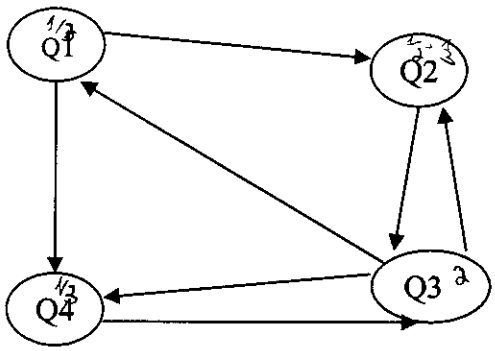
המחלקה להנדסת מערכות מידע והתוכנית להנדסת תוכנה
מבחן בקורס איחזור מידע 372.1.4406 מועד א'
ד"ר ברכה שפירא, ארז שלום
13:30 – 22.07.07

משך הבחינה: שעתיים וחצי
חומר עזר מותר – לא מחשב נייד!

יש לענות על כל השאלות

1. 20%

I. 10% נתון הגרף הבא המתאר רשת של דפי אינטרנט- יש לדרג את הצמתים על פי ערכי ה pagerank שלהם. אין צורך לבצע חישוב של הpagerank של כל צומת, אלא לנמק את הדרוג במדוייק על פי ההגיון של הנוסחה.



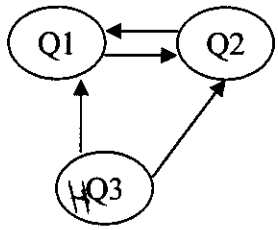
$Q_3 \to Q_2 \to Q_4 \to Q_1$

II. 4% הסבר כיצד לשנות את נוסחת ה pagerank כך שתהיה העדפה מלאכותית של אתרים מסויימים בדרוג על פני אחרים.

III. 6% נתון הגרף הבא המייצג רשת דפי אינטרנט: הפעל את אלגוריתם HITS ודרג את האתרים על פי ערכי ה authority וה -hub שלהם (דירוגים נפרדים ל –hub ו – authority). אין צורך לחשב את הערכים המדוייקים רק לנמק את הדירוג.

H $Q_3 \to Q_1 \to Q_2$

A $Q_1 \to Q_2 \to Q_3$



2. 30% נתון אינדקס הופכי. כל כניסה באינדקס עבור term מסויים כולל מופעים בפורמט הבא:
.. x:y,z כאשר המשמעות היא שבמסמך x ה– term מופיע במיקומים y,z.

d: d: :

a: 1:3, 2:1,4
everywhere: 3:1
go: 4:6
had: 1:2
lamb: 1:5,2:3,6,4:2
little: 1:4, 2:2,5
mary: 1:1,3:3
sure: 4:4
that: 3:2
the: 4:1
to: 4:5
was: 4:3
went: 3:4

1: 1, 1,

1 ||||| 5
2 |||||| 6
3 |||| 4
4 |||||| 6

$d_3 = (0, \frac{1}{4}\cdot 4, 0, 0, 0, 0, \frac{1}{4}\cdot 2, 0, \frac{1}{4}\cdot 4, 0, 0, 0, \frac{1}{4}\cdot 4)$

5/3

א. 5% חשב את ה IDF של כל Term (יש להתעלם מה log בנוסחה)
ב. 5% הראה את הווקטור של מסמך 2 (ערכי tf-idf – שוב ללא log – כאשר tf מנורמל לאורך המסמך)
ג. 10% דרג את המסמכים במאגר לשאילתא : "mary lamb" על פי נוסחת ה cosine (ללא log בחישוב Tf-idf). בשאילתא- משקל כל term הוא 1.
ד. 10% משתמש החזיר feedback למנוע על המסמכים שהמנוע החזיר כרלוונטים. המשתמש סימן את מסמך 2 כרלוונטי ואת מסמך 3 כלא רלוונטי. עדכן את השאילתא בהתאמה ל - feedback על פי נוסחת rocchio כאשר $\alpha=1$, $\beta=0.5$, $\gamma=0.25$

3. 20%

I. 6% סמן על כל אחד מהמשפטים הבאים "נכון" או "לא נכון" ונמק **במשפט אחד** את תשובתך:
   i. הפעלת stemming תמיד מעלה את ה recall
   ii. הפעלת stemming תמיד מורידה את ה precision
   iii. כדי לשפר את תהליך האיחזור ניתן לבצע stemming רק למסמכי המאגר ללא stemming לשאילתא.

II. 14% להלן סט של 3 חוקים מתוך אלגוריתם ה Stemming של פורטר. החוקים מיוצגים על ידי תבניות של - s1->s2 (s1 מוחלף ב s2) בביצוע ה stemming מתבצע תמיד רק אחד מהחוקים בסט, על פי ההתאמה ה"ארוכה ביותר" לתבנית (longest match).

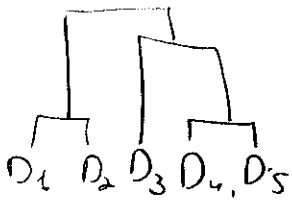
IES->I
SS->SS
S-> null

   i. 3% מדוע יש צורך בחוק : SS->SS
   ii. 3% הראה את תוצאות ה Stem על המילים הבאות: PONIES, CARESS,TIES
   iii. 4% אילו חוקים יש להוסיף לסט כדי לבצע stem נכון על המילים הבאות: CARESSES ,PONY
   iv. 4% תוצאת ה Stem של המילים PONIES ו-PONY יוצרות מילים לא "תקניות" באנגלית. האם לעובדה זו יש השפעה על תוצאות האיחזור? – נמק (במשפט)

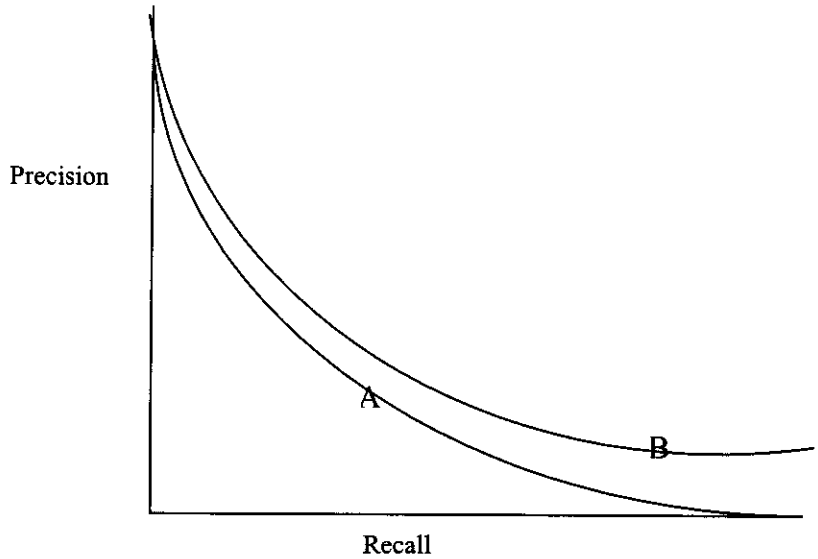
4. clustering: 14%

I. נתונה מטריצת **הדמיון** (לא מטריצת מרחקים!) הבאה בין מסמכים:
הראה את ה Clusters ההיררכים הצוברים (דנדוגרם) שיווצרו מהמטריצה הזו כאשר הדמיון בין ה clusters מחושב על פי complete link

	D1	D2	D3	D4	D5
D1	1.00	0.42	0.00	0.1	0.1
D2	-	1.00	0.00	0.07	0.2
D3	-	-	1.00	0.33	0.33
D4	-	-	-	1.00	0.36
D5	-	-	-	-	1.00



5. 16% נתונים גרפים Precision-Recall של שתי מערכות איחזור.
סמן על כל אחד מהמשפטים הבאים המתייחסים לתמונה : "**נכון**" , "**לא נכון**" או "**אי אפשר לדעת על פי הגרף**" (ללא הסבר)



I. מערכת B יעילה יותר ממערכת A
II. מערכת B איכותית (אפקטיבית) יותר ממערכת A
III. מערכת A איכותית יותר ממערכת B למאגרי מידע סטטיים (כגון מאגרים של רפואה)
IV. מערכת A לא הצליחה למצוא את כל המסמכים הרלוונטים לשאילתות שאותן הגרף מתאר

בהצלחה
ברכה וארז

I ①

$PR(Q1) = PR(Q3)/3$

$PR(q2) = PR(Q1)/2 + PR(Q3)/3$

לכן

$q2 > q1$

$q3 > q1$

$PR(Q2) = PR(Q4) = PR(Q1)/2 + PR(Q3)/3$

לכן גם

$q4 > q1$

לעומת q1 יקבל בוודאי.

$PR(Q3) = PR(Q2) + PR(Q4)$

לעומת

$Q3 > Q2$

$Q3 > Q4$

הדירוג הוא:

Q3

Q2, Q4 שווים

Q1.

II כדי להעדיף אתרים מסוימים ויש לשנות את ערכי d (decay factor)

כך שאתרים מועדפים יקבלו בחלוקת PR לאתרים מוקבלים מעל ודק.

Q1, Q2 מוגבלים

$H(Q3) = A(Q1) + A(Q2)$

$H(Q1) = A(Q2) = H(Q2) = A(Q1)$

$A(Q3) = 1$

HUB	AUTH
Q3	Q1, Q2
Q1, Q2	Q3

III

(3)

(1)

I - הפעלת stemming לכן מעלה את ה recall משום שהא גורמת לעיתים מסמכים להיות רלוונטי בכל סט של מסמכים יש אותו מספר של רלוונטים שגם נמצאו ולכן בהכרח, כמות הרלוונטים שנמצאו סה"כ (recall) תעלה.

יש מקרים שנגרמים את התפסה; כלומר, גם ה stem לא מוסיף מסמכים אז שבר $\frac{\text{המסמכים הרלוונטים נמצאו לפני ה stem}}{\text{recall, אבל ה stem}}$ לא מעלה, מי שעמן מקרים אלו (אז מקרי קצה אחרים בונים) קבלת את הנקודות.

II stem - מקצר מסמכים נספים ושמאזר חלקים רלוונטים, לא תמיד ה precision ירד, תלוי כמה שכן הרלוונטים לא רלוונטים לפני ה stem לוחם אחרי. למשל גם לפני ה stem הין מסמך אחד מתוך 2 (כלומר $\frac{1}{2}$ = pr) בסקבות ה- stem נמצאו לך 5 מסמכים שמתוכם 4 רלוונטים, הוא הוא $\frac{5}{7}$ שזה יותר מ $\frac{1}{2}$.

III stemming לא סימטרי (רק למסמכים) אינו מדויק, כי לשאלה התאמה בין השאלה למסמך יכן כאוכן שהיא לא משפר.

(2)

I החוק $SS \to SS$ מבלבל בין מילים בלשון הרבים (כשין students) שפרט להוריד להם את S הרבים לכן מילים שיש להם שני S כחלק מהמילה (כלומר ה-S הוא לא של רבים) ולכן לא נגיע להורדת S.

IV caress → caress, PONIES → PONI, TIES → TI;

III דרך לתוסף: SSES → SS, Y → I (אבל אם es אחרי פונה כמו gates)

IV - אין השפעה על התוצאות, לא משנה מה מופיע בשאילתה שהשאלה המסמך

שאלה 2

(א) מספר המסמכים במאגר = 4 לפי הנוסחא לחישוב $idf_j = \log_2(N/df_j)$: T_j

N הינו מספר המסמכים במאגר, df_j הינו מספר המסמכים בו $term_j$ מופיע.

אם מתעלמים מה-log נקבל $idf_j = \frac{N}{df_j}$

הביטויים המופיעים רק במסמך אחד הינם:

everywhere, go, had, sure, that, the, to, was, want.

לחישוב קבועה של ה-idf יהיה כפה והוא: $\frac{4}{1} = 4$

הביטויים המופיעים בשני מסמכים הם:

a, little, mary

לחישוב ה-idf הינו: $\frac{4}{2} = 2$

הביטויים המופיעים בשלושה מסמכים הינם:

lamb

לחישוב ה-idf הינו $\frac{4}{3} = 1\frac{1}{3}$

(2) אורך מסמך 2 הינו 6. ישנם שלושה ביטויים המופיעים בו. נחשב tf לכולם:

a: $tf = \frac{2}{6} = \frac{1}{3}$ $\quad tf \cdot idf = \frac{2}{3}$

lamb: $tf = \frac{2}{6} = \frac{1}{3}$ $\quad tf \cdot idf = \frac{1}{3} \cdot \frac{4}{3} = \frac{4}{9}$

little: $tf = \frac{2}{6} = \frac{1}{3}$ $\quad tf \cdot idf = \frac{1}{3} \cdot 2 = \frac{2}{3}$

לכל שאר הביטויים שלא מופיעים במסמך $tf=0$, לכן $tf \cdot idf = 0$.

לכן וקטור המסמך יראה כך: $Vd_2 = (\frac{2}{3}, 0, 0, 0, \frac{4}{9}, \frac{2}{3}, 0, 0, 0, 0, 0, 0, 0)$

(ג) נוסחת ה-cosine:

$$CosineSim(d_j, q) = \frac{d_j \cdot q}{|d_j| \cdot |q|}$$

$d_j \cdot q = \sum w_{ij} \cdot w_{iq}$

כאשר: $w_{ij} = tf \cdot idf$

$d_1 \cdot q = \frac{1}{5} \cdot 2 + \frac{1}{5} \cdot \frac{4}{3} = \frac{2}{3}$

$d_2 \cdot q = 0 + \frac{4}{9} = \frac{4}{9}$

$d_3 \cdot q = \frac{1}{4} \cdot 2 + 0 = \frac{1}{2}$

$d_4 \cdot q = 0 + \frac{1}{6} \cdot \frac{4}{3} = \frac{2}{9}$

לחישוב המכנה נחשב $|q|$

$$|q| = \sqrt{\sum w_{iq}^2} = \sqrt{1+1} = \sqrt{2}$$

$|d_j|$ נרמל

$$|d_1| = \sqrt{\overset{a}{\left(\tfrac{1}{5}\cdot 2\right)^2} + \overset{had}{\left(\tfrac{1}{5}\cdot 4\right)^2} + \overset{lamb}{\left(\tfrac{1}{5}\cdot \tfrac{4}{3}\right)^2} + \overset{little}{\left(\tfrac{1}{5}\cdot 2\right)^2} + \overset{mary}{\left(\tfrac{1}{5}\cdot 2\right)^2}} =$$

$$= \sqrt{\frac{4}{25} + \frac{16}{25} + \frac{16}{225} + \frac{4}{25} + \frac{4}{25}} = \sqrt{\frac{28}{25} + \frac{16}{225}} = \sqrt{\frac{268}{225}}$$

$$|d_2| = \sqrt{\overset{a}{\left(\tfrac{1}{3}\cdot 2\right)^2} + \overset{lamb}{\left(\tfrac{1}{3}\cdot \tfrac{4}{3}\right)^2} + \overset{little}{\left(\tfrac{1}{3}\cdot 2\right)^2}} = \sqrt{\frac{4}{9} + \frac{16}{81} + \frac{4}{9}} = \sqrt{\frac{8}{9} + \frac{16}{81}} = \sqrt{\frac{88}{81}}$$

$$|d_3| = \sqrt{\overset{everywhere}{\left(\tfrac{1}{4}\cdot 4\right)^2} + \overset{mary}{\left(\tfrac{1}{4}\cdot 2\right)^2} + \overset{that}{\left(\tfrac{1}{4}\cdot 2\right)^2} + \overset{went}{\left(\tfrac{1}{4}\cdot 4\right)^2}} = \sqrt{1 + \tfrac{1}{4} + 1 + 1} = \sqrt{\frac{13}{4}}$$

$$|d_4| = \sqrt{\overset{go}{\left(\tfrac{1}{6}\cdot 4\right)^2} + \overset{lamb}{\left(\tfrac{1}{6}\cdot \tfrac{4}{3}\right)^2} + \overset{sure}{\left(\tfrac{1}{6}\cdot 4\right)^2} + \overset{the}{\left(\tfrac{1}{6}\cdot 4\right)^2} + \overset{to}{\left(\tfrac{1}{6}\cdot 4\right)^2} + \overset{was}{\left(\tfrac{1}{6}\cdot 4\right)^2}} =$$

$$= \sqrt{\frac{4}{9} + \frac{4}{81} + \frac{4}{9} + \frac{4}{9} + \frac{4}{9} + \frac{4}{9}} = \sqrt{\frac{20}{9} + \frac{4}{81}} = \sqrt{\frac{184}{81}}$$

$$\text{CosSim}(d_1, q) = \frac{\frac{2}{3}}{\sqrt{\frac{268}{225}\cdot 2}} = 0.431$$

$$\text{CosSim}(d_2, q) = \frac{\frac{4}{9}}{\sqrt{\frac{88}{81}\cdot 2}} = 0.301$$

$$\text{CosSim}(d_3, q) = \frac{\frac{1}{2}}{\sqrt{\frac{13}{4}\cdot 2}} = 0.196$$

$$\text{CosSim}(d_4, q) = \frac{\frac{2}{9}}{\sqrt{\frac{184}{81}\cdot 2}} = 0.104$$

לפי נוסחת ה- Cosine הדירוג יהיה:

d_1
$\downarrow$
d_2
$\downarrow$
d_3
$\downarrow$
d_4

$$q_m = q_0 + 0.5 \cdot \frac{1}{|D_r|} \cdot \sum_{d_j \in D_r} d_j - 0.25 \cdot \frac{1}{|D_{nr}|} \cdot \sum_{d_j \in D_{nr}} d_j$$

③

המשתמש בירב מסמך אחד כרלוונטי, לכן $|D_r| = 1$

ומסמך אחד כלא רלוונטי, לכן $|D_{nr}| = 1$ אז לפי הנוסחא:

$$q_m = q_0 + 0.5 \cdot d_2 - 0.25 \cdot d_3$$

את וקטור השאילתא ואת d_2 אנו יודעים. נחשב את וקטור d_3:

$$d_3 = (0, 1, 0, 0, 0, 0, \tfrac{1}{2}, 0, 1, 0, 0, 0, 1)$$

$$d_2 = (\tfrac{2}{3}, 0, 0, 0, \tfrac{4}{9}, \tfrac{2}{3}, 0, 0, 0, 0, 0, 0, 0)$$

$$q_0 = (0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0)$$

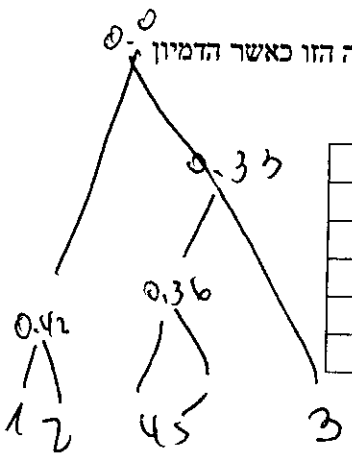
$$\Downarrow$$

$$q_m = (\tfrac{1}{2} \cdot \tfrac{2}{3}, -\tfrac{1}{4}, 0, 0, 1 + \tfrac{1}{2} \cdot \tfrac{4}{9}, \tfrac{1}{2} \cdot \tfrac{2}{3}, 1 - \tfrac{1}{4} \cdot \tfrac{1}{2}, 0, -\tfrac{1}{4}, 0, 0, 0, -\tfrac{1}{4})$$

$$q_m = (\tfrac{1}{3}, -\tfrac{1}{4}, 0, 0, \tfrac{11}{9}, \tfrac{1}{3}, \tfrac{7}{8}, 0, -\tfrac{1}{4}, 0, 0, 0, -\tfrac{1}{4})$$

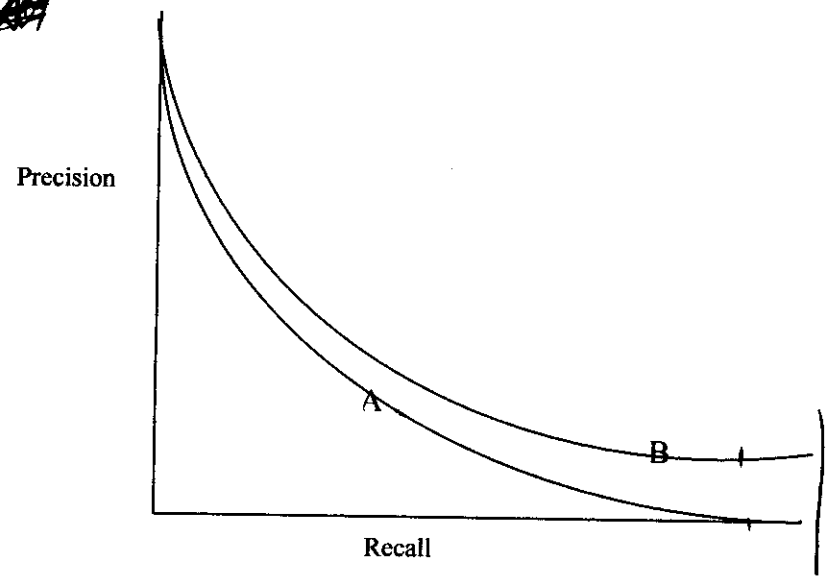
4. clustering: 14%

I. נתונה מטריצת **הדמיון** (לא מטריצת מרחקים!) הבאה בין מסמכים:
הראה את ה Clusters ההיררכים הצוברים (דנדוגרם) שיווצרו מהמטריצה הזו כאשר הדמיון בין ה clusters מחושב על פי complete link



	D1	D2	D3	D4	D5
D1	1.00	0.42	0.00	0.1	0.1
D2	-	1.00	0.00	0.07	0.2
D3	-	-	1.00	0.33	0.33
D4	-	-	-	1.00	0.36
D5	-	-	-	-	1.00

5. נתונים גרפים Precision-Recall של שתי מערכות איחזור. 16%
סמן על כל אחד מהמשפטים הבאים המתייחסים לתמונה : **"נכון"** , **"לא נכון"** או **"אי אפשר לדעת על פי הגרף"** (ללא הסבר)



I. מערכת B יעילה יותר ממערכת A – אי אפשר לדעת
II. מערכת B איכותית (אפקטיבית) יותר ממערכת A נכון
III. מערכת A איכותית יותר ממערכת B למאגרי מידע סטטיים (כגון מאגרים של רפואה) לא נכון
IV. מערכת A לא הצליחה למצוא את כל המסמכים הרלוונטים לשאילתות שאותן הגרף מתאר נכון.

בהצלחה
ברכה וארז

= ואולם מתייחסת לקיקופים של מהירות, שטח אחסון וכו'.
איכות אי אפקטיבית מתייחסת לאיכות ונראות הלתאור, כלומר
PR ו Recall.

= מכיוון שלמערכת A יש 0=PR בנקודת ה Recall הגבוה
(קרוב ל-1 ו-0), היא לא הצליחה למצוא את כל המסמכים