אוניברסיטת בן-גוריון בנגב

הוראות לנבחן בצידו השני של הדף

אין לכתוב מעבר לשוליים משני צידי הדף

מס' כיתה 306 בנין 32

מס' נבחן 020313

20/02/2012 1345302 1 12/12

הנדסת מערכות מידע

אחזור מידע וספריות דיגיטליות

03721440601101

המחלקה מע' מידע שנה 2

תאריך בחינה 20/2/12

מקצוע בחינה אחזור מידע

לשימוש המרצה הבודק

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

ציון הבחינה 76

שם המרצה

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.

12/12

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
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.....

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

12/12

שאלה 2

(1)

28/32

אלגוריתם 5 – q1

Precision	recall	rel?	doc#
1	0.2	✓	1
0.5	"	×	2
0.33	"	×	3
0.25	"	×	4
0.4	0.4	✓	5
0.5	0.6	✓	6
0.57	0.8	✓	7
0.625	1	✓	8
0.55	"	×	9
0.5	⋮	×	10
	⋮	×	⋮

א 0.5 ✓

ב 0.5 ✓

ג $r_{precision} = 0.4$ ✓

$MRR = \frac{1}{1} = 1$ ✓

ד אין עניינה בה ה- recall הוא 30% לכן לא ניתן לדעת.

ה עשרון אקראי בוטל (-4)

$DCG_5 = 3 + 0 + 0 + 0 + \frac{1}{\log_2 5} = 3.43$ ✓

ו אמור הכל ✓

מלבנית 15 - q2

1 (1)

Precision	recall	rel?	doc #
1	0.06	✓	1
1	0.13	✓	2
1	0.2	✓	3
	"	✗	4
0.8	0.26	✓	5
0.83	0.33	✓	6
0.85	0.4	✓	7
	"	✗	8
	"	✗	9
	"	✗	10
0.63	0.46	✓	11
	"	✗	12
	"	✗	13
0.57	0.53	✓	14
	"	✗	15

~~cutoff~~ Avg precision	q2 Precision	q7 precision	interpolated
1	1	1	0
1	1	1	0.1
1	1	1	0.2
0.625	0.85	0.4	0.3
0.625	0.85	0.4	0.4
0.53	0.57	0.5	0.5
0.25	0	0.5	0.6
0.28	0	0.57	0.7
0.28	0	0.57	0.8
0.31	0	0.625	0.9
0.31	0	0.625	1

interpolated average precision $= \dfrac{1 + 1 + 1 + 0.625 \cdot 2 + 0.53 + 0.25 + 0.28 \cdot 2 + 0.31 \cdot 2}{11} =$

$= 0.56$

לבדוק!

5

$$MAP(q_1) = \frac{1+0.4+0.5+0.57+0.625}{5} = 0.619$$

ח

אם המסמך ירוץ למקום האחרון (או הקטן), נקבל את ה- MAP המקסימלי:

$$\begin{matrix} Max \\ Map \end{matrix} = \frac{1+1+\frac{3}{6}+\frac{4}{7}+\frac{5}{8}+\frac{6}{9}}{6} = 0.72$$

מעולה!

אם המסמך ירוץ במקום ה- 20, נקבל את ה- MAP המינימלי:

$$\begin{matrix} Min \\ Map \end{matrix} = \frac{1+0.4+0.5+0.57+0.625+\frac{6}{20}}{6} = 0.56$$

(2)

	מסמכים	מסמכים
d_1	(, , ,)	(, , ,) user1
d_2	(, , ,)	(, , ,) user2
		(, , ,) user3

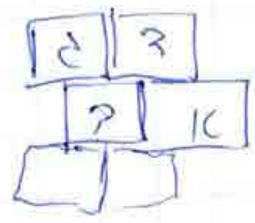
idf-tf פה כרונל מתאמים:

ניתן לייחס את חיסרון ה idf למאגר המתאמים כלומר אם יש מילה שמופיעה אצל כל המתאמים, אז המקדם שלה יהיה 0. למה? מה ההבדל? (-2)

(3)

נגדיר $dist(x,y)$ - מרחק מנהטן במרחב בין האות x לאות y. כלומר כמה כפתורים האצבע צריכה לעבור כדי להגיע מכפתור x לכפתור y. המרחק בין אותה האות הוא באופן 0.

לדוגמא:



$dist(א, ב) = 2$

$dist(א, ד) = 1$

אין מחסרים את ערך המנהטן אין מוציאים את המקלדת יד שאבשר לחשוב בקלות.

m - כמות הכפתורים בקלידת.

$$dist(w_1, w_2) = \alpha \cdot \frac{\sum_{i=1}^{\min(|w_1|,|w_2|)} \overbrace{m - dist(w_1[i], w_2[i])}^{\text{המרחקים בין כל אותיות מקלדת}}}{m \cdot \min(|w_1|, |w_2|)} + (1-\alpha) \cdot \overbrace{\frac{\min(|w_1|, |w_2|)}{\max(|w_1|, |w_2|)}}^{\text{התחלות בהבדל האורכים}}$$

(-2)

$w[i]$ - האות במקום ה-i במילה w.

$|w|$ - אורך המילה w.

α - פרמטר. נקבע את הערך קבוע ל- 0.9.

הסבר בסמוך הבא

הסבר לנוסחה:

הקנסת α - ניתן גבול (0.9) לחיזוק המרחק במקלדת בין האותיות כי זהו נתון חשוב בהרבה שההפרש בין אורך המילים (יש המון מילים באותו אורך השונות זו מזו...)

הנוסחה עוברת על כל אות במילה הקצרה יותר ובודקת את המרחק לאות במילה המקבילה. ככל שהמרחק קטן יותר כך טוב יותר לכן נחסיר מ- m מרחק זה.

מנרמלים ע"י חלוקה ב- m כפול כמות האותיות שעברנו עליהם.

⑨

חסרון 1 - מתמקדים רק באות הראשונה ומקודדים את הקוד לאותה המילה.

-2.5

למשל למילים Khris ו- chris יהיה קוד שונה למרות שבמילה אחת היתה טעות והכוונה היתה לאותה המילה.

דווקא בדיוק הפוך. זה יכול לתקן שגיאה ולהציג תוצאה למרות שיש שגיאת כתיב

חסרון 2 - יש מספר מוגבל של ספרות (3) לכל אות, כלומר לכל אות יש 999 אפשרויות שונות. יתכן וקיימות שפות מלבד אנגלית שכמות זו לא מספיקה.

-2.5

אם יש יותר אותיות
הייצוג יהיה יותר טוב
אבל מה שרלבנטי לאנגלית.

יתרון 1 -

יש מקרים בהם ה- stem יהפוך 2 מילים שונות לגמרי במשמעות לאותה המילה בגלל חיתוך לפי כללי דקדוק ברורים. דבר זה יתכן בעיקר בשמות של אנשים ומקומות. שיטת ה- Soundex מונעת זה. לפי אין פעולת [illegible] ברוב המקרים

-2

חסרה כאן דוגמא. זה דוקא מאוד דומה ל- stem אך
בחוקיות שונה ובכלל אם הכוונה היתה להראות מתי
זה עדיף מ- stem הלוא דוקא בדוגמא.

יתרון 2 - לאחר שקודדנו מילה, החיפוש שלה לפי הקידוד יהיה מהיר ~~כי~~ כי ניתן למיין את המילים ולחפש רק במקום המתאים כלומר איפה שהאות הראשונה זהה.

-2.5

נכון? גם Inverted index
מאפשר חיפוש מהיר, וגם אנדקס
היכן היתרון?
רגיל אפשר למיין.

$\frac{0.5}{10}$

* אני מניח לפי ניסוח השאלה ש-"זהה" = בדיוק אותה השאילתה.

נשמור מאגר של וקטורים לכל שאילתה (לאחר תהליך stems והורדת stop words).

האם להכלול את המסמכים הקשורים לשאילתא או רק את המילים בלבד

אם לא מצאנו שאילתה זהה לשאילתה שהכניס המשתמש, נוסיף את השאילתה שלו כוקטור גם כן.

משקל "מילה" בוקטור השאילתה יהיה 1 אם המילה בשאילתה.
של שאילתות במאגר יחושב לפי tf-idf
כש-tf "ינורמל" לפי גודל השאילתה. ; idf יחושב
למאגר של השאילתות הקיימות.

נחשב דמיון ע"י cos sim בין שאילתת המשתמש לבין כל אחת מהשאילתות במאגר.
ניקח סף התחלתי של 0.5, כאשר רק שאילתות שרמת הדמיון שלהן לשאילתה גבוהה מסף זה, נשתמש בתוצאות שלהן.
השימוש בתוצאות ישקה לפי רמת הדמיון.

ניעזר בתוצאות של השאילתה (או השאילתות) שהכי קרובות לשאילתת המשתמש (בתנאי שעברו כמובן את סף הדמיון).
אם אף שאילתה לא עברה את סף הדמיון אז לא נעזר בתוצאות של שאילתות אחרות.

נבקש מהמשתמשים משוב כדוגמת "האם התוצאות עזרו?" במידה והמשתמשים מגיבים התקפה.
ונראה שהתוצאות לא מספיק טובות והמשתמשים אינם מרוצים, ננסה לשנות את סף הדמיון ולבדוק שוב נוסף וכך הלאה.

איך נעזרים בתוצאות? אין פה שיטה לכך.

-2

$$dist(word\,1, word\,2) = \sum_{i=1}^{\min(|W_1|,|W_2|)} dist(W_1\{i\}, W_2\{i\})$$

word

d a g
d o b

apple
adpple ?

$$dist(w_1, w_2) = \frac{\sum_{i=1}^{\min(\ )} m - dist}{m \cdot \min(|w_1|,|w_2|)} + (1-\alpha)\frac{\min(\quad)}{\max(\quad)}$$

$\alpha = 0.9$