שאלה 1. 15-
שאלה 2. 8-
שאלה 3. 8-

חלק ב
-9 62/2

אוניברסיטת בן-גוריון בנגב

הוראות לנבחן בצידו השני של הדף

אין לכתוב מעבר לשוליים משני צידי הדף

מס' כיתה 16 בנין 34

מס' נבחן 023093

מדבקה

1

מדבקה

המחלקה _____ שנה _____

תאריך בחינה _____

מקצוע בחינה _____

לשימוש המרצה הבודק

	יחידות	עשרות	מאות
0	☐	☐	
1	☐	☐	☐
2	☐	☐	
3	☐	☐	
4	☐	☐	
5	☐	☐	
6	☐	☐	
7	☐	☐	
8	☐	☐	
9	☐	☐	

ציון הבחינה 76

שם המרצה _____

חתימה רונה שפנר

תאריך _____

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)

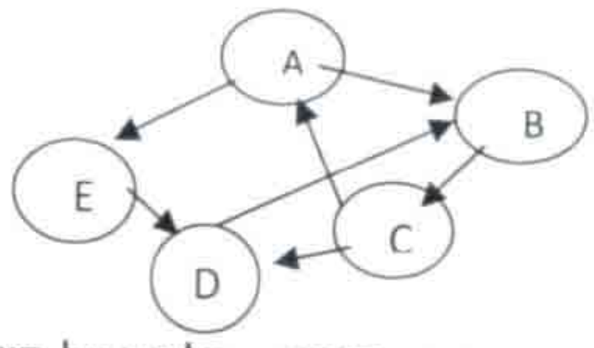
2. 8% נניח crawler שעובד ב batch ושומר את ה Repository באופן מבוזר. הביזור מתבצע תוך כדי ה crawling כאשר ה crawler מוריד את הדף ואז שומר אותו בצומת לפני הטיפול שמתבצע על הדף ("טיפול" למשל הוצאת הלינקים וכו'). מנה שתי סיבות שבגללן עדיף ל crawler לצורך ייעול עבודתו, לבזר את הנתונים בצמתים כך שבצומת אחת 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

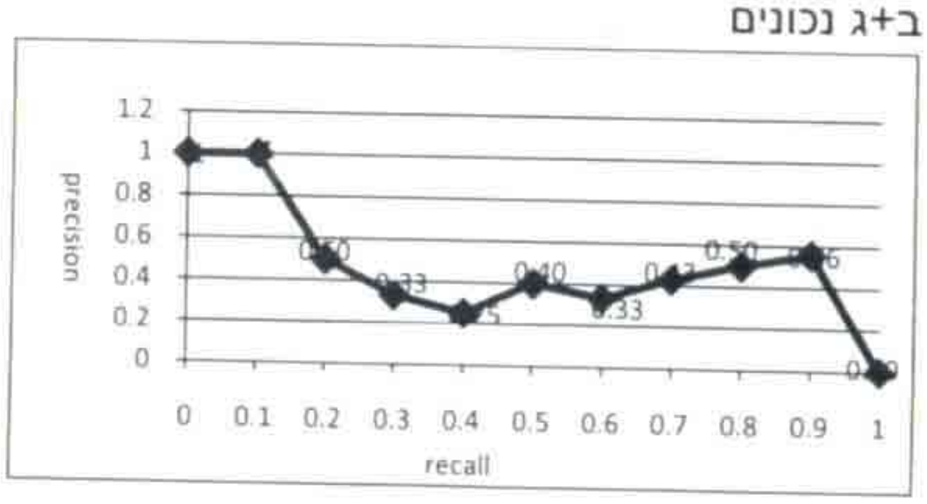
ב. הגרף התקבל בהכרח לאחר אינטרפולציה של ערכים

ג. כל המסמכים הרלוונטים לשאילתא חזרו.

ד. 90% לפחות מהמסמכים הרלוונטים לשאילתא חזרו

ה. ב+ד נכונים

ו. ב+ג נכונים



בהצלחה

ברכה ואורלי

חלק ב' -

(4) תחילה נחשב את הצומת בעלת ערך ה- hub הגבוה ביותר עפ"י HITS:

$H(A) = A(B) + A(E)$

$H(B) = A(C)$

$H(C) = A(A) + A(C)$

$H(D) = A(B)$

$H(E) = A(D)$

צמתים A ו- C מצביעים על יותר צמתים מכל שאר הצמתים האחרים ולכן הם המועמדים להיות עם ה-HUB הכי גבוה. נתמקד רק בשניהם וננסה להבין מי יותר גדול:

(*) $H(A) = A(B) + A(E) = 2H(A) + H(D)$

$A(B) = H(A) + H(D)$ $\qquad$ $A(E) = H(A)$

$H(C) = A(A) + A(C) = H(C) + H(B)$

$A(A) = H(C)$ $\qquad$ $A(C) = H(B)$

ניתן לראות כי A מצביעה על יותר צמתים ובנוסף יותר צמתים מצביעים עליה ולכן זוהי הצומת בעלת ערך ה- Hub הגבוה ביותר.

מבין הצמתים E,B,D - הצומת בעלת Authority הנמוך ביותר היא:

$A(B) = H(A) + H(D)$

$A(D) = H(E) + H(C)$

$A(E) = H(A)$

ניתן לראות באופן מיידי כי $A(E) < A(B)$ ולכן רק נשאר לבדוק מה היחס בין $A(E)$ ל- $A(D)$.

(*) $A(E) = H(A) = A(B) + A(E)$

(*) $A(D) = H(E) + H(C) = A(D) + A(A) + A(C)$

לכן ניתן לראות שצומת E בעלת ערך Autority נמוך יותר כי היא מצביעה על פחות צמתים וגם פחות צמתים מצביעים עליה.

המשך שאלה 4

2 הצמתים בעלי ה- Pagerank הגבוה ביותר הם:

$$PR(A) = \frac{PR(C)}{2}$$

$$PR(B) = \frac{PR(A)}{2} + PR(D)$$

$$PR(C) = PR(B)$$

$$PR(D) = \frac{PR(C)}{2} + PR(E)$$

$$PR(E) = PR(A)$$

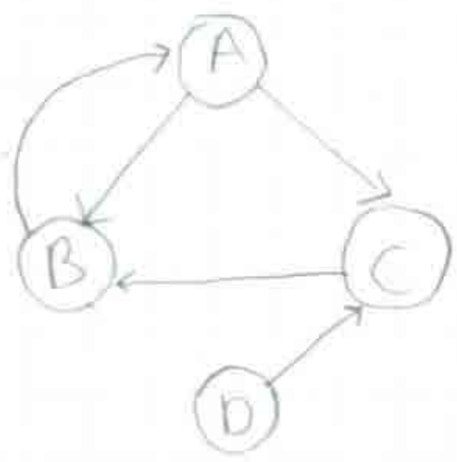
ניתן לראות כי צומת B מקבלת את כל ה- PR מצומת D ומדרג חצי מה- PR של A וכנ"ל גם צומת D מקבלת את כל PR(E) ומדרג חצי מ- PR(C), כ"א שהצמתים האלו מקבלים את ה- PR הכי גבוה בגרף

שאלה 5

א. נכון, צריך לה יכול להיות גדול מ-1 וכה תלוי אם מחליטים לנרמל
את tf / idf או את שניהם יחד. גם בלמידת הביט התבקשנו לחשב
idf מבלי לחשב את ה-log ולכן קיבלנו צריך idf גדול מ-1 מה
שהביא אותנו לתוצאה של W_{ij} שגדול מ-1. לכן לה תלוי מימוש.

ב. נכון, שימוש ב-Stemmer בדרך מביא יותר תוצאות רלוונטיות כי
לכל מילה בשפה יש הטיות רבות ואם נרצה להתייחס רק לשורש המילה אז נוכל לקבל
הרבה יותר תוצאות מתאימות. כמובן שכאשר ישנן יותר תוצאות רלוונטיות אל גם
שביעות הרצון של המשתמש מהמנוע עלה. יש לציין שבמקרים מסוימים Stemmer
יכול גם לטעות מילה מסוימת וכתוצאה מכך תתקבל תוצאה לא רלוונטית אבל ברוב
המקרים Stemmer דווקא מצר לשפר את הרלוונטיות.

ג. לא נכון.



$P(A) = P(B)$

$P(B) = \frac{P(A)}{2} + P(C)$

$P(C) = P(A) + P(D)$

$P(D) =$

ד. נניח שבמסמך d1 שיש הופיע - teach פעמים 1 - taught פעם אחת
נניח שסך אורך המסמך הוא 4
הנרמול הוא לפי אורך המסמך
אם המילה taught 1 - teach
היו באותו מסמך אז צריך
לספור אותן לאחר השינוי
פעם אחת עבור ה-idf 0.84
ולעומת זאת לפני השינוי
צריך לספר פעמיים

$tf(teach) = \frac{2}{4}$

$tf(taught) = \frac{1}{4}$

$idf(teach) = \log_2 \frac{5}{2}$

$idf(taught) = \log_2 \frac{5}{3}$

$d_1 \rightarrow$ teach $tf \cdot idf = 0.66$
taught $tf \cdot idf = 0.18$

אחרי השינוי $tf = \frac{3}{4}$, $idf = \log_2 \frac{5}{3} \Rightarrow tf \cdot idf = 0.55$

סה"כ הבעיות להרשאות מקרה שבו חיפוש התוכאות שונה ולכן התשובה היא <u>שגוי נכון</u>

ה)

$$\overrightarrow{q_m} = (0, 0, 0.8, 0.8, 0.2, 0.3)$$

שאלה 6

E1 ו- E3 מתקיים לפי מודל בולינאני טהור ומודל וקטורי מורחב בהתאמה.

Malaga and Tibet ושלום ל- E_1 ו- E_2

Malaga Tibet ושלום ל- E2

א) מנוע E1 יחזיר את D1, D2, D4

נבדוק מה יחזיר מנוע E2 לשאילתא "Malaga Tibet":

נניח שהמודל הוקטורי מנרמל את tf לפי אורך המסמך אז

$$Sim(D1, q) = \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 1 = 1$$

$$Sim(D2, q) = \frac{1}{7} \cdot 1 + \frac{4}{7} \cdot 1 = \frac{5}{7} = 0.71$$

$$Sim(D4, q) = \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1 = 0.5$$

מנוע E2 יחזיר D1, D2, D4

ב) E1 ו- E3 הם מנועים שמתקיים לפי מודל בולינאני ולכן בשלב החזרת מסמכים מתאימים הם יחפשו התאמה מלאה לכל המסמכים שיש בהם גם Malaga וגם Tibet ← לכן הם יחזירו את אותם מסמכים בדיוק אבל לא בטוח שלה יהיה באותו סדר כי לה תלוי בדירוג שיקבל המודל הבולינאני המורחב.

שאלה 7

לא תמיד ה- precision ישתפר בעקבות הפידבק שיתקבל מהמשתמש כי לה תלוי גם במנוע עצמו והאופן שבו הוא אינדקס את המסמכים. בפוטנציאל הרעיון של הפידבק הוא לשפר את ה- recall וה- precision אך זה לא תמיד קורה

שאלה 8

נחשב את המדדים:

$$MAP1 = \frac{1/1 + 2/3 + 3/9 + 4/10}{4} = 0.6$$

$$MAP2 = \frac{1/2 + 2/5 + 3/6 + 4/7}{4} = 0.49$$

$$R\text{-precision}(E1) = \frac{2}{4} = 0.5$$

$$R\text{-Precision}(E_2) = \frac{1}{4} = 0.25$$

תשובה ב+ג נכונה.

שאלה 9

	recall	Precision
1	$1/10 = 0.1$	$1/1 = 1$
2		
3		
4	$\frac{2}{10} = 0.2$	$2/4 = 0.5$
5		
6		
7		
8		
9	$\frac{3}{10} = 0.3$	$3/9 = 0.33$
10		
11		
12		
13		
14		
15		
16	$\frac{4}{10}$	$4/16 = 0.25$
17		
18		
19		
20		

הגרף היה חייב להתקבל אחרי תהליך אינטרפולציה ואפשר להוות לפי הטבלה ההתחלתית שכתבתי.

לפי תוצאות הגרף רואים שלפחות 90% מהתוצאות תגרו.

חלק א'

שאלה 1-

בקובץ log יש לכל שאילתא ← 30 מסמכים שחזרו לשאילתא

על אילו מסמכים המשתמש הסתכל מה שחזר

א) תחילה נחלץ מבנה נתונים שיכין עבור שאילתא מסוימת לאיזה נושא (concept) היא שייכת, לשם נפרק את ה- terms של השאילתא ונשייך אותה לנושא מסוים.

השיוך לנושא מסוים יעשה ע"י המסמכים שעליהם הסתכל המשתמש ← למשל אם משתמש הסתכל על $d_1, d_{13}, d_{14}, d_{29}$ אז אנחנו יודעים שמסמכים אלו שייכים לנושא שמתאים לשאילתא.

בדומה למודל LSI שלמדנו שיודע לקשר בין terms-concept-document כאשר נפרק שאילתא ל- terms ונדע מה המשתמש הסתכל אז נדע אילו מסמכים שייכים לאיזה ונוכל לקשר את ה- concept

התחלת טוב! חבל שלא המשכת!

-15-

חבל שמענת על השאלות בפרק ד' בפרוטרוט

הלא כן בהחלט לפי את התשובות

ללא הסברים אבל תוסבר היטב למנוע

ב) לא הספקתי לסיים את סעיף ג' מחוסר זמן אבל המגבלות של הפתרון שרציתי

לה יש קשורות למצב שבו שתי משתמשים הריצו שאילתות שונות שמתייחסות

לנושאים שונים אבל הנושאים האלו מאוד קרובים (דומים) אחד לשני ולכן יהיה

קשה לצהות את השוני ביניהם.

V

שאלה 2

הסיבות שבגללן עדיף ל- Crawler לשמור בזמנת אחת את כל הדפים מאתר מסוים:

1) שליפת URLs - אם ה- Crawler ישמור את כל הדפים של אתר מסוים בזמנת אחת אל יהיה לו קל יותר בתהליך של הוצאת הלינקים מהדפים כי במידה ויש לינק מסוים אחד באתר לעמוד אחר באותו אתר (דבר שהוא מאוד נפוץ באתרי web) אל אם הדף האחר יהיה ממוקם בזמנת אחרת ייקח יותר זמן ל- crawler לגשת לדף הזה מכל מיני סיבות: תעבורת רשת, עומס על הזמנת וכו' ברוב המוחלט של המקרים, ישנו קשר חזק בין דפים ששייכים לאותו אתר אינטרנט וכמות הלינקים שיהיו בתורם לדפים אחרים באותו אתר כנראה שיהיה גבוה.

2) בדיקת להוראות הדפים (אתיקה) - בעלי אתרים נוהגים לשים קובץ robots.txt או לציין ב- Metadata של דף מסוים האם מותר לאינדקס דף מסוים ואיזה חלקים מתוכו בכלל מותר לאינדקס. אם נבצר את האתר על פני כמה זמנים אל יהיה לנו קשה לנהל את הבדיקות להוראות הדפים כי בכל זמנת נצטרך לבדוק את התיקיות שאוסרו עלינו לאינדקס, התחבוק של כל הנושא הזה יהיה הרבה יותר קשה מאשר אם הכל יהיה על זמנת אחת ונדע בדיוק מה מותר ומה אסור לאינדקס בכל האתר כולו.

באופן כללי, התחבוקה של ביצור אתר אחד על פני כמה אתרים הוא קשה במיוחד, צריך לדעת להתנהל במקביל כל הזמן ולפעמים יש בעיות שלאו תלויות ב- Crawler שלאו יאפשרו לו לעשות את זה.

לפעמים יכול להיות מצב שבו זומנת אחת נפלה ואיחסנו שם חלק מהאתר ובעת לא נוכל להמשיך בפעולת ה- crawler כי הוא תלוי בדף הזה שנמצא באתר שנפל.

שאלה 3

א) עד היום חישבנו $Precision = \frac{\text{מס' רלוונטים שחזרו}}{\text{מס' המסמכים שחזרו}}$

מכיוון שכעת צריך לקחת בחשבון גם את הסקאלה של הדרגת מסמכים

אנו נתבסס על זמה נוסחה שאנחנו מכירים רק שכעת במונה נתייחס למס' הרלוונטים

שחזרו באופן קבלת שונה:

1) אם ההדרגה היא 0 אנו מס' רלוונטים שחזרו = 0

2) אם ההדרגה היא 1 אנו מס' הרלוונטים = 0.5

3) אם ההדרגה היא 2 אנו מס' הרלוונטים = 1

בעצם נירמלתי את ההדרגה לפי הבין הגבוה ביותר.

$$Precision @ k = \frac{\Sigma(\text{רלוונטים עד לנק' } k)}{k}$$

רלוונטים עד לנק' k ייספרו באופן שציינתי למעלה.

הרעיון הוא שאם יש מסמך שמתאים רק באופן חלקי אנו בעצם הוא חצי רלוונטי

ולכן ניתנים לו משקל נמוך יותר בספירה.

ב) נבדיק את האלגוריתם על $Precision @ 5$.

1, 0, 2, 0, 0 — תוצאות השאילתא (עד k=5)

0.5, 0, 1, 0, 0 — אחרי נרמול המשקל

$$Precision @ 5 = \frac{0.5 + 1}{5} = 0.3$$