אוניברסיטת בן-גוריון
המחלקה להנדסת מערכות מידע
מבחן בקורס איחזור מידע - 372-144-06

מרצה : ד"ר ברכה שפירא
אסיסטנטית : גב' מיטל טובי
תאריך : 19.01.05
משך המבחן: שעתיים וחצי, חומר עזר מודפס או כתוב - מותר
יש לענות על כל 6 השאלות.

1. (20) נתון המאגר הבא – הכולל רק את 3 המסמכים הבאים:

D1: readers read books if they are not tired
D2: a child reads childish booklets if they are readable and with childish pictures
D3: reading a book for children is very tiring.

נתונה רשימת Stopwords: are ,very ,they ,not ,if ,a ,b ,is ,for ,with ,has ,many ,and,
נתונה טבלת stems (look-up table) :

readers	read
reading	read
reads	read
readable	read
children	child
childish	child
books	book
booklets	book
tiring	tired

א. (10) הראה את ייצוג המאגר על פי מודל vector-space (ללא inverted File) לאחר תהליך של הסרת stopwords ולאחר ביצוע stem . לחישוב המשקל יש להשתמש ב- tf*idf, כאשר הנרמול מחושב על פי תדירות מקסימלית במסמך (אין לכלול log בנוסחת ה – idf).

ב. (5) הראה את ייצוג המאמר הנ"ל לאחר stem במודל ה vector-space כאשר נוסף לו Inverted-File הכולל הצבעות למסמכים בלבד (כלומר רק vocabulary ולא postings) – הצע סדר יעיל למסמכים המוצבעים בכל כניסה של term וסדר את ההצבעות לפיו.

ג. (5) מה תהיה תוצאת איחזור של השאילתא " children books" – כאשר מניחים ביצוע של אותו stem, וכן משקל אחיד ל-terms בשאילתא. הדמיון בין השאילתא למסמך מחושב על ידי inner-product תשובתכם צריכה לכלול את סדר המסמכים המוחזרים ואת חישוב הדרוג. (אין לכלול log בנוסחת ה – idf).

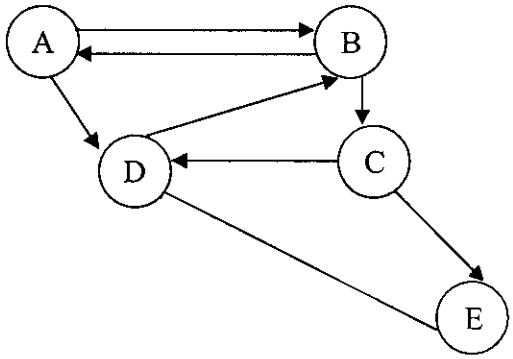
	read	book	tired	child	pictures
D_1	$\frac{2}{2}\cdot\frac{3}{3}=1$	$\frac{1}{2}\cdot\frac{3}{3}=\frac{1}{2}$	$\frac{1}{2}\cdot\frac{3}{2}=\frac{3}{4}$	0	0
D_2	$\frac{2}{3}\cdot\frac{3}{3}=\frac{2}{3}$	$\frac{1}{3}\cdot\frac{3}{3}=\frac{1}{3}$	0	$\frac{3}{3}\cdot\frac{3}{2}=\frac{3}{2}$	$\frac{1}{3}\cdot\frac{3}{1}=1$
D_3	$\frac{1}{1}\cdot\frac{3}{3}=1$	$\frac{1}{1}\cdot\frac{3}{3}=1$	$\frac{1}{1}\cdot\frac{3}{2}=\frac{3}{2}$	$\frac{1}{1}\cdot\frac{3}{2}=\frac{3}{2}$	0

ב) לפי tf·idf ולפי האות-כיא

book : $D_3 \to D_1 \to D_2$
child : $D_2 \to D_3$
pictures : D_2
read : $D_1 \to D_3 \to D_2$
tired : $D_3 \to D_1$

ג) $sim(d_1,q) = \frac{1}{2}+0 = \frac{1}{2}$ $\quad sim(d_3,q) = 1+\frac{3}{2} = \frac{5}{2}$
$sim(d_2,q) = \frac{1}{3}+\frac{3}{2} = \frac{11}{6}$ $\quad \Rightarrow D_3 \to D_2 \to D_1$

2. א. (12) crawler מסויים מסדר את תור הדפים שלו על פי ה pagerank שחישב ב-crawling קודם. נתונה הרשת הבאה :



Crawler מתחיל מצומת A

מה יהיה סדר סריקת הדפים- כלומר מה יהיה סדר הדפים שייצאו מהתור? – אין לחשב ערכים מדוייקים של pagerank אלא להשוות בין רמת pagerank של צמתים במידת הצורך. יש להסביר את התוצאה.

3.(23) שני מנועי חיפוש החזירו את התוצאות הבאות עבור שאילתא מסויימת (כאשר x מסמן מסמך רלוונטי)

במאגר 6 מסמכים רלוונטים לשאילתא זו.

מנוע 1	מסמכים רלוונטים	מנוע 2	
D1	X	D6	X
D2	X	D7	
D3		D3	
D4	X	D8	
D5	X	D9	
D6	X	D1	X
D7		D2	X
D8		D4	X
D9		D5	X
D10	X	D10	X

א) precision
recall
E-measure

ב) precision@5
avg-precision
R-precision

ג) כאשר אנו מחפשים מאמרים איקאמיים ונוכלה לעבור על כל המסמכים הרלוונטים

א. (5) ציין וחשב שני מדדים שבהם ההערכה של מנוע 1 תהיה זהה להערכה של מנוע 2.
ב.(15) ציין וחשב 3 מדדים שבהם ההערכה של מנוע 1 תהיה שונה (טובה יותר) ממנוע 2.
הערה לסעיפים א-ב : ציור של גרף נחשב כחישוב מדד.
ג. (3) תאר מציאות שבה מנוע 1 **אינו** עדיף על מנוע 2.

4. (15) א. (10) השווה בין שיטת עידון השאילתא המסתמכת על מילון (thesaurus) לעומת השיטה המוסיפה מילים שכיחות ביותר מבין המסמכים שקיבלו ציון גבוה ביותר בדרוג. ציין יתרונות וחסרונות של כל שיטה תוך ציון של לפחות 2 הבדלים ברורים בין השיטות.
ב. (5) ציין שתי סיבות שבגללן מנועי חיפוש מסחריים לא נוטים להשתמש ב relevance-feedback (משוב חוזר) לעידון השאילתא.

5. (20) נתונים המרחקים הבאים בין 5 מסמכים:

	D1	D2	D3	D4
D2	0.2			
D3	0.1	0.9		
D4	0.3	0.98	0.4	
D5	0.7	0.95	0.32	0.25

א. (15) הראה כיצד יתבצע על המסמכים הנ"ל תהליך clustering היררכי צובר (agglomerative) על פי Single-Link ועל פי complete-Link. הראה רק 3 איטרציות בכל אלגוריתם (כלומר צרוף של 3 clusters ראשונים)

ב. (5) תאר סוג של יישום הקשור לאיחזור מידע שעבורו רצוי להשתמש ב -clustering מסוג single-link – הסבר מדוע. יצירת clusters של פרופילים של משתמשים.

6. (10) בשני הסעיפים הבאים יש לבחור את התשובה הנכונה

6.1. מערכות IR מבוזרות שולחות את השאילתות למספר מנועים, מקבלות את אוסף התשובות ואז ממזגות את התשובות. בהנחה ששני מנועים בונים את המאגרים בשיטת ה vector-space (כולל tf*idf) – ובהנחה ששני מנועים הם זהים ורק המאגרים שהם מאנדקסים הם שונים, האם אפשר פשוט לדרג את התשובות על פי הערכים המתקבלים משני המנועים?

א. אפשר למיין לפי הדרוג של שני המנועים משום שהם מדרגים ומאנדקסים באותה שיטה בדיוק

ב. אי אפשר משום ש idf של כל מנוע מחושב על פי המאגר שלו וכדי שהדרוג יהיה נכון צריך לחשב את ה idf על פי האוסף הכולל המצוי בשני המאגרים.

ג. אפשר משום ש tf*idf מנורמל לגודל המאגר

6.2 ככל ש term מופיע ביותר ויותר מסמכים

א. ערך ה idf שלו יורד וכן רמת הדמיון בין המסמכים יורדת.

ב. ערך ה idf שלו יורד ורמת הדמיון בין המסמכים עולה

ג. ערך ה idf שלו עולה ורמת הדמיון בין המסמכים עולה.

1c)

	(1,3)	2	4
(1,3)	0		
2	0.2	0	
4	0.3	0.98	0
5	0.32	0.95	0.25

→

	(1,2,3)	4	5
(1,2,3)	0		
4	0.3	0	
5	0.32	0.25	0

בהצלחה
ברכה ומיטל.

	(1,3)	2	4
(1,3)	0		
2	0.9	0	
4	0.4	0.98	0
5	0.7	0.95	0.25

→

	(1,3)	2
(1,3)	0	
2	0.9	0
(4,5)	0.7	0.98