בס"ד

מבחן בקורס איחזור מידע וספריות דיגיטליות – ד"ר ברכה שפירא, ליהי נעמני
תשס"ט סמסטר ' מועד א', 29.03.09 13:30

חומר עזר מודפס מותר - לא מחשב נייד
משך המבחן: שעתיים וחצי
יש לענות על כל השאלות.

1. 15% כחלק מתהליך הכנת מאגר להערכת מנועי איחזור, נתנו לשני שופטים להעריך רלוונטיות של מסמכים עבור שאילתא מסויימת.
להלן תוצאות הערכתם ל- 4 מסמכים:

מסמך 4	מסמך 3	מסמך 2	מסמך 1	
לא רלוונטי	רלוונטי	רלוונטי	רלוונטי	שופט 1
רלוונטי	רלוונטי	לא רלוונטי	רלוונטי	שופט 2

על סמך נתונים אלה, בחר את המשפט הנכון משלושת המשפטים הבאים, נמק את תשובתך:

א. בין השופטים יש הסכמה טובה
ב. ההסכמה בין השופטים אינה טובה
ג. אי אפשר להסיק על ההתאמה בין השופטים

2. 20% למנוע חיפוש מסויים יש אינפורמציה על מספר הכניסות ביום של כל דף שמופיע באינדקס שלו.
   א. 10%הצע נוסחת דירוג המתחשבת גם באינפורמציה זו. הנח שנוסחת דירוג הנוכחית של המנוע כוללת כבר מרכיב IR ומרכיב של מבנה הרשת (כגון :pagerank). הנוסחא החדשה צריכה לשלב את כל המרכיבים.
   ב. 10%כיצד ניתן להעריך האם הנוסחא החדשה משפרת את תוצאות החיפוש (הצע שיטת הערכה **ישימה**).

3. 10%אחת השיטות להערכת גודל יחסי של מנועי חיפוש ב web היא לדגום דפים המאנדקסים במנוע אחד ולבדוק האם הם נמצאים באינדקס של המנוע השני.
ציין לפחות שתי סיבות להטיה אפשרית של הבדיקה הזו.

4. 20% שני מנועי חיפוש: e1,e2 הורצו על שתי שאילתות: Q1,Q2 . ידוע שלשאילתא Q1 10 מסמכים רלוונטים במאגר, לא ידוע מספר המסמכים הרלוונטים במאגר לשאילתא Q2.

להלן תוצאות ההרצה של השאילתא לשני המנועים (+ מסמן מסמך רלוונטי שהוחזר, - מסמן מסמך לא רלוונטי):

מנוע E1		מנוע E2	
Q1	Q2	Q1	Q2
+	+	-	+
+	+	+	-
+	+	-	-
+	+	-	+
+	-	+	-
-	+	+	+
-	-	+	+
-	-	+	+
-	-	+	+
-	-	+	-

א. 3% האם ניתן להסיק על מספר המסמכים הרלוונטים הנמצאים במאגר לשאילתא Q2 מתוך התוצאות לעיל.

ב. חשב, או הסבר למה אי אפשר לחשב:

a. 3% mean average precision על פני שתי השאילתות.

b. 3% mean average precision לשאילתא Q1 למונע E2.

c. 5% E-measure לשאילתא Q1 למנוע E1 כאשר ניתן משקל כפול ל precision לעומת ה recall.

ג. 6% אפיין את התנהגות כל אחד מהמנועים, (בהנחה ששתי השאילתות מייצגות את התנהגות הכללית).

5. 20% נניח מאגר הכולל רק את המסמכים : . d1,d2,d3 להלן נתונים על תדירות terms במסמכים על פ הפרוט הבא:

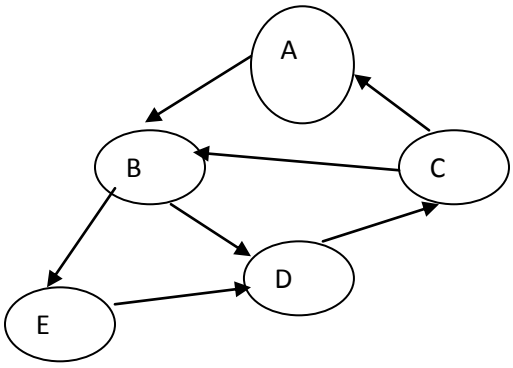
Terms	D1 -	D2	D3
Information	35	0	200
Retrieval	0	100	0
course	25	40	0

א. 6% ייצג את המסמכים באמצעות ווקטור משקולות על פי נוסחת tf*idf יש לנרמל את tf ל termהשכיח ביותר.

ב. 6% אילו שני מסמכים קרובים ביותר ביניהם (השתמש בנוסחת קוסינוס).

ג. 8% מדוע כדאי לחשב דמיון בין מסמכים במאגר? (תן דוגמא לשימוש אפשרי).

6. 15% נתונה הרשת הבאה:



א. 5% מהו הצומת/הצמתים בעלי ה pagerank **הנמוך** ביותר (אין צורך לחשב, אלא רק להעריך)

ב. 5% איך ישפיע ביטול הקישור בין D ל-C על ערכי ה Pagerank ברשת?

ג. 5% איך ישפיע ביטול הקישור בין B ל D (בנוסף על הביטול בסעיף ב') על ערכי ה pagerank ברשת?

בהצלחה

ברכה וליהי