נבחרו כ Seeds תחיליים. יש לשקלל את הterms על פי tf עם נרמול לאורך המסמך (ללא idf). יש להדגים חישוב של k-means clusters עד להתכנסות.

Doc1: movie movie Potter
Doc 2 : Movie Monsters
Doc3: Movie Potter
Doc4: Monsters Monsters

ב. 5% האם היה הבדל בתהליך אם ה seeds הראשוניים היו המסמכים 1, 4 , אם כן מה ההבדל?

5. 10% נתונה טבלת הדירוגים הבאה של 5 משתמשים (user1..user5) לחמישה items (A..E). כמו כן נתונה הקרבה ($W_{5,i}$) בין כל אחד מה-users ל user5 (המוגדר כ Active user). יש לנבא את הרלוונטיות של D item ל -user 5 על פי **שני** השכנים הקרובים אליו ביותר (יש לחשב את ($P_{5,D}$)).

Item	User1	User2	User3	User4	User5 (active user)
A	10	5	9		9
B	6	9		5	5
C	2	7	3		1
D	4	8	3	3	
E	8	1	9	2	
דמיון ל user5 $W_{5,i}$	1	-0.5	0.9	0.7	

6. 12% ציין נכון או לא נכון ליד כל אחד מהמשפטים הבאים:

א. 4% במנוע חיפוש באינטרנט r-precision הוא מדד חשוב

ב. 4% נניח מאגר מסמכים המכיל את המסמך השגוי הבא: "שלום שלום מה העניינים ואיך אתה מרגיש" – נניח שהטעות (מופע פעמיים של המילה שלום") התגלתה. יש לחשב מחדש את כל האינדקס כדי לתקן את הטעות.

ג. 4% שאילתא של מילה שיש לה מילים נרדפות רבות עלולה ל"סבול" מ recall נמוך.

בהצלחה
ברכה וארז

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

Midterm Exam

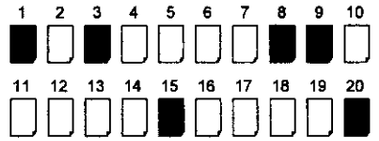
Name: Jimmy Lin

- 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?

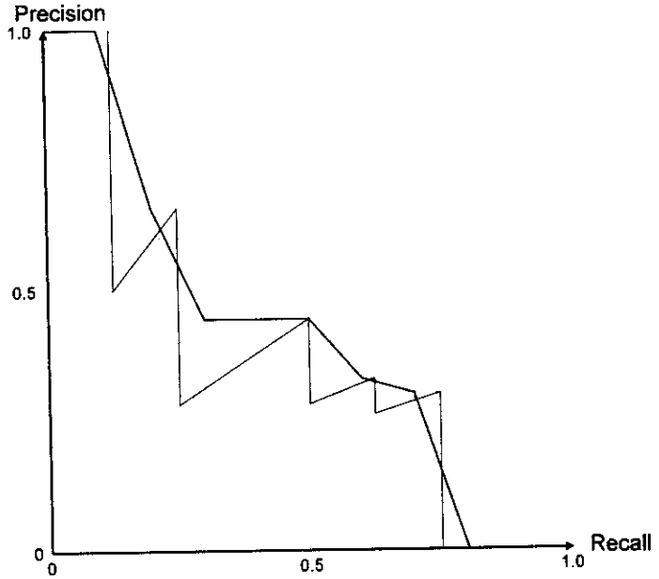
$3/8 = 0.375$

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

$4/10 = 0.4$

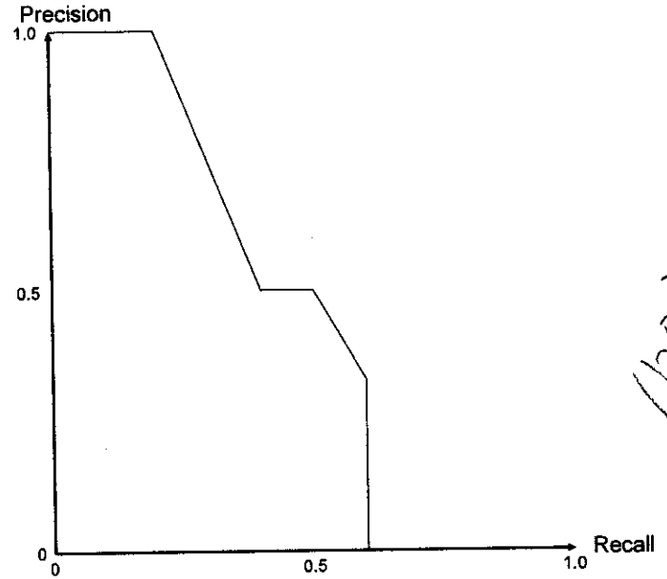
3 -1

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.

k **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.

7 **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.

4

שלב בחירת

- מחשבים tf של כל המילים.
- מחשבים sim בין ה seeds (Doc2, Doc4) לכל
שאר המסמכים האחרים. מסמך שאינו קרוב ל seed מסוים
מקבלת אליו. (Doc1, Doc3, Doc2 - יהיו ~~[illegible]~~ [illegible])
Doc4 יהיה ב cluster.

- מחשבים וקטור מרכזי של כל cluster - שנקרא
בשלב הקבוצה (ממוצע של המרכזים בכל כיתה).
- שוב בודקים קירבה של כל מסמך לוקטור המרכזי
של ה cluster
יתקבל

cluster1
(Doc4, Doc2)

cluster2
(Doc3, Doc1)

אפשר לראות שזה תקבוצה סטטי - לא פי הוקטור המרכזי
של ה cluster.

tf.

	movie	Monsters	Potter3
Doc1	$\frac{2}{3}$	0	$\frac{1}{3}$
Doc2	$\frac{1}{2}$	$\frac{1}{2}$	0
Doc3	$\frac{1}{2}$	0	$\frac{1}{2}$
Doc4	0	1	0

Cluster 1

Doc2 $(\frac{1}{2}, \frac{1}{2}, 0)$

Doc1, Doc3

$sim(Doc_1, cluster_1) = 0$

Cluster 2 : Doc4

Doc4 $(0, 1, 0)$

---

centroid $(0, 1, 0)$

Doc4, Doc2

centroid $\left(\left(\frac{2}{3} + 1\right)/3,\right.$

$\left(\frac{5}{9}, \frac{1}{6}, \frac{5}{18}\right)$

Doc1, Doc3.

$sim(Doc_2, cluster_1) = \left(\frac{5}{9} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{6} + 0\right)$ $\frac{5}{18}$ $\frac{1}{12}$

$\frac{4}{2}$

5 1,3 הם הדומים

$$(P_{r,u}) = \frac{\bar{r}_a + \sum_{u_i} + \sum (r_{u_i} - \bar{r}_u) \times w_{a,u}}{\sum w_{a,u}}$$

$$P_{r_1} = \frac{15}{3} + \frac{[(1 \times (4 - 30/5)) + (0.9 (3 - 24/4))]}{1 + 0.9}$$

$$= 2.526$$

6 אלא נכון

כי לא נכון

לא נכון.