בס"ד

איחזור מידע תשע"א – 372.1.4406
סמסטר חורף מועד ב' 17.02.11
ד"ר ברכה שפירא, אורלי מורנו

משך המבחן : שעתיים וחצי
חומר עזר: מותר (**לא מחשב נייד**)
יש להחזיר את השאלון

מס נבחן 023093

חלק א – נא לענות במחברת הבחינה

1. 26% נתון log של שאילתות שמשתמשים שלחו למנוע, ולכל שאילתא סט של 30 מסמכים שחזרו לשאילתא. כמו כן, נתון לכל שאילתא על אילו מסמכים מתוך אלו שחזרו המשתמש הסתכל. כדי לנתח את האיכות של המנוע צריך לזהות מתי המשתמש חיפש באותו נושא – כלומר מבין השאלתות של המשתמש- אילו מהן היו בהקשר לאותו צורך מידע בהנחה שאין זיהוי של משתמש על השאילתא (לא ידוע אילו מהשאילתות התבצעו על ידי אותו משתמש ברצף).
   א. 18% הצע פיתרון (אלגוריתם בפסאודו קוד) שיקבץ את השאילתות של משתמש ברצף ל session – , כלומר לזהות באותו רצף את השאילתות של משתמש אחד המנסה למצוא מידע בנושא מסוים ומתי מתחיל session של משתמש אחר (או אותו משתמש בנושא שונה).
   ב. 8% הסבר מהן המגבלות של הפתרון – כלומר באילו מקרים הוא לא יצליח לזהות את ה Session (אם אין כאלה מה טוב.....הסבר מדוע אין מגבלות)

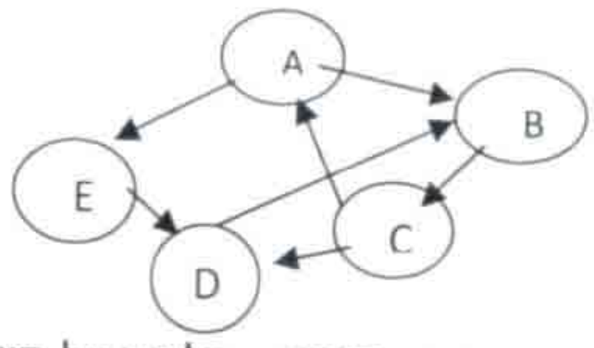
2. 8% נניח crawler שעובד ב batch ושומר את ה Repository באופן מבוזר. הביזור מתבצע תוך כדי ה crawling כאשר ה crawler מוריד את הדף ואז שומר אותו בצומת לפני הטיפול שמתבצע על הדף ("טיפול" למשל הוצאת הלינקים וכו'). מנה שתי סיבות שבגללן עדיף ל crawler לצורך ייעול עבודתו, לבזר את הנתונים בצמתים כך שבצומת אחת יישמרו כל הדפים של אתר מסוים –על פני ביזור של דפים שונים של אותו אתר בצמתים שונים. (הסיבות צריכות להיות קשורות לייעול עבודתו של ה crawler – לא לאינדוקס בהמשך).

3. ✓ 16% למאגר מסוים ואוסף שאילתות התבצעה הערכה של מומחים על הרלוונטיות של כל מסמך לשאילתא (לצורכי הערכה של מנועים). ההערכה כללה 3 סוגי תוצאות (במקום ה 2 הסטנדרטיות) : 0 – כאשר המסמך לא רלוונטי, 1 כאשר מסמך רלוונטי חלקית ו -2 כאשר המסמך רלוונטי לגמרי.
   - א. 13% כתוב נוסחה מדויקת שמגדירה את מדד precision@k בהתאמה לתוצאות כאלו.
   - ב. 3% הדגם את המדד precision@5 שהגדרת על התוצאות הבאות (משמאל לימין):

1,0,2,0,0,2,1,1

חלק ב – יש לענות במחברת הבחינה במקומות המסומנים

4. 15% נתונה רשת :



השלם: (תשובה לא נכונה קונסת בחצי נקודה – כלומר על תשובה לא נכונה יופחתו 4.5 נקודות).

- 5% הצומת בעלת ערך ה hub **הגבוה ביותר** על פי אלגוריתם HITS הוא : A
- 5% בין הצמתים E,B,D הצומת בעלת ערך authority **נמוך ביותר** על פי אלגוריתם HITS היא E
- 5% 2 הצמתים בעלי ה pagerank הגבוה ביותר הם B,D

-3

B,C

סמן נכון או לא נכון (הסברים במחברת הבחינה)

5. 17%

א. 3% ערך tf*idf של term במסמך יכול להיות גבוה יותר מ- 1. נכון/לא נכון

ב. 3% שימוש ב Stemmer יכול להשפיע על שביעות רצון המשתמש מהמנוע. נכון/לא נכון

ג. 3% צומת בגרף מסוים שהיא בעלת ה pagerank הגבוה ביותר בגרף, בהכרח תהיה גם בעלת ה Authority הגבוה ביותר. נכון/לא נכון

ד. 4% הנח שתי מילים "teach" ו "taught" ששקללו אותן כשתי מילים נפרדות במסמכים בהן הן הופיעו, על פי tf*idf סטנדרטי. לאחר מכן הסתבר ששתי המילים הן בעצם stem של אותה מילה ושאפשר לייצג אותן במשותף כמילה אחת (בהנחה שהמנוע תומך ב stem). כדי לתקן את הטעות, צריך לסכום את ה tf*idf שלהן לייצוג המשותף.
נכון/לא נכון

ה. 4% נתון פרופיל משתמש לאחר עדכון על פי תגובת משתמש על פי אלגוריתם rocchio לאחר ניתוח תגובתו למסמכים d1 d2 d3. בהנחה שאין הפחתה שלילית (כלומר **γ=0**). אפשר לראות שהמשתמש העדיף פוליטיקה ומדע על פני מוזיקה. נכון/לא נכון

	ספורט	מוזיקה	פוליטיקה	מדע	רמטכ"ל	לימודים
פרופיל לאחר עדכון	0.3	0.2	0.8	0.8	0	0
מסמך d1	0.2	0	0.8	0.9	0	0
מסמך d2	0	0.9	0.2	0	0.2	0.1
מסמך d3	0.4	0	0.8	1	0	0

6. 6% נתון מאגר ובו 4 מסמכים :

D1- Tibet Tibet Malaga

D2- Malaga Rimini Salvador Tibet Tibet Tibet Tibet

D3- mexico Sun

D4- Mexico Malaga Tibet Sun

נתונים שלושה מנועים – E1 עובד לפי המודל הבוליאני הטהור, E2 לפי המודל הווקטורי, E3 לפי מודל בוליאני מורחב שמדרג את המסמכים לאחר הפעלת האופרטורים הבוליאנים בשאילתא.

השאילתא Malaga and Tibet נשלחה ל E1 ול E3

השאילתא Malaga Tibet נשלחה ל E2

א. 3% E1 ו E2 יחזירו תשובה זהה לשאילתות שנשלחו אליהם. נכון/לא נכון

ב. 3% E1 ו E3 יחזירו את אותם מסמכים לשאילתא שנשלחה אליהם. נכון/לא נכון

-6

12% סמן תשובה אחת נכונה לשאלות הבאות

7. 4% מנועים מסוימים משתמשים בתגובות משתמשים לתשובות על שאילתא q1 ומריצים בעקבותיה שאילתא נוספת q2. (relevance feedback). שיטה זו :

א. משפרת תמיד את ה Recall של שאילתא q1

ב. עשויה לשפר את ה precision של שאילתא q1

ג. אינה משפיעה על ביצועי המנוע אלא רק משפרת את שביעות רצון המשתמשים

ד. משפרת precision ו recall של שאילתא q2

8. 4% לשאילתא Q 4 מסמכים רלוונטים במאגר. להלן תוצאות שחזרו משני מנועים: E1 ו-E2.

- התוצאות משמאל לימין כשאר N מסמל מסמך לא רלוונטי, R מסמן מסמך רלוונטי:
  - E1: R N R N N N N N R R
  - E2: N R N N R R R N N N

R = 4

א. ל MAP E1 גבוה יותר מאשר ל E2

ב. R-precision של שני המנועים שווה

ג. R-precision של E1 גבוה מ R-precision של E2

ד. MAP של E2 גבוה מ MAP של E1

ה. א+ג נכונים

ו. א+ב נכונים

ז. ג+ד נכונים

9. 4% נתון גרף precision-recall. הגרף משקף תוצאות של שאילתא אחת.

א. הגרף התקבל בהכרח מנתונים אמיתיים ללא אינטרפולציה

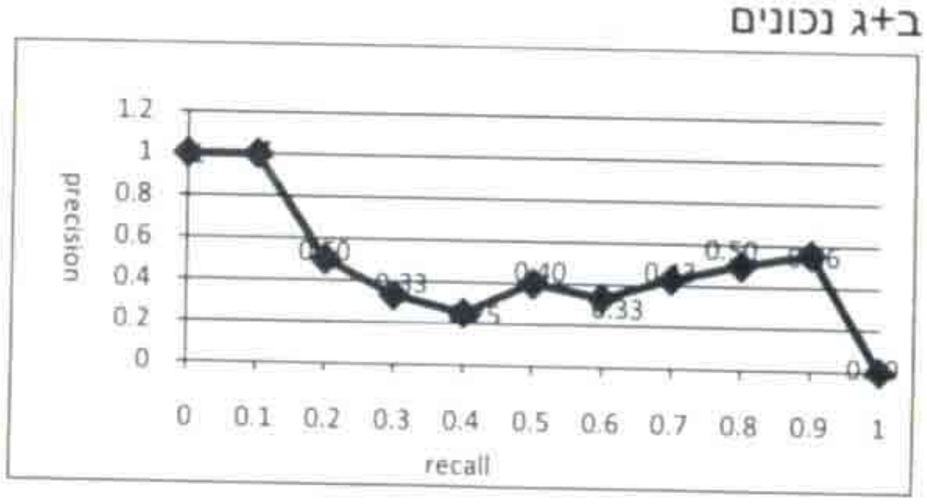
ב. הגרף התקבל בהכרח לאחר אינטרפולציה של ערכים

ג. כל המסמכים הרלוונטים לשאילתא חזרו.

ד. 90% לפחות מהמסמכים הרלוונטים לשאילתא חזרו

ה. ב+ד נכונים

ו. ב+ג נכונים



בהצלחה

ברכה ואורלי