אוניברסיטת בן-גוריון - המחלקה להנדסת מערכות מידע
קורס איחזור מידע וספריות דיגיטליות
סמסטר חורף תשס"ח- 01.06.08 – מועד ב' - פתרון
ד"ר ברכה שפירא, ליהי נעמני

משך המבחן: שעתיים וחצי. יש לענות על כל 6 השאלות.
חומר עזר מודפס מותר ללא מחשב נייד (מחשבון מותר)

1. 20% נתונות התוצאות הבאות של הרצה של מנוע חיפוש מסויים שמתארות את מס' התוצאות הרלוונטיות שחזרו מתוך k המסמכים הראשונים שחזרו (המנוע החזיר סה"כ 50 מסמכים).

מס. הרלוונטים שחזרו	מס' המסמכים שחזרו
5	10
12	20
15	30
13	40
27	50

ענה על השאלות הבאות:

א. 3% בטבלת התוצאות שלעיל, יש תוצאה לא הגיונית. מהי התוצאה? הסבר מדוע היא לא הגיונית.

פתרון: לא יתכן שחזרו 15 מסמכים רלוונטים, ואחר כך חזרו רק 13. 13 לא הגיוני

ב. 5% שנה את המספר הלא הגיוני על ידי הגדלת ספרת העשרות שלו ב-1 . חשב: Precision בכל אחד מנקודות המסמכים (10,20,30,40,50,) על פי התוצאה המתוקנת.

פתרון: 5/10, 12/20, 15/30, 23/40, 27/50

ג. 5% הנח שתיקנת את המספר הלא הגיוני. האם אפשר לחשב MAP על פי התוצאות הנתונות. אם אפשר, חשב, אחרת הסבר אילו נתונים חסרים לך לחישוב.

פתרון: אי אפשר לחפש כי חסר מקום החזרה של כל מסמך וכן חסר מספר סך כל המסמכים הרלוונטים במאגר

ד. 7% בהנחה שבמאגר 30 מסמכים רלוונטים ושתיקנת את המספר הלא הגיוני, האם אפשר לשרטט גרף precision-recall המתאים לשאילתא על פי הנתונים הנתונים בשאלה. **אם אפשר** ציין את ערכי הגרף, **אם אי אפשר**, הסבר אילו נתונים חסרים, הנח נתונים מתאימים וציין את ערכי הגרף על פי הנתונים שהנחת.

פתרון: אי אפשר מאותם הסיבות כמו בסעיף ג'. ניתן עדין לפתור את השאלה אם מניחים הנחות. פתרון לדוגמא: נניח כי קיימים 30 מסמכים רלוונטים במאגר וכן נניח כי הטבלה למטה מציינת את המיקום של חזרת כל מסמך:

doc #	relevant y/n	recall	precision		doc #	relevant y/n	recall	precision
1	y	0.033333	1		31	y	0.533333	0.516129
2	y	0.066667	1		32	y	0.566667	0.53125
3	y	0.1	1		33	y	0.6	0.545455
4	y	0.133333	1		34	y	0.633333	0.558824
5	y	0.166667	1		35	y	0.666667	0.571429
6	n	0.166667	0.833333		36	y	0.7	0.583333
7	n	0.166667	0.714286		37	y	0.733333	0.594595
8	n	0.166667	0.625		38	y	0.766667	0.605263
9	n	0.166667	0.555556		39	n	0.766667	0.589744
10	n	0.166667	0.5		40	n	0.766667	0.575

11	y	0.2	0.545455
12	y	0.233333	0.583333
13	y	0.266667	0.615385
14	y	0.3	0.642857
15	y	0.333333	0.666667
16	y	0.366667	0.6875
17	y	0.4	0.705882
18	n	0.4	0.666667
19	n	0.4	0.631579
20	n	0.4	0.6
21	y	0.433333	0.619048
22	y	0.466667	0.636364
23	y	0.5	0.652174
24	n	0.5	0.625
25	n	0.5	0.6
26	n	0.5	0.576923
27	n	0.5	0.555556
28	n	0.5	0.535714
29	n	0.5	0.517241
30	n	0.5	0.5

41	y	0.8	0.585366
42	y	0.833333	0.595238
43	y	0.866667	0.604651
44	y	0.9	0.613636
45	n	0.9	0.6
46	n	0.9	0.586957
47	n	0.9	0.574468
48	n	0.9	0.5625
49	n	0.9	0.55102
50	n	0.9	0.54

מטבלה זו ניתן לחשב את הנקודות עבור הגרף: *******

standard recal points	precison
0	1
0.1	1
0.2	0.64
0.3	0.68
0.4	0.7
0.5	0.65
0.6	0.58
0.7	0.6
0.8	0.61
0.9	0.61
1	0

. 2 15%ציין על כל אחד מהמשפטים הבאים אם הוא **נכון** או **לא נכון**.

א. 5% אלגוריתם **k-means** מוצא את מספר ה Clusters האופטימלי

ב. 5% בניסוי להערכת מנוע שכלל 100 שאילתות, המדד שהשתמשו בו היה precision ב-30 מסמכים. נתון שהתוצאה הממוצעת על פני השאילתות של מנוע A גבוהה יותר מאשר התוצאה של מנוע B. אפשר להסיק מכך שעבור רוב השאילתות של מנוע A יהיה precision ב-30 גבוה יותר מאשר למנוע B.

ג. 5% מדד ה F-measure מחושב כממוצע הרמוני של precision ו-recall ולא כממוצע אריתמטי כדי לוודא ש F-measure גבוה מעיד גם על ערכים גבוהים גם של precision וגם של recall .

פתרון: לא נכון, לא נכון, נכון

3. מיפתוח 25%

א. 10%מיפתוח של ידיעות שהופיעו בשירות של חדשות דיגיטאליות של NY-Times בין השנים 1991-1995 הראה שהמאגר כולל בערך 400 מליון מילים וגודל ה vocabulary (של האינדקס) הוא מליון מילים. הערך כמה מילים יהיו במאגר ומה יהיה גודל ה vocabulary אם נמפתח את המאגר של NY-Times בין השנים 1991-2000 בהנחה של קצב קבוע של יצירת הודעות בכל שנה.

בהנחה כי הקצב נשאר קבוע, כמות המילים תוכפל ויעמוד על 800M. נשתמש בנוסחא:

$v = k\sqrt{N}$ כאשר V=1M , N=400M. ולכן: k=50

ולכן: $50\sqrt{800M} = 1.4M$

ב. 15% כאשר רוצים לבזר אינדקס ישנן 2 דרכים עקרוניות לחלק את קובץ ה Posting של האינדקס לצורך ייעול על ידי ביזור האינדקס. בשיטה אחת הקובץ מחולק לחלקים על פי טווח המילים שהוא כולל (כלומר כל חלק של הקובץ מכיל מופעים של אוסף אחר של מילים) ואילו בשיטה השנייה כל חלק של הקובץ כולל את כל טווח המילים באינדקס אבל רק חלק של המופעים שלהן במסמכים. ציין לכל שיטה 2 יתרונות (שונים) לפחות.

פתרון:

יתרונות אינדוקס לפי טווח:

1. חיפוש מהיר אם שתי המילים אימן נמצאות באותו החלק
2. ניתן לבנות אינדקס חכם עבור ביטויים או מילים שקשורות אחת לשניה

יתרונות אינדוקס לפי חלקי מופעים:

1. ניתן ליצור הקבלה גמורה בחיפוש וכך ליצור חיפוש מהיר יותר
2. אם כל קובץ מאוחסן על שרת אחר, ושרת אחד נופל, עדיין ניתן להחזיר תשובה, גם אם היא חלקית

4. 20% הנח את הגרף הבא המתאר חלק ב Web:

A מצביע ל C ו-D
B מצביע ל A ו-C
C מצביע ל –B
D מצביע ל – E
E מצביע ל – A

א. 5% הגדר את מטריצת המעבר (transition matrix) של הגרף הזה (רמז: המטריצה שעל פיה ניתן לחשב pagerank)

פתרון:

	A	B	C	D	E
A	0	0.5	0	0	1
B	0	0	1	0	0
C	0.5	0.5	0	0	0
D	0.5	0	0	0	0
E	0	0	0	1	0

ב. 5% הראה חישוב של איטרציה אחת של חישוב pagerank באמצעות המטריצה (**לא** באמצעות נוסחת pagerank), הנח ערכים תחילים – 0.2 לכל צומת (לצורך השאלה **לא** להתחשב ב surfer model, כלומר לא להתחשב בהסתברות מעבר לצמתים באופן רנדומי)

PR(a) = (0.5*0.2)+(1*0.2)
PR(b)= 0.2*1
PR(c)= (0.5*0.2)+(0.5*0.2)
PR(d)=0.5*0.2
PR(E)=0.2*1

ג. 5% הראה את המטריצות שבאמצעותם מחשבים את אלגוריתם HITS (מטריצת הקשר וההופכית שלה).

פתרון:

המטריצה:

0	0	1	1	0
1	0	1	0	0
0	1	0	0	0
0	0	0	0	1
1	0	0	0	0

ההופכית:

0	1	0	0	1
0	0	1	0	0
1	1	0	0	0
1	0	0	0	0
0	0	0	1	0

ד. 5% הראה איטרציה אחת של חישוב Auithority על פי אלגוריתם HITS באמצעות המטריצות הנ"ל. הנח ערכים תחיליים של 1 לכל צומת.

פתרון:

$M^T x M x A$

3
1
4
2
1

10%.5 להלן כמה כללים מתוך ה Stemmer של פורטר.

```
Step 1a
   SSES -> SS
   IES  -> I
   SS   -> SS
   S  ->
   ……..
```

Step 4

```
   (m>1) AL    ->
   (m>1) ANCE  ->
   (m>1) ENCE  ->
```

```
(m>1) ER    ->
(m>1) IC    ->
(m>1) ABLE  ->
(m>1) IBLE  ->
(m>1) ANT   ->
(m>1) EMENT ->
(m>1) MENT  ->
(m>1) ENT   ->
(m>1 and (*S or *T)) ION ->
(m>1) OU    ->
(m>1) ISM   ->
(m>1) ATE   ->
(m>1) ITI   ->
(m>1) OUS   ->
(m>1) IVE   ->
```

הראה עבור המילים הבאות מהו ערך של m (הראה את חלוקת המילה לפי האלגוריתם) ומהו ה Stem שלהם , הראה אילו חוקים הפעלת:

Placement
rigorous

פתרון:

Plac, m=3 (pl-ac-em-ent)
Rigor, m=3 (r-ig-or-ous)

6. 10%

נתון המאגר הבא:

D1: John gives a book to Mary
D2: John who reads a book loves Mary
D3: who does John think Mary loves?
D4: John thinks a book is a good gift.

הנח שמהמאגר נופו Stop-Words שכיחות באנגלית וכן הופעל עליו suffix stemmer .
הראה את המבנה ההופכי (vocabulary+posting) המתאים למנוע על פי המודל הווקטורי, כאשר על כל מילה נשמר ערך tf*idf שלה. tf מנורמל לאורך המסמך. (אין צורך לחשב את ה log- מספיק רק להציב בנוסחא).

פתרון:

****

	Log(N/dft)				
book	log(4/3)	d1:1/4	d2:1/5	d4:1/5	
gift	log(4/1)	d4:1/5			
give	log(4/1)	d1:1/4			
good	log(4/1)	d4:1/5			
John	log(4/4)	d1:1/4	d2:1/5	d3:1/4	d4:1/5
love	log(4/2)	d2:1/5	d3:1/4		
Mary	log(4/3)	d1:1/4	d2:1/5	d3:1/4	
read	log(4/1)	d2:1/5			
think	log(4/2)	d3:1/4	d4:1/5		

בהצלחה

ברכה וליהי