אוניברסיטת בן-גוריון בנגב

הוראות לנבחן בצידו השני של הדף

אין לכתוב מעבר לשוליים משני צידי הדף

מס' כיתה 307 בנין 32

מס' נבחן 020255

20/02/2012 1345303 1 53/10

הנדסת מערכות מידע

אחזור מידע וספריות דיגיטליות

03721440601101

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 ברכה שבשי

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.

53/10

בס"ד

איחזור מידע תשע"א – 372.1.4406

סמסטר חורף מועד א' 31.01.11 - גירסא ב'

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

חלק א ' 30% - יש לענות על גבי הטופס

סמן תשובה אחת נכונה

1. 5% נתון מאגר עם 4 המסמכים הבאים:

D_1: alpha bravo charlie delta echo foxtrot golf
D_2: golf golf golf delta alpha
D_3: bravo charlie bravo echo foxtrot bravo
D_4: foxtrot alpha alpha golf golf delta

לשאילתא alpha bravo

א. מנוע שמבוסס על vector space יחזיר את כל המסמכים ואת מסמך D1 במקום הראשון
ב. מנוע שמתבסס על vector space יחזיר את כל המסמכים ואת מסמך D3 במקום הראשון
ג. ~~מנוע שמתבסס על מודל בוליאני טהור יחזיר את מסמך d1~~
ד. ~~מנוע שמתבסס על מנוע בוליאני טהור לא יחזיר את מסמך d2~~
ה. ~~ג+ד נכונים~~
ו. ~~ב+ג נכונים~~

2. 4% הפרמטר d בנוסחת Pagerank

א. מאפשר למדל באופן נכון את התנהגות ה Random surfer (הגולש האקראי) ולפתור את בעיית ה spider trap
ב. ~~מאפשר לחשב pagerank באופן מהיר~~
ג. ~~מאפשר לנרמל את תוצאת ה Pagerank~~
ד. ~~הוא הווקטור העצמי של מטריצת המעברים~~
ה. מאפשר למשתמש לעבור לדף באופן אקראי — באופן סמאוי לא אקראי
ו. א+ה נכונים

3. 4% מדדי ההערכה הבאים מתאימים להערכת מנועי חיפוש אינטרנטיים:

א. ~~DCG , F-measure,precision@k ,R-precision~~
ב. MAP ,interpolated average precision ,precision@k ,DCG
ג. Reciprocal Rank ,precision@k ,MAP ,DCG,
ד. ~~MAP ,precision@10 ,Fallout ,NDCG~~
ה. ~~כל המדדים שהוזכרו בהרצאות יכולים להתאים גם למנועי חיפוש אינטרנטים~~
ו. ~~אף אחד מהמדדים שהוזכרו בהרצאות לא יכול להתאים בלי שינויים בהגדרות~~

4. 4% עדכון של האינדקס מיד עם הגעה של מסמך חדש מה crawler
   א. מאט את פעולת האינדוקס
   ב. משפר בהכרח את ה precision של שאילתות רלוונטיות
   ג. משפר בהכרח את ה recall של שאילתות רלוונטיות
   ד. מעלה את העדכניות של האינדקס אך מסבך את הניהול שלו
   ה. משפר את המהירות של ביצוע השאילתא.

5. 5% ההשפעה של חוק zipf באה לידי ביטוי באורך רשימות postings של terms על פי הפירוט הבא (רשימות posting הן הרשימות הכוללות אינפורמציה על מופעים של Terms במסמכים):
   א. אין קשר בין הדברים
   ב. רשימות של מעט מילים תהיינה מאוד ארוכות ושל הרבה מילים תהיינה מאוד קצרות
   ג. רשימות של הרבה מילים תהיינה מאוד ארוכות ושל מעט מילים תהיינה מאוד קצרות
   ד. ~~החוק יבוא לידי ביטוי רק אם יופעל תהליך stemming על המסמכים~~
   ה. ~~החוק יבוא לידי ביטוי רק אם יופעל תהליך של הסרת stopwords~~
   ו. ~~ב+ד נכונים~~
   ז. ב+ה נכונים
   ח. ~~ג+ה נכונים~~

סמן נכון או לא נכון:

6. 4% כדי לאפשר מענה לשאילתות של ביטויים האינדקס חייב להיות במבנה של biword index . נכון/לא נכון

7. 4% Relevance feedback יכול להתבצע באמצעות האלגוריתם של Rocchio אשר מעדכן את ווקטור השאילתא שהמשתמש הקליד על פי תגובת המשתמש לתוצאות שהמנוע החזיר לשאילתא בהתאם לערכי הפרמטרים שקובעים את רמת ההתחשבות בשאילתת המשתמש לעומת תגובתו לתוצאות השאילתא . נכון/לא נכון

חלק ב ' 70%

ענה על השאלות הבאות:

1. 30%

הנח שאילתא q1 שלה 5 מסמכים רלוונטים במאגר על פי הפירוט הבא,

מסמך d13 רלוונטי באופן מושלם (ציון 3 על סקאלה של 0-3 כאשר 0 הוא לא רלוונטי ו-3 רלוונטי באופן מושלם) , מסמכים d10 ו d9 רלוונטים מאוד (ציון 2) , מסמכים d1 ו d3 רלוונטים באופן סביר (ציון 1) וכל שאר המסמכים במאגר לא רלוונטים. (הערה: כל המסמכים שקיבלו ציון גבוה או שווה ל-1 נחשבים רלוונטים). מנוע E1 הפעיל את שאילתא q1 על המאגר הנ"ל והחזיר 20 מסמכים. ידועים רק 9 המסמכים הראשונים שהמנוע החזיר ואת הסדר שבו החזיר אותם (משמאל לימין):

d13, d2, d8,d15,d3,d10,d1,d9, d4.....

3/10

ענה על השאלות הבאות:

א. 2% מהו ה precision ב 10 מסמכים של המנוע, e1 על פי השאילתא הנ"ל

ב. 3% אם המנוע היה אידיאלי לשאילתא הנ"ל (כלומר היה מחזיר את כל המסמכים הרלוונטים ראשונים, מדורגים לפי רמת הרלוונטיות שלהם), מה היה אז ה precision ב10 מסמכים?

ג. 4% מהו ה r-precision, ומהו ה Reciprocal Rank של מנוע e1 על פי השאילתא הנ"ל

ד. 4% מהו ה precision ב 30% recall?

ה. 5% חשב DCG ב 5 , השתמש ב discount של $1/\log_2(rank)$

ו. 5% מנוע e1 הריץ שאילתא q2 שלה 15 מסמכים רלוונטים במאגר, המנוע החזיר 15 תוצאות על פי הסדר הבא (משמאל לימין, על כל מסמך שחזר מסומן אם הוא רלוונטי או לא, כאשר "ר" מסמן מסמך רלוונטי ו "ל" מסמן מסמך לא רלוונטי):

ל,ר,ל,ל,ר, ל, ל ,ל,ר, ר,ר,,ל,ר,ר,ר,

חשב את interpolated average precision על פי שתי השאילתות.

ז. 3% חשב MAP של המנוע על פי שאילתא q1.

ח. 6% איך ישתנה ה MAP אם מתווסף מסמך רלוונטי נוסף למאגר והמנוע מחזיר אותו בין ה 20 המסמכים הראשונים. (הראה את ה MAP המקסימלי והמינימאלי האפשריים בעקבות השינוי).

2. 5% במערכת סינון שמקבלת באופן קבוע זרימה של מסמכים חדשים, לכל מסמך שמגיע למערכת מחושב הדמיון שלו עם פרופיל המשתמש. נניח שכל מסמך מיוצג כווקטור על פי גישת tf*idf וכך גם פרופיל המשתמש . מהו idf במקרה זה, כלומר לאיזה מאגר אפשר לייחס את חישוב ה idf?

3. 15% לצורך תיקון שגיאות של שאילתות, מחשבים מרחק בין מילת שאילתא שמשתמש הקליד ואיננה נמצאת באינקדס ומילים פוטנציאליות שאליהן התכוון המשתמש. הצע **נוסחא** לחישוב מרחק בין מילה שגויה לבין מילה פוטנציאלית מתוקנת שתתחשב במרחק בין אותיות במקלדת (בין אותיות שאולי הוחלפו בין המילים לתיקון) וכן בהפרש בין אורך המילה המקורית לאורך המילה המוצעת כתיקון. הראה באופן מפורט את כל חלקי הנוסחא הנדרשים , למשל הראה כיצד מחושב מרחק בין האותיות במקלדת). (כתשובה יש להראות נוסחה ולא הסברים מילוליים)

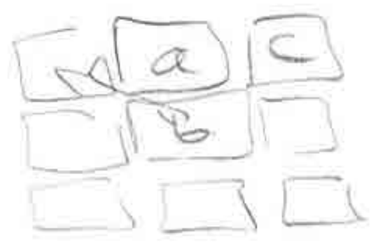
4. 10% קיימות שיטות שונות לקודד מילים על פי איך שהן נשמעות. השיטה הידועה ביותר היא שיטת ה Soundex – שהיא שיטה המבוססת על פונטיקה שהומצאה בשנת 1918.
כל מילה מקודדת לאות שאחריה 3 ספרות. האות היא האות הראשונה של המילה, והספרות נקבעות על ידי כללים שונים שמקבצים כמה אותיות דומות לאותה ספרה. כך למשל, M ו N מקודדות לסיפרה 5. B, F, P, ו V מקודדות לסיפרה 1. אותיות ניקוד אינן מיוצגות, אלא אם כן האות הראשונה של המילה היא אות ניקוד. להלן כמה דוגמאות למילים מקודדות בשיטת soundex:

הקוד S-530 הוא הקוד של המילים: smith ו-smythe
הקוד a-450 הוא הקוד של המילים : Allan, Allen, Alan, Allynn ועוד.....
Retrieve ו retrieval מקודדות ל r-361 (משום שמתעלמים מאותיות עודפות מעבר לשלוש ספרות קוד).
ציין שני יתרונות ושני חסרונות (שונים) לשימוש ב Soundex כשיטה לייצוג המילים באינדקס-במקום להשתמש ב Terms עצמם (או ב stem שלהם).

5. 10% מנועי חיפוש משתמשים בתוצאות של שאילתות זהות קודמות כדי לשפר תוצאות של שאילתא נוכחית. כלומר, המנועים שומרים במאגר מיוחד את השאילתות של המשתמשים, את התוצאות שהם החזירו וגם את התנהגות המשתמש עם התוצאות, כלומר אילו מהתשובות הוא אהב. כאשר משתמש מקליד שאילתא, המנוע מחפש במאגר המיוחד שאילתא זהה (אחת או יותר) ומשתמש בתוצאות של השאילתות הזהות כדי לשפר את תוצאות השאילתא הנוכחית. הבעיה היא שכ 50% מהשאילתות שנשלחות למנועי חיפוש הן ייחודיות (כלומר, אין להן שאילתות זהות). הסבר כיצד אפשר להשתמש בתוצאות של שאילתות קודמות, גם כאשר השאילתא שהמשתמש הקליד אינה זהה לשאילתא שקיימת במאגר המיוחד. (אין צורך להציג אלגוריתם מדוייק בפסדו-קוד אלא לספק הסבר ברור על השיטה שאתם מציעים)

בהצלחה - ברכה ואורלי

a c



53/10

שאלה 1

$\frac{22}{32}$

נתונים - שאילתא q_1 ו 5 רלוונטיים

d_{13}	רלוונטי באופן מושלם	3
d_{10}, d_9	רלוונטיים מאוד	2
d_1, d_3	רלוונטיים באופן סביר	1
השאר	לא רלוונטיים	0

$\Rightarrow$

1
$\frac{2}{3}$
$\frac{1}{3}$
0

$\Leftarrow_1$ התיבור 20 מסמכים.

$d_{13}, d_2, d_7, d_5, d_3, d_{10}, d_1, d_9, d_4 \ldots\ldots$ פלט

3 1 0 0 0 1 2 1 2 0

לפי הפורמט:

$$\frac{1+\frac{1}{3}+\frac{2}{3}+\frac{1}{3}+\frac{2}{3}}{10} =$$

לפי רמת מסמכים רלוונטים שהתקבלו

(א) precision at 10 - $\frac{5}{10} = 0.5$

מספר המסמכים הרלוונטיים

סך המסמכים

(ב) אני מניחה שבכלל שיש 5 רלוונטים לשאילתא הוא יחזיר בהתחלה 5 מסמכים שיהיו רלוונטים באופן מושלם. בהנחה שאין עוד רלוונטים

precision at 10 = $\frac{\overbrace{1+1+1+1+1}^{\text{רלוונטים באופן מושלם}}}{10} = 0.5$

לפי רמת מסמכים

לפי פורמט

(ג) r-precision = $\frac{1+0+0+0+1}{5} = \frac{2}{5} \rightarrow \frac{1+0+0+0+}{5}$

Reciprocal Rank = $\frac{1}{1} = 1$

פורמט של רמת רלוונטיות בינארית

(3) $recall_{0.3} \rightarrow \frac{3}{10} \rightarrow \frac{1.5}{5}$ ✗ $\boxed{0.55}$ -4

(ד) DCG_5

Rank	1	2	3	4	5
reli	3	0	0	0	1
disc	3	0	0	0	$\frac{1}{\frac{1}{\log_2(1)}}$

-3

התאמה לפי הקומבינציה הוא אותו דבר

$\Rightarrow \frac{\log_2(1)}{1} = 0$

$\frac{1}{\log_2 5}$ לא לאנרמל גם נדרש

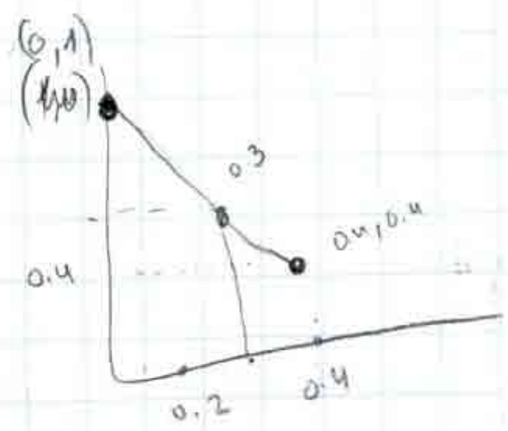
אני מניחה שהרווח שהמסמך תרם כך:

$$\frac{rel}{\frac{1}{\log_2(rank)}}$$

לא הבנה

$DCG_5 = 3.43$

$\frac{0.6}{0.4}$



$-\frac{0.4}{0.6}$ -

$y = -\frac{2}{3}(x-0)$

$\boxed{y-1 = \frac{3}{2}x}$

$y = \frac{3}{2}x + 1$

$\frac{6}{7}$ $\frac{7}{11}$

(4)

e1

מספר מסמכים

2a

מסמכים רלוונטיים 15

		recall	precision	Stanta	interput.
1	✓	$\frac{1}{15}=0.06$	1	0	1
2	✓	$\frac{2}{15}=0.13$	1	0.1	1
3	✓	$\frac{3}{15}=0.2$	1	0.2	1
4	X			0.3	$\cancel{\frac{5}{6}}$ $\frac{6}{7}$
5	✓	$\frac{4}{15}=0.26$	$\frac{4}{5}$		
6	✓	$\frac{5}{15}=0.33$	$\frac{5}{6}$	0.4	$\frac{6}{7}$
7	✓	$\frac{6}{15}=0.4$	$\frac{6}{7}$		
8	X			0.5	$\frac{8}{14}$
9	X				
10	X			0.6	0
11	✓	$\frac{7}{15}=0.466$	$\frac{7}{11}$	0.7	0
12	X			0.8	0
13	X				
14	✓	$\frac{8}{15}=0.53$	$\frac{8}{14}$	0.9	0
15	X			1	0

e2

		recall	pre	stat.	interpolated
1	✓	$\frac{1}{5}=0.2$	1	0	1
2	X		$\frac{1}{2}$	0.1	1
3	X			0.2	1
4	X			0.3	$\frac{2}{5}$
5	✓	$\frac{2}{5}=0.4$	$\frac{2}{5}$	0.4	$\frac{2}{5}$
6	✓	$\frac{3}{5}=0.6$	$\frac{3}{6}$	0.5	$\frac{3}{6}$
7	✓	$\frac{4}{5}=0.8$	$\frac{4}{7}$	0.6	$\frac{3}{6}$ $\frac{4}{7}$
8	✓	$\frac{5}{5}=1$	$\cancel{\frac{5}{8}}$	0.7	$\frac{4}{7}$
9	X			0.8	$\frac{4}{7}$ $\cancel{\frac{5}{8}}$ $\cancel{\frac{5}{8}}$
	X			0.9	
	⋮			1	
20					

$\downarrow$ 1, 1, 1, $\frac{37}{60}$, $\frac{22}{35}$, $\frac{15}{28}$, $\frac{1}{4}$, $\frac{2}{7}$, $\frac{255}{7,616}$ ← 3 INN

המשך

$$\frac{1+1+1+\frac{37}{60}+\frac{22}{35}+\frac{15}{28}+\frac{1}{4}+\frac{2}{7}+\frac{2}{7}+\frac{5}{16}+\frac{5}{16}}{11} = \boxed{0.566}$$

יפה

ב) $$\frac{1+\frac{2}{5}+\frac{3}{6}+\frac{4}{7}+\frac{5}{8}}{5} = \boxed{0.619}$$

map =

ג) נוסף הלוונטי 1

במערכה הכי טוב הלוונטי אתווסף ממש בהתחלה אחרי ה 1 שהנתקלים.

		recall	pr
	1 ✓	$\frac{1}{6}$	1
	2 ✓	$\frac{2}{6}$	1
	3		
	4		
	5 ✓	$\frac{3}{6}$	$\frac{3}{5}$
	6 ✓	$\frac{4}{6}$	$\frac{4}{6}$
	7 ✓	$\frac{5}{6}$	$\frac{5}{7}$
	8 ✓	$\frac{6}{6}$	$\frac{6}{8}$

שינוי

לא רשום איפה בכל

לגל התשובה הכי נכון לא נכון כי כל המסמכים האלה תגרים לקרוסלא נמוך יותר

$$map_{Max} = \frac{\frac{3}{5}+\frac{4}{6}+\frac{5}{7}+\frac{6}{8}+1+1}{6}$$

$$= 0.788$$ ✗

-3

במערכה הכי גרוע ותלה במסמך הכי

	prec
1 ✓	1
2 ✗	
3 ✗	
4 ✗	
5 ✓	$\frac{2}{5}$
6 ✓	$\frac{3}{6}$
7 ✓	$\frac{4}{7}$
8 ✓	$\frac{5}{8}$
⋮	
20 ✓	$\frac{6}{20}$

$$map_{Min} = \frac{1+\frac{2}{5}+\frac{3}{6}+\frac{4}{7}+\frac{5}{8}+\frac{6}{20}}{6}$$

$$= \boxed{0.566}$$

7.5/10

שאלה 4

בתרונות - א) מקטין את גודל האינדקס - הרבה הודעות מילים מקובצות למילת קוד אחת - וכך ניתן לחסוך במקום

ב) יכול להקטין את זמן החיפוש כתוצאה מהקטנת גודל האינדקס. אם משתמש הכניס מילה וחפש את מילת הקוד שלה.

חסרונות - א) גם כאמר האם באינדקס סמור כל מילת קוד נשמרת הקבוצת המילים שלה. אבן יכול להיות מצב שמשתמש יכניס מילה ששייכת כבר למילת קוד מסוימת ומה יביא לו את כל המסמכים שקשורים לכל מילות הקוד

לכן כדי סמור כל מילה בקבוצת מילת הקוד לשמור את המסמכים הרלוונטים שלה. לכן בקטע הזה, השיטה לא חוסכת דבר. עדיין צריך לתרגם סמור כל מילה בודדת

ב) אולי לבנות מדדים המודגל - המילים המקובצות בן מילים ללא קשר לשורר לשאר המילים בקבוצה

tf · Idf < פרופיל משתמש < רעיון

tf · Idf < מסמך חדש < רעיון

לה ל הקטיים נשמר רק קוד ומכן לכ אפשר סעיר אקיים מתעדר ומל מבודה נוסכן בנו שרקבן בסעיף 1

שאלה ב

חישוב ה אלו יהיה לפי המאגר של פרופיל המשתמש כלומר אנו מנתח שוש לפרופיל המשתמש רשימת מסמכים שאהב/ ואשר מבטא מסמך חדש ונחשב אלו מול רשימת המסמכים של המשתמש

במובן וקיים רק פרופיל משתמש ולא רשימת המסמכים שלו חישוב ה אלו יהיה מול מאגר המסמכים הכולל

אין דבר כזה מאגר מסמכים כללי!

כולל מה? המסמכים שהגישו לכל המשתמשים?

ממתוך

-7

שאלה 3

M = מולה הקבה ותה

L = המולה האורכה ביותר

אין שום רמז להניח לה לגיל מקום.

הבן לה ביותר חלק.....

לכל אות; X נשמור את המולה שמחזיקה אליה ושמרת והמיקום מתחיל

1. בשרשרת השנייה יהיה מרחק של $\frac{1}{2}$, אחר $\frac{1}{3}$ .....

במודר והמטרה להשיג מרחק, נתחל מהשרשרת ה1 ונבדוק אם יש האמה לשה.

אם לא, נעבור לשרשרת ה2 ..... ונחזור להאמה הכי גדולה.

נחלק את המילה ל2 משרשרת

למה -1

$$\left(\max\left(\sum_{j=1}^{L}\sum_{i=1}^{M}(y_i \cdot x_j)\right)\right)\cdot\left(1-\frac{|L-M|}{L}\right)\cdot\left|\frac{\omega\,n\,z}{w\,\theta\,y}\right|$$

למה!

y_i: המרחק של האות שנרצה לחתיך למרחק במילה

x_j: האות הספציפית שיכול להיות שהוחלפה.

$x_j = \Delta$

אין לו זמן לספר את החלק הלבן. הכוונה היא שנחשב לאות סיבוב המילה לאות סיבוב המרחק שלה שאות שיכול לחיות לחלפה ואז נחשב המילה כמידה ולא נעבוד מומן ומרחק אלא הכי אות אולם

מתחשב באורך המילה

מכל שמות מילים באותו אורך והיה הפרש קטן יותר ולכן לא הכיוון בגלל יש

הנוסחא:

$\omega\,n\,z$

$w\,\theta\,y$

בסכול מתחשב אותם

שמות משתנים לשרשרת

אותיות של השמות

לא הכיוון מאשר אם דרך להתחשב רק עם אחד הוספה אחד

כמה מקום מכין המילים

(קטר

מילים

שקיבל שמקום את

הכיוון לשון אלף אלף מילים

מדוקה מדוקה מדוקה

מדוקה מדוקה מדוקה

מדוקה מדוקה מדוקה

שאלה 5

ב-

שמעון:

נאצ' של שאילתות משתמשים, תשובות והתנהגות

אחד

יש שאילתא

צריך לעשות נתון! שיתרום את השאילתא-אל לא לפי

נתון בשאלה שמור כל משתמש שמורות השאילתות ששאל, תשובות והתנהגות.

בגלל שהשאילתות הן ייחודיות לא נחפש דומין בין שאילתא אחת לאחרת. אלא בין שאילתא מסוימת למסמכים ששמרים מידע כל אחת מהשאילתות, תשובות שאחרי המשתמש.
דמיון בין שאילתא למסמכים ניתן לבצע בדרך וחסור פשוט הי בשימוש ב COSSIM. כושר מציאה שאילתא של משתמש נבדוק מול המסמכים שאחרי למי יש את הדימון הכי גבוה לשאילתא. נקבע ערך סף לדוגמא של 0.8.

משתמש שאילתא 1 — אוסף מסמכים ותשובות, התנהגות.

שאילתא 2

מעלה בשאלה ...

נבדוק מול כל מסמך בשאילתא COSSIM ונעשה מטריצה, נבדל שקלול של COSSIM מעבר ל שאילתא. אם הערך גבוה מ0.8 אז נראה שהשאילתא הנוכחית לשאילתא אותה בדקנו. וכך נוכל להתאים תשובות כמו שהמשתמש אחרי / רבים.

אם קטן מ0.8 אז נראה שמדובר בשאילתא חדשה לגמרי שלא קשורה לגמרי לאחד השאילתות הקודמות.
ניתן לבדוק במקרה זה מול משתמשים שדומים למשתמש שלנו. ולבדוק מולם COSSIM לראות מה הם אהבו ואיך להתאים למשתמש את מה שהם אהבו. ואז נשקול פידבק מהמשתמש נוסף כדי את השאילתא האישית שלו

* בנוסף ניתן לשמור מידע על שאילתא ותוצאותיה את ה concept שלה ע"י שימוש במודל LSI
דרך שמביאה שאילתא חדשה לבדוק אם היא קשורה לנושא הספציפי ששאל בעבר המשתמש
אם אין בכל מאגר נתונים לגבי המשתמש
נבדוק אצל משתמשים שדומים לו.

מה נכון, עדיף לא למשתמש אלא בכלל לשאילתא!

מוצר

1 1

	alpha	brov		Idf $\frac{4}{2}$
d_1 $\frac{1}{7}$	1	1		log
D_2 $\frac{1}{5}$	$\frac{1}{3}$	0 $\frac{3}{5}$		
D_3 $\frac{0}{6}$	0	1 $\frac{3}{6}$		
D_4 $\frac{2}{6}$	1	0 0		

$\log\frac{4}{3}$ $\log\frac{4}{2}$

$= 0.34$

lie

$$\begin{cases} 1\log\frac{4}{3} + \log\frac{4}{2} \\ \frac{1}{7}\cdot\log\frac{4}{3} + \frac{1}{7}\log\frac{4}{3} \end{cases}$$

a	b	c
d	e	f
g	h	i