בס"ד

מבחן איחזור מידע וספריות דיגיטליות

מועד ב' 17.04.09 09:00, – ד"ר ברכה שפירא, ליהי נעמני

משך המבחן - שעתיים וחצי

חומר עזר מותר (לא מחשב נייד!)

יש לענות על כל השאלות.

1. 25% מערכת סינון מבוססת תוכן מחזיקה פרופילים של תחומי עניין של שני משתמשים, U1, U2.

	ספורט	כלכלה	מוזיקה	פוליטיקה
U1	0.3	0.4	0.9	0.2
U2	0.7	0.6	0.1	0.8

סף הדמיון (הסינון) של המערכת הוא : 0.6 (פונקציית הדמיון מחושבת על ידי נוסחת קוסינוס)
למערכת הגיע המסמך הבא:
" נאמר על מר כהן שהוא עוסק בפוליטיקה, ולא אוהב ספורט ומתעניין בכלכלה בינלאומית"

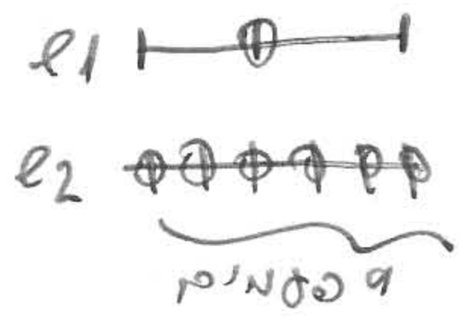
1.1 15% האם המסמך:
   א. יוצג לשני המשתמשים
   ב. יוצג ל-U1
   ג. יוצג ל-U2
   ד. לא יוצג לאף אחד מהמשתמשים.

הסבר (באמצעות חישוב) כיצד הגעת לתשובתך. הנח הנחות (מתאימות) לגבי אופן חישוב משקל המילים במסמך.

1.2 10% בהנחה שהמסמך הוצג למשתמש U2 והוא סימן אותו כלא רלוונטי. עדכן את הפרופיל שלו בהתאמה. (יש לעדכן רק מילים קיימות בפרופיל ולא להוסיף לפרופיל מילים חדשות).
הראה את החישוב ואת הפרופיל החדש. (המשקל של מסמך לא רלוונטי 0.25 והמשקל של פרופיל נוכחי 1).

2. 15% נתונים שני דפי Web : e1, e2. התוכן של e1 משתנה פעם ביום ושל - e2 משתנה 9 פעמים ביום. Crawler יכול לבצע סה"כ 2 עדכונים ביום. מהי אסטרטגיית העדכון המומלצת כדי למקסם freshness :
   א. לעדכן פעמיים ביום את e2 ו- 0 פעמים את e1.
   ב. לעדכן פעמיים ביום את e1 ו-0 פעמים את e2.
   ג. לעדכן פעם ביום את e1 ופעם ביום את e2.

יש לנמק את התשובה על ידי חישוב הערכה של ה freshness (expected freshness increase).



Authority - אז כמה תלוי מלבים

3. 30% נתונים 20 מסמכים ראשונים שחזרו כתוצאה משאילתא מסויימת Q של מנוע חיפוש מסויים (**לפי סדר הופעתם משמאל לימין**) כאשר R מסמן מסמך רלוונטי ו- N מסמן מסמך לא רלוונטי.

המאגר כולל בסך הכל 10,000 מסמכים וקיימים במאגר 6 מסמכים רלוונטים לשאילתא Q.

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

→

על פי נתונים אלה:

30%

א. 3% חשב precision@20

ב. 3% חשב r-precision

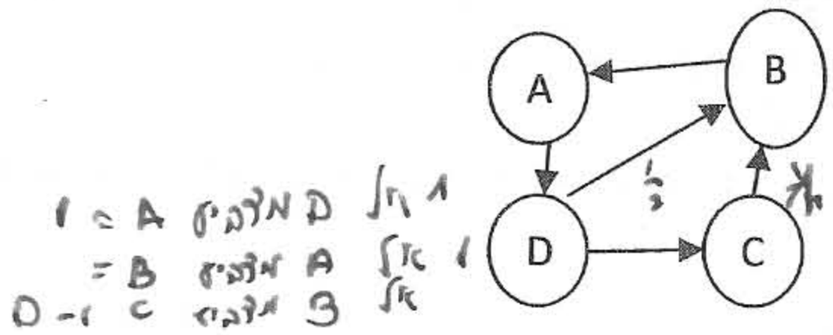
ג. 3% חשב f-measure ל 20 המסמכים שחזרו.

ד. 8% חשב precision בנקודות recall 0.5 ו- 0.8 (השתמש באינטרפולציה אם יש צורך).

ה. 3% חשב MAP

ו. 10% נניח שהמנוע החזיר את כל ה 10,000 המסמכים כתשובה לשאילתא Q, כאשר 20 המסמכים הראשונים שחוזרים הם אלה שלמעלה. מה יכול להיות ה MAP המקסימאלי שיושג על ידי המנוע, ומה ה MAP המינימאלי. הראה את החישוב.

4. 15% נתונה הרשת הבאה:



אז כמה מלבים
D = 2
A = 1
B = 1
C = 1

1 אז D מלבים A = 1
1 אז A מלבים B =
אז B מלבים C ו- D

4.1 5% הסבר מיהו הצומת/ צמתים עם ה Authority הגבוה ביותר על פי אלגוריתם HITS (לא לחשב-אלא להסביר על פי ההגיון של האלגוריתם)

4.2 5% הסבר מיהו הצומת/ צמתים עם ה Authority הגבוה ביותר על פי אלגוריתם Pagerank (לא לחשב, אלא להסביר על פי ההגיון של האלגוריתם)

4.3 5% הסבר למה הצומת עם ה Authority הגבוה ביותר זהה או שונה על פי שני האלגוריתמים.

הבדל בין PR ל- Hits
אי תלוי בשאילתא
תלוי בשאילתא

5. 15% משתמש שלח למנוע שאילתא שכוללת את ה term - sophistecatid שאינו מופיע באינדקס.

הסבר, איך אפשר לקבוע (על פי שתי שיטות שונות), האם המשתמש טעה והתכוון ל - term: sophisticated שנמצא באינדקס. ציין מה הם הפרמטרים הדרושים לכל שיטה, קבע פרמטרים בהתאמה וחשב רק על פי אחת מהשיטות האם המשתמש התכוון ל sophisticated.

בהצלחה
ברכה וליהי.