-22- 8/30 חלק א'

חלק ב'

20/20 שאלה 1

20/20 שאלה 2

אוניברסיטת בן-גוריון בנגב

הוראות לנבחן בצידו השני של הדף

אין לכתוב מעבר לשוליים משני צידי הדף

מס' כיתה 226 בנין 90

מס' נבחן 0085564

שאלה 1 15

שאלה 2 15

25/03/2012 1352180 1 9/9

הנדסת מערכות מידע

אחזור מידע וספריות דיגיטליות

03721440601402

המחלקה מ. מדע מידע שנה 3

תאריך בחינה 25/3/12

מקצוע בחינה אחזור מידע וספריות דיגיטליות

לשימוש המרצה הבודק

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

ציון הבחינה 78

שם המרצה רפית שפשש

חתימה

תאריך

אוניברסיטת בן-גוריון בנגב

הוראות לנבחן

1. בהגיעך למקומך יש להניח את כרטיס הנבחן ותעודה מזהה על שולחנך.
2. אסור להביא למקום הבחינה תיקים, ספרים, מחברות, טלפון נייד או רשימות פרט למותר על פי שאלון הבחינה.
3. עזב תלמיד את האולם אחרי חלוקת השאלונים, דינו כדין "נבחן" בבחינה.
4. אסור לנבחן לשוחח בזמן הבחינה, או לעזוב את מקומו ללא נטילת רשות.

מס נבחן 9/9

בס"ד

איחזור מידע תשע"א – 372.1.4406

סמסטר חורף מועד ב' 25.03.2012

פרופ' ברכה שפירא, אורלי מורנו

משך המבחן : שעתיים וחצי

חומר עזר: מותר (**לא מחשב נייד**), **מותר מחשבון**

יש להחזיר את גיליון הבחינה. המבחן כולל 4 דפים. יש לענות על כל השאלות

חלק א 30% - יש לענות על הגליון

בחר תשובה אחת נכונה

1. 4% הפרמטרים בנוסחת rocchio ל relevance feedback :

א. קובעים האם המשתמש סימן את המסמכים כרלוונטים או לא.
ב. נקבעים דינמית לפי תגובות המשתמש למסמכים
ג. צריכים להסתכם ל – 1
ד. קובעים את רמת ההתחשבות בשאילתא הנוכחית, במסמכים הרלוונטים ובמסמכים הלא רלוונטים בעדכון השאילתא
ה. ג+ד נכונים
ו. ב+ג נכונים

2. 6% נתונה טבלת הדירוגים הבאה של משתמשים על מוצרים בדירוג 1-5:

	User 1	User 2	User 3	User 4	User 5
Item1	1			1	
Item2		2	3		3
Item3	1	1	5	1	
Item4				4	
Item5	5		1		5

כדי לנבא דירוג של item5 עבור User4 על פי שיטת הדמיון בין המשתמשים:

א. יש לחשב דמיון בין user4 לכל אחד מהמשתמשים האחרים ולחשב את הדירוג על פי המשתמשים הדומים ביותר
ב. יש לחשב דמיון בין user4 ל user3 | User1 ולהשתמש בדירוגים של שניהם ל item5 כדי לחשב את הדרוג של Item5 ל user4
ג. יש לחשב דמיון בין user4 ל user3 ו User1 ולהשתמש רק בדרוג של user1 ל item5 כדי לחשב את הדרוג של Item5 ל user4
ד. יש לחשב דמיון בין user4 ל user3 ו User1 ולהשתמש רק בדרוג של user3 ל item5 כדי לחשב את הדרוג של Item5 ל user4
ה. על סמך הנתונים הקיימים אי אפשר לנבא דירוג של item5 עבור User4
ו. יש לחשב דמיון בין user4 ל user 5 ובין user 1 ו user 3

3. 5% המשקל של term בווקטור של מסמך במודל הווקטורי
   א. מבטא את מידת הייצוגיות של ה term למסמך
   ב. צריך להיות מחושב באותו אופן שבו מחשבים את משקל ה Terms בשאילתא
   ג. מחושב תמיד על פי tf*idf
   ד. חייב להיות בין 0-1
   ה. ב+ג נכונים
   ו. ג+ד נכונים
   ז. א+ד נכונים

4. 4% Benchmark
   א. מגדיר מדדי ביצוע
   ב. משמש להשוואה של ביצועי אלגוריתמים שונים מול מדדים חדשים שהוגדרו
   ג. משמש להשוואה של ביצועי אלגוריתמים שונים מול תוצאות נכונות ידועות
   ד. א+ג נכונים
   ה. מחזק את הקשר בין האקדמיה לתעשיה
   ו. א+ג+ה נכונים

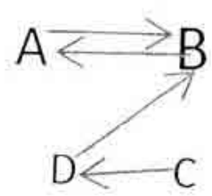
5. 3% - ההחלטה אם להתחשב ב Robots.txt תלויה ברמת האתיקה של אלגוריתם ה Crawling
   ענה נכון או לא נכון
   נכון/לא נכון

6. 4% - DCG הוא מדד לבדיקת יכולת של מנוע לדרג גבוה דפים רלוונטים, תוך התחשבות ברמת הרלוונטיות של הדף
   נכון/לא נכון

7. 4% stemming הוא תהליך התלוי תמיד בכללי שפה ולכן יש צורך להגדיר לכל שפה stemmer נפרד.
   נכון/לא נכון

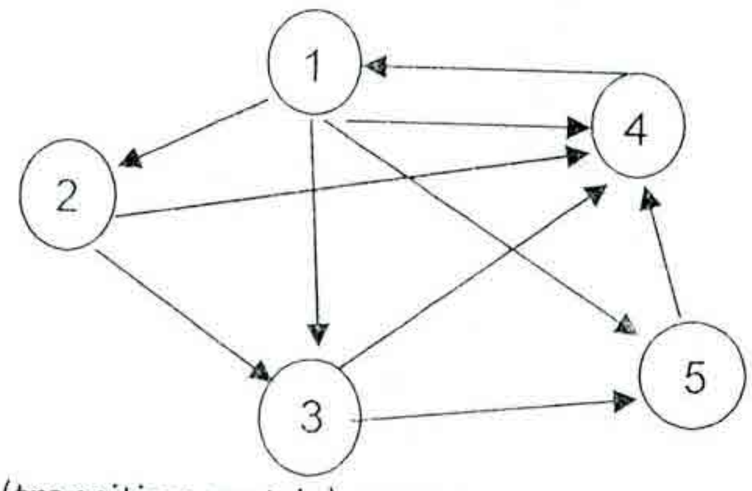
חלק ב 70%

1. 20% link analysis
   א. % 11% נתונים 3 גרפים שונים המתארים קישוריות של צמתים ברשת:
   4% גרף אחד: A->B->C->D
   5% גרף שני: כמו הגרף הראשון, רק שבנוסף D מצביע על A (כלומר נוצר מעגל)
   5% גרף שלישי :



   עבור כל אחד מהגרפים הסבר מה יהיו ערכי pagerank של כל הצמתים (אין לחשב ערכים מדוייקים, אלא להעריך ולהסביר את היחסים בין ערכי כל הצמתים)

ב. 6% נתון הגרף הבא:



הראה את מטריצת המעברים (transition matrix) שעל פיה יחושב Pagerank לדפים (ללא התחשבות בפרמטר d.)

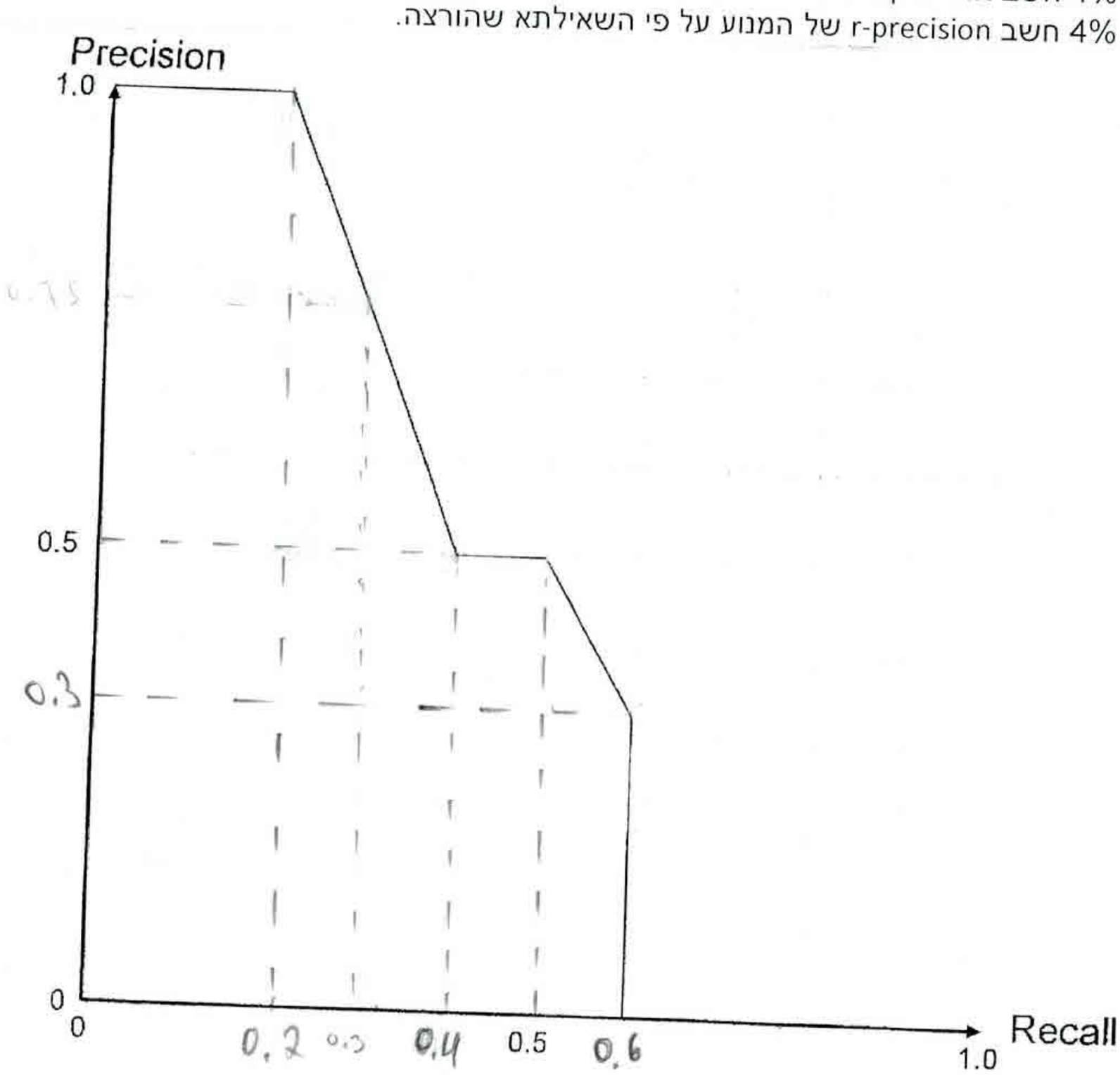
2. 20% נתון גרף Precision-recall על תוצאות של הרצת מנוע מסויים על שאילתא מסויימת. המנוע החזיר 20 תוצאות. ידוע שבמאגר 10 מסמכים רלוונטים לשאילתא.

א. 4% מהו ה precision לאחר שהמנוע החזיר 4 מסמכים רלוונטים.

ב. 8% הראה את הרשימה שהמנוע החזיר (הצג רשימה של 20 מסמכים מדורגים וסמן מי מהם רלוונטי ומי לא)

ג. 4% חשב את ה map של המנוע על פי השאילתא שהורצה.

ד. 4% חשב r-precision של המנוע על פי השאילתא שהורצה.



3. 15%

א. 10% הועלה רעיון להשתמש ב jaccard similarity לצורך דירוג מסמכים לשאילתא (כלומר, לדרג את המסמך על פי מספר המילים המשותפות בין שאילתא ומסמך מחולק בסכום של מספר המילים של השאילתא והמסמך). אפשר להניח שכל המילים בשאילתא יחודיות (אין חזרה על מילים) ושבמסמך סופרים כל מילה רק פעם אחת. באופן פורמאלי הדרוג למסמך D נקבע על פי:
jaccard(Q,D) = |Q ∩ D| / |Q ∪ D| כאשר Q הוא מספר המילים הייחודיות בשאילתא, ו D הוא מספר המילים הייחודיות במסמך.
תן 2 סיבות לכך שהרעיון אינו מוצלח, כלומר שתי סיבות לכך שהשיטה אינה מדרגת טוב מסמכים.

ב. 5% הדגם את חסרונה של השיטה מסעיף א. כלומר, תן דוגמא ספציפית לשאילתא Q כלשהי ושני מסמכים D1 ו D2 . כאשר ברור ש D1 רלוונטי לשאילתא ו D2 לא רלוונטי. בחישוב על פי jaccard יוצא ש D2 יותר רלוונטי מ D1.

4. 15% בניסיון לשפר תוצאות של מנוע שעובד לפי השיטה הווקטורית, הוצע להוסיף לשאילתא מילים נרדפות למילות השאילתא מתוך מילון. לכל מילה בשאילתא הוצע להוסיף עד 10 מילים נרדפות.
למשל: אם השאילתא היא : אחזור מידע – המילים הנרדפות לאחזור במילון הן: דלייה, אשוב, אשנה. מילים נרדפות למידע הן: ידיעה, אינפורמציה. השאילתא שתישלח למנוע תהיה:
אחזור, דלייה, אשוב, אשנה, מידע, ידיעה, אינפורמציה

א. 8% תן סיבה אחת לסיכוי שהשיטה תשפר תוצאות וסיבה אחת לסיכוי שהשיטה תפגע באיכות התוצאות

ב. 7% הצע רעיון לשיפור השיטה שתגדיל את סיכוייה לשפר את התוצאות, הסבר את ההיגיון ברעיון שאתה מציע.

בהצלחה

ברכה ואורלי

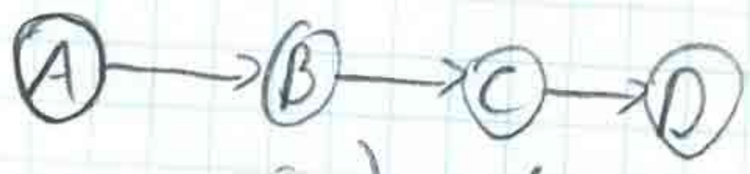
מס' נבחן

9/9

חלק ב

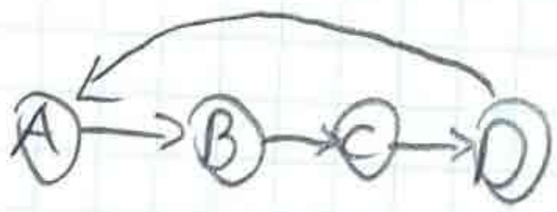
שאלה 1

א. (1)



ה- PR של A שווה ל- $1-d$ (ההסתברות שנגיע לצומת A כאומן ישר) דבר שבמאמר של דגר יאפס את כל ה- PR של הצמתים בגרף.

(2)



$PR(A) = PR(D)$

$PR(B) = PR(A)$

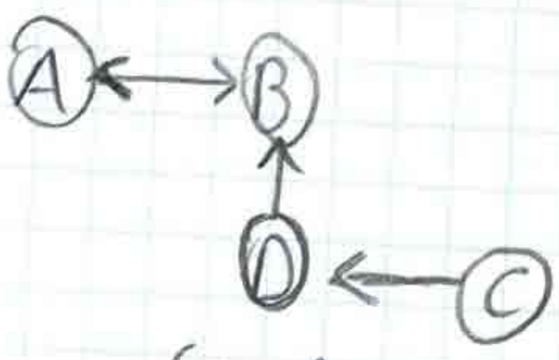
$PR(C) = PR(B)$

$PR(D) = PR(C)$

$\Rightarrow PR(A) = PR(B) = PR(C) = PR(C)$

נקבל מעגל כך שכל צומת מצביע ל צומת הבאה אחריה ובסוף צומת D מצביע חזרה ל A דבר אשר גורם לשיווין בין כל ה- PR של הצמתים.

(3)



ה- PR של C שווה ל- $1-d$ (במו בסעיף א) , ה- PR של D שווה ל PR של C דבר אשר יאפס במאמר של דגר אם ה PR של D, ובמאמר.

$PR(A) = PR(B)$

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

לפי מה שבנינו למעלה ה- PR של D במאמר א דבר יהיה שווה ל אפס ולכן $PR(A) = PR(B)$.

המשך שאלה 1

(א) מטריצת המעברים:

מצבים	1	2	3	4	5
1	0	0	0	1	0
2	$\frac{1}{4}$	0	0	0	0
3	$\frac{1}{4}$	$\frac{1}{2}$	0	0	0
4	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	0	1
5	$\frac{1}{4}$	0	$\frac{1}{2}$	0	0
	$\shortparallel$	$\shortparallel$	$\shortparallel$	$\shortparallel$	$\shortparallel$
	1	1	1	1	1

בדיקה

0
0.1
0.2
0.3
0.4
0.5
0.6

חלק ב'

שאלה 2

נתונים: המנוע החזיר 20 תוצאות

באשר יש 10 מסמכים רלוונטיים לשאילתה.

א. $Precision = \frac{4}{8} = 0.5$ אחרי 4 מסמכים רלוונטים

ב. √ - רלוונטי

X - לא רלוונטי

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

ג. $MAP = \frac{1+1+0.75+0.5+0.5+\frac{1}{3}}{6} = 0.6805$

ד. $R\text{-}precision = \frac{5}{10} = 0.5$

שאלה 4

א. <u>סיבה לשיפור</u> - במקרה זה כאשר לשאילתא מוסיפים מילים נרדפות לכל מילה בשאילתא המנוע יחזיר מסמכים אשר קשורים לשאילתא אך לא בוקשו בצורה מפורשת ע"י המשתמש, הרי כאשר משתמש שולח שאילתא לדוגמא: אחזור מידע. אין זה משנה אם הוא יקבל מסמכים אשר כוללים את המילים אחזור אינפורמציה, מפני שלאינפורמציה ומידע ישנה אותה משמעות, דבר זה מגדיל את כמות כמות המסמכים הרלוונטים אשר המנוע יכול להחזיר. ✓

<u>סבה שהפעולה תפגע באיכות</u>- אם המשתמש שולח שאילתא מסוימת והוא רוצה לקבל מסמכים אשר המילים הספציפיות בשאילתא מופיעות במסמכים, הוספת המילים הנרדפות תגרום לשינוי במספר המסמכים וכך המסמכים הרלוונטים לשאילתא הספציפית יכול להיות שלא יחזרו בכלל ע"י המנוע או שיחזרו אחרי הרבה מסמכים שלא רלוונטים דבר אשר יפגע באיכות התוצאות. ✓

ב. כאשר המשתמש שולח שאילתא אפשר "לבקש" מהמשתמש לבחור האם הוא רוצה את המילים הספציפיות הללו או האם לחפש גם מילים נרדפות, כאשר המשתמש יבחר את התשובה הרצויה לו, לפני שהשאילתה תשלח למנוע נדע האם לחפש מילים נרדפות במילון ולהוסיף לשאילתא או לשלוח את השאילתא כמו שהמשתמש שלח בהתחלה. ✓
הדבר יכול להתגבר על הבעיות שכתבתי בסעיף א, ולשפר את התוצאות.

a. Dog

b. D_1 Dog Dog Dog Cat Cat

D_1 Dog Cat

$D_1\ q = \frac{1}{2}$

$D_2\ q = \frac{1}{2}$

שאלה 3

א. **סיבה 1-** באשר ישנה שאילתא בעלת מילה משותפת אשר נמצאת במסמך אחד מס' רב של פעמים ובמסמך שני מס' מועט של פעמים אזי מסמך אחד יתר רלוונטי לשאילתא ובמסמך השני אך יכול להיות מצב שבו המסמך השני מקבל את אותו הציון כמו המסמך הראשון גם קיום בבירה יתר

סיבה 2- אורך המסמך מאוד משפיע על קיום המסמך בשיטת jaccard, כאשר מסמך הוא ארוך ובעל הרבה מילים שונות ה- QUD לא יהיה קבוע וכך יקטן קיום המסמך, גם אם מילה שם מופיע בשאילתא חוברת עם עצמה מס' רב של פעמים עדיין ה DUQ נשאר ב קבוע ומקטין את קיום המסמך.

ב. שאלה - q_1 : Dog

מסמך כשהם D_1 : Dog Dog Dog Dog cat dog boy dog

ומסמך שני D_2 : Dog Cat Cat Cat Cat

$$jaccard(Q, D_1) = \frac{1}{3} = \frac{1}{3}$$

$$jaccard(Q, D_2) = \frac{1}{2} = \frac{1}{2}$$

ולכאורה שמסמך D_1 יתר רלוונטי לשאילתא מ- D_2 אך D_2 מקבל משקל יתר גבוה מ- D_1.