בס"ד

איחזור מידע תשע"א – 372.1.4406

סמסטר חורף מועד א' 31.01.11

ד"ר ברכה שפירא, אורלי מורנו

מספר נבחן 28/1

משך המבחן : שעתיים וחצי

חומר עזר: מותר (**לא מחשב נייד**) – **יש להחזיר את הטופס**

חלק א.

סמן תשובה אחת נכונה –**יש לסמן תשובות על טופס המבחן**.:

1. 4% מנוע חיפוש מחזיר לכל שאילתא 10 מסמכים (מתוכם, חלקם רלוונטים). הוחלט לנסות לשפר את ביצועי המנוע על ידי הגדלת מספר המסמכים המוחזרים לכל שאילתא מ 10 ל 1000 , כלומר, יוחזרו עוד 990 מסמכים על 10 המוחזרים. (המספר נשאר).

אם אף מסמך נוסף לא זוהה כרלוונטי בקבוצת המסמכים הנוספת המוחזרת אז :

א. ה- Precision ו ה Recall של המנוע יישארו אותו דבר גם לאחר הוספת המסמכים לתוצאה
ב. לאחר ההוספה, ה precision יקטן ו ה Recall יישאר אותו דבר
ג. ה Precision וה Recall יקטנו לאחר ההוספה
ד. ה r-precision יקטן
ה. אף תשובה לא נכונה

Recall- גדל
Precision- יקטן

2. 4% (בהמשך ל -1). אם לפחות מסמך אחד נוסף זוהה כרלבנטי בקבוצת המסמכים הנוספת שהוחזרה, אז:

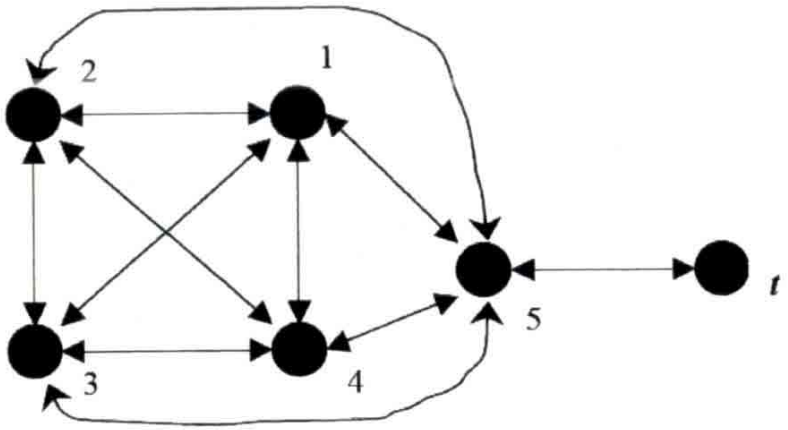
א. ה precision לאחר ההוספה יהיה בהכרח נמוך יותר
ב. ה precision לאחר ההוספה יהיה בהכרח גבוה יותר
ג. ה Recall לאחר ההוספה יהיה בהכרח גבוה יותר
ד. ב ו-ג נכונים
ה. הprecision ב – 10 מסמכים יגדל

Recall - יעלה

3. 4% Interpolated average precision הוא מדד יעיל ל-

א. השוואה בין מנועים ששונים רק בסדר המסמכים שמוחזרים
ב. השוואה בין מנועים שמתחשב גם ב Recall
ג. השוואה רק בין ה precision של המנועים
ד. ציור של גרף precision-recall על פני הרבה שאילתות
ה. ב+ד נכונים

4. 8% נתונה הרשת הבאה :



כאשר הסימון  שווה ל 

לרשת שכללה רק את הצמתים 1-5 , נוספה הצומת t , כיצד ישתנו ערכי ה pagerank של הצמתים ברשת בעקבות ההוספה:

א. כל ערכי ה pagerank של כל הצמתים ישארו שווים לערכם לפני ההוספה, גם ערך צומת t יהיה שווה לערכי צמתים .
ב. הערך של הצמתים 1-4 יקטנו במידה שווה
ג. הערך של צומת 5 יגדל ויהיה הגבוה ביותר ברשת
ד. הערך של צומת t יהיה הקטן ביותר ברשת
ה. הערך של צומת 5 יקטן
ו. ה+ב נכונים
ז. ב+ג+ד נכונים
ח. רק ב+ג נכונים

5. 16% ציין נכון או לא נכון ליד כל אחד מהמשפטים הבאים:

4% א. במנוע חיפוש באינטרנט F- measure הוא מדד חשוב — נכון
4% ב. נניח מאגר מסמכים המכיל את המסמך השגוי הבא: "שלום טמבל מה העניינים ואיך אתה מרגיש" – . נניח מנוע הפועל לפי המודל הווקטורי. נניח שהטעות (מופע של המילה "טמבל" לא רלוונטית למסמך) התגלתה. תיקון הטעות ישפיע רק על המסמך השגוי. — נכון
4% ג. שאילתא הכוללת מילה שיש לה מילים נרדפות רבות בהכרח תסבול מ precision נמוך. — לא נכון
4% ד. תור של Crawler של מנוע המיועד לבניית מאגר בנושא כלשהו יכלול רק דפים הרלוונטים לנושא המאגר — נכון

חלק ב.

ענה על השאלות הבאות:

6. 7% נתונה טבלת הדירוגים הבאה של 5 משתמשים (user1..user5) לחמישה items (A..E). כמו כן נתונה הקרבה ($W_{5,i}$) בין כל אחד מה-users ל user5 (המוגדר כ Active user). יש לנבא את הרלוונטיות של D item ל -user 5 על פי **שני** השכנים הקרובים אליו ביותר (יש לחשב את ($P_{5,D}$)).

Item	User1	User2	User3	User4	User5 (active user)
A	10	5	9		9
B	6	9		5	5
C	2	7	3		1
D	4	8	3	3	2.526
E	8	1	9	2	
דמיון ל user5 $W_{5,i}$	1	-0.5	0.9	0.7	

7. 15% הצע מדד Discount Cumulative Gain (DCG) קבוצתי, שמפחית ציון של מסמכים בקבוצות של 5. כלומר הפחתה מסוימת ל- 5 המסמכים הראשונים, הפחתה מוגדלת ל-5 מסמכים השניים וכו'. במסמכים שבתוך כל קבוצה ההפחתה שווה.
כתוב נוסחה מדויקת ל DCG קבוצתי לדרגה K כלשהי, הנוסחה צריכה לכלול תמיכה ב- K שאינו בכפולות של 5. יש להסביר בקצרה את ההיגיון של הנוסחה.

כמות למה

$$DCG = \sum_{j=0}^{\lfloor K/5 \rfloor} \sum_{i=1}^{5} \frac{rel(5j+i)}{\log_2(j+2)} + \sum_{i=1}^{K-y} \frac{rel(y\times 5+i)}{\log(y+3)}$$

$y = \frac{K}{\lfloor 5 \rfloor}$ אוגדת למה אם 3 בריק.

ההסבר בצבים

8. 42% במאגר של 25000 מסמכים צירפו לכל מסמך רשימה מדורגת (לא ריקה) הכוללת 1-10 מילים לכל מסמך של מילות מפתח מתוך מאגר של 100 מילות מפתח הרלוונטיות למאגר (המילים לא בהכרח מופיעות במסמך אליו הן מצורפות). סדר המילים המצורפות למסמך מעידה על חשיבותן למסמך (המילה הראשונה חשובה ביותר, השנייה פחות מהראשונה, השלישית פחות מהשנייה וכו'). שאילתא במנוע שיועד למאגר היא רשימה של מילות מפתח מתוך 100 המילים (ללא חשיבות לסדר המילים בשאילתא), אין משקל למילים בשאילתא.

- פונקציית הדירוג של מסמך לשאילתא צריכה :

עבור שאילתא של מילה אחת להחזיר מסמכים שברשימה של מילות המפתח שלהם מופיעה מילת המפתח של השאילתא ולדרג את המסמכים המוחזרים לפי המיקום של המילה ברשימת מילות המפתח של המסמך.

עבור שאילתא של יותר ממילה אחת, פונקציית הדירוג צריכה להתחשב במספר המילים המשותפות בין מילות המפתח של המסמך ומילות השאילתא וכן בסדר של מילות המפתח במסמך.

(לדוגמא, מילות המפתח של מסמך A כוללות שתי מילים של השאילתא- מילה אחת של השאילתא היא ראשונה במילות המפתח של המסמך, והמילה השניה של השאילתא היא השניה במילות המפתח של מסמך A,

מילות המפתח של מסמך B גם כוללות שתי מילים מהשאילתא- מילה אחת של השאילתא נמצאת במקום השלישי במילות המפתח של המסמך, והשניה במקום ה -5 במסמך. A ידורג יותר גבוה מ B.)

א. 10% הצע מודל ווקטורי (עם 100 ממדים) שיתמוך בפונקציית הדרוג הנ"ל: תאר מה יכלול ווקטור של מסמך. תאר בקיצור כיצד נקבע המשקל של מילת מפתח בווקטור של מסמך (כלומר, מהי פונקציית המשקול של מילה במסמך). אין צורך לחשוב על מודל חסכני בזכרון.

ב. 12% כתוב פונקציות דרוג מדוייקות שיבטאו את דרישת הדרוג לעיל. פונקציה אחת לשאילתות של מילה אחת, ופונקציה שניה לשאילתות עם יותר ממילה 1.

ג. 14% כיצד ניתן לשלב את המודל שהצעת ב-א עם מודל ווקטורי סטנדרטי – כלומר מסמך ייוצג גם על ידי מילות מפתח מהרשימה וגם על ידי המילים המופיעות במסמך ושאילתא תוכל לכלול גם מילים מהמסמך וגם מילות מפתח.

- 6% כיצד ייראה ווקטור של מסמך במודל המשולב
- 8% כיצד יחושב הדמיון בין שאילתא ומסמך (נוסחא מדויקת)

ד. 6% ציין שיפור אפשרי של ביצועי מנוע המבוסס על המודל המוצע (כלומר, מדוע כדאי להשתמש במודל כזה, מה הוא יכול לשפר).

בהצלחה - ברכה ואורלי

שאלה 6

ניקח את user 1 ו- user 3 מכיוון שהם הכי דומים ל- user 5

$$P_{5,0} = 5 + \frac{(4-6)\cdot 1 + (3-6)\cdot 0.9}{1+0.9} = 5 - 2.473 = 2.526$$

שאלה ד

נסמן $g = \frac{k}{\lfloor 5 \rfloor}$

* הנוסחה גם נכונה כרגע
ניתן בטלפון.

$$DCG = \sum_{j=0}^{\frac{k}{\lfloor 5 \rfloor}-1}\left[\sum_{i=1}^{5}\frac{rel(5\cdot j+i)}{\log_2(j+2)} + \sum_{i=\frac{k}{\lfloor 5 \rfloor}}^{k-\frac{k}{\lfloor 5 \rfloor}}\frac{rel(g\cdot 5+i)}{\log_2(g+2)}\right]$$

הסימון הראשון עובר על הקבוצות (הקבוצה הבודדה)
כאשר הקבוצה האחרונה $j=0$ ואחרונה היא $\frac{k}{\lfloor 5 \rfloor}-1$ ולכן עבור
לדוג' $k=17$ מהיא קבוצה $\left(\frac{17}{\lfloor 5 \rfloor}-1=2\right)$ 3 בפעם מ-0 עד 2 כולל.
בתוך הסימון שלו j יש סימון של i שרץ עבור כל איבר
בקבוצה. לדוג' עבור קבוצה 0 יהיה לו rel1, rel2, rel3 ומתחלק
ב-1 ($\log_2 2$) ובמקום ה- log יכול ב-1 כי הוא ה- j שלו.
ובמקרה שיש לי שארית אז יש לי סימון חיצוני לסמר'
אשר רץ רק על השארית ובמקום DCG גם מקבל דגל
ב-1 מהיום שהפסקתי את ה- $\sum$ של j.
נסכם את הכל ונקבל DCG עבור k נקודות להבנת
של 5.

ב

שאלה 8

	Term 1	Term 2	.....	Term 100
DOC 1	$\frac{1}{doc(1,1)}$			
DOC 2				
.....				
DOC 25,000				

(א)

לכאורה כל מסמך יכול לכלול 100 מילות. כאשר בכל אחד יש את המשקל של ה Term שבחרנו לציין אותו. כאשר משקל נתבע $w_{ij} = \frac{1}{doc(i,j)}$

כאשר $doc(i,j)$ הוא מספר המילה (בפרק). כלומר אם המילה שהמילה הראשונה מופיע ה-50 שמיעות היא "מקבלת" את המשקל שלה היה $\frac{1}{7}$ (המשקל של המילה הבאה יהיה $\frac{1}{2}$ נניח אולי זו $\frac{1}{10}$ (המילה האחרונה) שאר המילים יקבלו משקל 0 כי לא בחרנו אותם לציין המסמך.

(ב) נוכל את השאילתה גם כוקטור בגודל 100 מימדים כאשר אם ה- Term מופיע בשאילתה אז משקלו יהיה 1

$Q = (x, y, \ldots\ldots, z)$

לאחר מכן נבצע בוקטוריות cosim ומכן הכוקטוריה הנתונה תגבה שאילתה עם איבר 1, 2 מופיעים 100 איברים.

$$cosim(d_j, q) = \frac{\sum (w_{ij} \cdot w_{iq})}{\sqrt{\sum (w_{ij}^2) \cdot (\sum w_{iq})^2}}$$

התרמול בקוסינוס מעלה את הדמיון. לדוגמא שאילתה לחשבו מסמכים שמופיעים שם מילה מסוימת הנמצאה במסמך A במקום תדר יותר מ-1 מאשר במסמך B כי - w_{ij} פלה יהיה גבוהה יותר ולכן תדירות גבוהה יותר.

(-2) אולי דבר לשאילתה לא מולה המבחה הקשרים לחם לוקן מולה מפתח, התרמול ובחלה את השאילתה, לא הקשר לתחשב בכמה אלה רק במספר המילים והחשיבות שלהם, לא בשאר המילים!

ב) נניח שיש לנו וקטור רשימת ת"ז מתן משקל נוסף למילות מפתח בנוסף למילים הרגילות (גם לשאילתה).

$doc_i = (x, y, \ldots, z)$

כאשר האינדקס הוא מספר המילים במאגר הכולל **אחוד** לכל המילות מפתח (לא נרצה כפילויות מילים). כאשר משקל יחושב כך:

$$w_{ij} = \begin{cases} d(tf \times idf) + (1-d)\cdot k_{ij} & \leftarrow \text{אם מדובר במאגר מילות מפתח} \\ (tf \times idf) & \leftarrow \text{אם מילה לא הופיעה במילות מפתח} \end{cases}$$

$\boxed{k_{ij}}$ - זה המשקל ברשימת הקישים

$\boxed{d}$ - אחוז המשקל שנרצה לתת למילות מפתח ומילים במאמרים.

אם בהמשך נרצה להגדיל למצב שבו נשווה רק מילות מפתח נציב $d=0$ ולאמוץ וקטורי רשימה נציב $d=1$. וכל d אחר ניתן לי אפשרות לשמר לכל שילוב 2 המודלים. בנוסף נרצה שהמשקל יהיה מנורמל.

- מבינים שכאשר התמשבנו במילות מפתח במסמכים אנו מוציאים לא לתת להם גם משקל בשאילתה לכן בשאילתה יראה כמו במודל וקטורי רגיל

$q = (x, y, \ldots, z)$

כאשר אורכה הוא במספר ה-terms באוצר כולו. ולכן משמש ה similarity יתבצע שוב בעזרת cosim

והמשקלים? על הכינוסים בשאילתה אין משקל?

$$cosim = \frac{\vec{d_j}\cdot\vec{q}}{|\vec{d_j}|\cdot|\vec{q}|}$$

יפה מאוד!

ג) כפי שראינו לעיל משתמשים ב- keywords (בדרך כלל 5-10) מילים כשבכל לשאיר למנוע חיפוש להבין מה הנושא של הדף. לא תמיד הנוכחות של הדף מעיד על הנושא שלו. לדוגמא דף של עוד מרמים לא מכיל מילים שמרמזות על מרמים ולכן נרצה לתת keywords שיעזרו לנו את הדף שלנו למנוע.

לה נכון

אפשר בהרבה

אפשר להשתמש בשיטה זה בשביל להעלות את הדירוג במנוע. בעיקר במאגרים פרטיים כמו web בשביל מאגרים גדולים

כאן מאחרים רבים אולי לתת יותר משקל למסמך ופחות למפתחות
ובמאחרים מסוימים נתת יותר משקל למפתחות (אם הייתם).

-2

התחלה יפה
אבל קצת חבל
בהשאלה.
לה בכן שם אם סל
תופס מלה סבבפסן ונוכל
לקלף עדיין שהמסמך בנושא
לא ברור כל כך מהמתה
ההגשה ולא שאלנו אתי
רצאי להשתמש בו אלא
נה השפור והשפור
העיקרי הוא התיחסות ל-concept