אוניברסיטת בן-גוריון - המחלקה להנדסת מערכות מידע
קורס איחזור מידע וספריות דיגיטליות
סמסטר חורף תשס"ח- 21.04.08 – מועד א'

מרצה: ד"ר ברכה שפירא, אסיסטנטית : ליהי נעמני
פתרונות

.א1

נניח ששאילתא מקורית של משתמש היתה "cheap CD DVDs extremely cheap CDs". המשתמש כתגובה העריך שמבין המסמכים שהמנוע החזיר לו מסמך d1 הוא רלוונטי, ומסמך d2 אינו רלוונטי לשאילתא.

להלן המסמכים:

d1: CDs cheap software cheap CDs
d2: cheap thrills DVDs

השתמש בנוסחת Rocchio כדי לשפר את שאילתת המשתמש עם הערכים הבאים: $\alpha=1, \beta=0.75, \gamma=0.25$. כדי לחשב את הווקטורים של המסמכים והשאילתא יש להשתמש רק בתדירות המילים ללא נורמליזציה וללא idf. הצג את וקטור השאילתא לפני ואחרי שיפורה.

פתרון:

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

	cheap	CDs	DVDs	extremely	software	thrills
query	2	2	1	1	0	0
d1	2	2	0	0	1	0
d2	1	0	1	0	0	1
RESULT	3.25	3.5	0.75	1	0.75	0

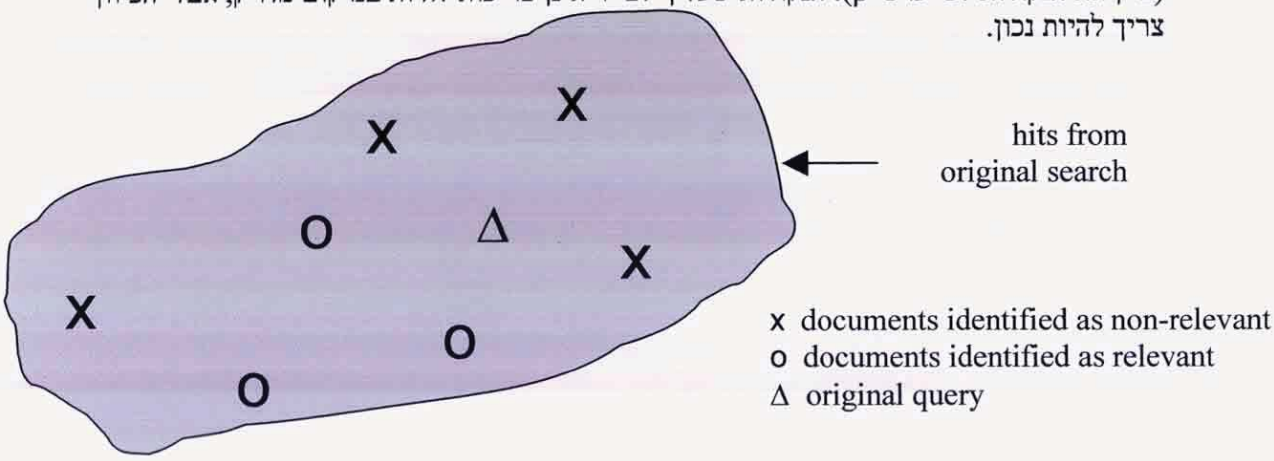
alpha	1
beta	0.75
gama	0.25

אין משקל שלילי עבור ה term האחרון.

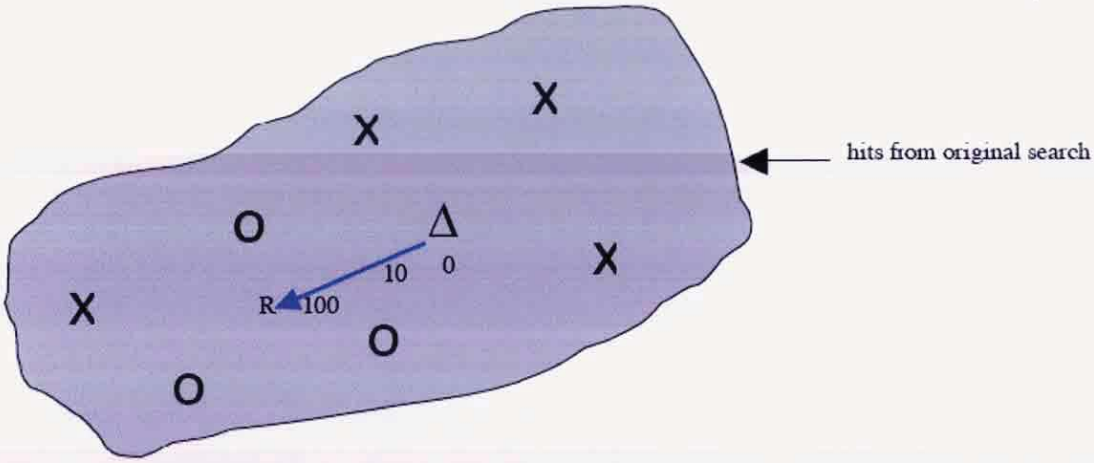
1ב. מה צריכים להיות ערכי α, β, γ כדי לממש במנוע חיפוש אופציה שמחזירה מסמכים דומים למסמך מסומן – כלומר המשתמש מסמן מסמך ומבקש מסמכים נוספים הדומים לו (more like this).

תשובה: מכיוון שהשאילתא לא חשובה וגם לא המסמכים הלא רלוונטים – הערכים הם beta=1 ואלפא וגמא=0.

1ג. נתון הציור הבא המתאר השטחה של המרחב הווקטורי של מסמכים . הציור מתאר את השאילתא המקורית של המשתמש, את המסמכים הרלוונטים לשאילתא ואת הלא רלוונטים.
הנח שהופעל אלגוריתם Rocchio על השאילתא 3 פעמים עם $\alpha=10$, $\gamma=0$ וערכי β של 0,10,100 (בכל הפעלה ערך אחר). הראה על הציור את השתנות המיקום של השאילתא בכל הפעלה של האלגוריתם (ציין את הנקודות לפי ערכי β). הנקודות שעליך לצייר אינן צריכות להיות במיקום מדויק, אבל הכיוון צריך להיות נכון.



פתרון:



X documents identified as nonelevant
O documents identified as relevant
Δ original query
R reformulated query

2. מדד הערכה אפשרי נוסף במקום **precision/recall** הוא מדד ה **accuracy** שמשתמשים בו כדי להעריך תוצאות סיווג. נניח מסווג שיודע לסווג משמכים כרלוונטים או לא רלוונטים לשאילתא (בעצם מערכת איחזור בוליאנית). אם המסווג סיווג **c** מסמכים באופן נכון, ו – **i** מסמכים לא נכון, **acuuracy** מוגדר כ : $c/(c+i)$

א. הסבר מדוע **recall/precision** מבטאים טוב יותר מאשר **accuracy** את התועלת שיש למשתמש משימוש במנוע.

פתרון: ברוב המאגרים ישנם מעט מסמכים רלוונטים. לכן מנוע שלא יחזיר אף תוצאה יהיה ברוב המקרים בעל **accuracy** גבוה. **Recall** ו **precision** עוזרים למדוד את ה **tradeoff** שקיים בין החזרת יותר מסמכים רלוונטים לבין החזרת פחות מסמכים לא רלוונטים שזה בעצם מה שמעניין משתמשים של מנוע חיפוש.

ב. נניח מאגר של 10 מסמכים ושני מנועים בוליאנים : **B,A** . תן דוגמא של תוצאות הרצה של המנועים על שאילתא **q**: **Aq**, **Bq**, כל ש **Precision(Aq)> Precision(Bq)** וגם התועלת של **Aq** גבוהה משל **Bq**. אבל, **Accuracy(Aq)=Accuracy(Bq)**

פתרון:
יש הרבה תשובות נכונות. אחת פשוטה היא: נניח כי מבין 10 המסמכים, מסמך מס' 1 הוא המסמך הרלוונטי היחיד במאגר
$Bq = \{3\}$ $Aq = \{1,2,3\}$
Aq טעה פעמיים, וכן Bq טעה פעמיים. לכן לשניהם אותו accuracy של 0.8
ה precision a של Aq הוא 1/3 ושל Bq 0.

ג. במאגר יש שני מסמכים **זהים** שהם המסמכים היחידים הרלוונטים לשאילתא מסויימת. מנוע מצא רק אחד מהם (ובעצם את כל האינפורמציה הרלוונטית). מהו ה **recall-** של המנוע לשאילתא זו? . נמק את תשובתך.

פתרון: ה **recall** הוא: 1/2 משום שהוא מוגדר על פי מספר המסמכים הרלוונטים.

3. נתונה המטריצה הבאה שמתארת שכיחות של terms במסמכים.
א. הצע שיטה המתבססת על מופע משותף במסמכים כדי להסיק ממטריצה זו מרחקים בין ה Terms לבין עצמם. תאר את השיטה בפסדו-קוד.

פתרון:
שיטה אחת פשוטה היא פשוט לספור בכמה מסמכים משותפים כל זוג של terms הופיעו (אולי בשכיחות מינימלית מסוימת) ולנרמל את הסכום למספר המסמכים במאגר.

ב. נניח שעל פי שיטה כלשהי הוסקה מטריצת המרחקים הבאה בין ה Terms לבין עצמם, הפעל אלגוריתם Clustering היררכי צובר על המטריצה בשיטת ה Complete-link, צייר את העצים שנוצרו בכל שלב.

פתרון:
שלב 1: t1 --- t2
שלב 2: t3---t5
שלב 3: t4---t6
שלב 4: חיבור של t1---t2 עם t4---t6
שלב חמש: כל העץ

4. הסיבה לביצוע אינטרפולציה של ערכי recall-precision ב -11 נקודות סטנדרטיות:

א. כדי שאפשר יהיה למצע ערכי precision-recall על פני שאילתות שונות שלהן נקודות recall שונות לסימון של נקודה אחת בגרף precision-recall

ב. כדי לאפשר השוואה בין precision ו-recall

ג. כדי שאפשר יהיה לחשב Mean Average Precision

ד. כדי שהגרפים של precision-recall יראו דומים.

פתרון:

התשובה הנכונה היא ב.

5. חוקר פיתח אלגוריתם חדשני לשיפור הדירוג של תוצאות מנועי חיפוש על ידי בדיקה של המקטעים במסמך שבו מופיעות מילות השאילתא במסמך. האלגוריתם סופר את מספר התווים במקטעים אלו ומניח שמקטעים חשובים מאופיינים על ידי מספר מצומצם של מילים. ככל שהמילה מופיעה במקטע חשוב יותר היא מקבלת משקל גבוה יותר. החוקר רוצה לבדוק האם ההתחשבות במקטעים משפרת את ה precision והאם היא פוגמת ביעילות (מבחינת מהירות).

לחוקר יש מאגר מסמכים עם שאילתות שעליהם הוא יכול להריץ ניסויים, אבל אין לו תוצאות ידועות לשאילתות (כלומר אין לו gold standard). הצע תוכנית ניסוי לחוקר כדי לבדוק את האלגוריתם שלו:

א. תאר תהליך ניסוי שעל פיו הוא יוכל להגיע למסקנה האם השיטה שלו משפרת precision (כולל תיאור המדדים)

פיתרון:

החוקר צריך ליצור gold-standard באמצעות הרצה של הרבה מנועים אמינים ולחתוך את התשובות הרלוונטיות שלהן. ואז להריץ את המנוע שלו עם ובלי השיטה החדשה ולהשוות את ה precision על פי ה Gold-standard שיצר.

שיטה אחרת היא לעבוד עם 2 קבוצות הומוגניות של משתמשים, אחת מריצה שאילתות עם מנוע עם השיטה החדשה ואחת עם מנוע ללא השיטה החדשה. המשתמשים מתבקשים לסמן האם x המסמכים הראשונים שחזרו הם רלוונטים או לא. .אפשר אז להשוות את ביצועי המנוע עם ובלי השיטה על פי precision ב – x מסמכים.

בהקשר ליעילות, פשוט מריצים עם ובלי השיטה החדשה (באותם תנאי ריצה) ומודדים זמני ביצוע ומחשבים בכמה גרוע או טוב יותר זמן הריצה של השיטה החדשה.

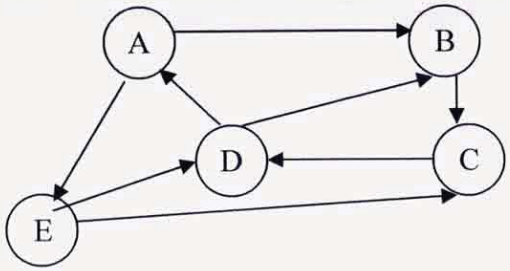
ב. האם ניתן לבצע את הניסוי ב Web ולא במאגר סגור. אם כן כיצד? ומה היתרונות והחסרונות של שימוש ב Web לעומת ניסוי המבוסס על מאגר סגור.

בשיטה של יצירת gold-standard אי אפשר לבצע ב web- משום שאי אפשר לחתוך מסמכים ממנועים שונים כי החיתוך לא משמעותי, כי מנועים שונים מאנדקסים מסמכים שונים והמאגרים שלהם שונים. זה ישים בשיטה של המשתמשים אם אפשר להוסיף את השיטה על מנוע שקיים ב-web כדי לא להצטרך לאנדקס (כי אז שוב חוזרים למאגר סגור).

היתרון של ה web הוא ההתמודדות עם הסביבה האמיתית שעליה המנוע אמור לרוץ והבדיקה היא אז בתנאים הכי אמיתיים שאפשר. החיסרון המרכזי הוא חוסר הבקרה והכנסת רעשים שונים שעלולים להטת את הבדיקות הסטטיסטיות.

6. Pagerank –

2. נתונה הרשת הבאה המתארת צמתים באינטרנט:



א. חשב את ערכי ה Pagerank של הצמתים (ללא נרמול) לאחר 2 איטרציות של חישוב (2 איטרציות מלבד השמת ערכים תחיליים, כל צומת = 1) אין צורך לחשב באמצעות המטריצה. D=0.85.

פתרון:

P(A) = 0.85*0.2+0.15 (אפשר לחלק את 0.15 ב-5)
P(B) = 0.85*0.2/2+0.15
P(C) = 0.85*(0.2+0.2/2)+0.15
P(D) = 0.85*(0.2+0.2/2)+0.15
P(E) = 0.85*0.2/2+0.15

ב. ציין מיהו הצומת בעל ערך ה authority הגבוה ביותר , והסבר על פי ההגיון של האלגוריתם את התוצאה (כלומר, להסביר לא על פי תוצאת החישוב)

1)H(a)=a(a)+a(b)
2) H(b) = a(c)
3) H(c)=a(d)
4) H(d)=a(a)
5) H(e)=a(d)+a(c)

H(a)>h(d) (1,4)
H(e)>h(d) (5)
H(e) > h(c) (3)
H(e) > h(b) (5)
H(a) > h(d) (1)

מזה אפשר לראות ש h(a) ו-h(e) בעלי ערך hub גבוה מהשאר ויש לבחון את שניהם:

H(a) = aut(b) + aut(e)
H(e) = aut(c) + aut(d)

אפשר לראות ש auth(c)+aut(d)> aut(e)+ aut(b) משום שאל c ו-d מצביעים שני צמתים (אל כל אחד מהם) ואילו אל B ו-e מצביע צומת אחד אל כל אחד מהם עם Hub בחוזק מתחרה. לכן:

H(e) הוא בעל הערך הגבוה ביותר.