3/3

בס"ד

איחזור מידע תשע"ג – 372.1.4406
סמסטר חורף מועד א' 19.01.2014
פרופ' ברכה שפירא, אביגיל פרדיס, יוסי בן-שלמה

משך המבחן : שעתיים וחצי
חומר עזר: מותר (**לא מחשב נייד**), **מותר מחשבון**

יש להחזיר את גיליון הבחינה. יש לענות על כל השאלות, למבחן 3 דפים

1. **16% יש לענות לשאלה זו על גבי גליון הבחינה**
סמן תשובה אחת- נכונה (נא לא לצרף הסברים - הם לא ייקראו)

1.1 4% החיסכון **המשמעותי** מביצוע stem נובע מ:
א. חיסכון בגודל המילון ובזמן חיפוש עליו (dictionary)
ב. חיסכון בגודל ובזמן חיפוש ב posting
ג. חיסכון בזמן של ביצוע פעולות Text על המסמכים בשלב יצירת האינדקס
ד. א+ב
ה. כל התשובות נכונות

1.2 4% אתיקה של crawler של מנוע חיפוש מתבטאת ב:
א. הפעלה של איסוף נתונים מאתרים על פי זמנים שנוחים לאתר על פי הגדרתו (של האתר)
ב. ציות למה שמוגדר ב Robots.txt באתר
ג. הגבלה של מספר הדפים הנסרקים בו זמנית מאותו אתר
ד. השמטה של נתונים פרטיים של משתמשים מהדפים שנאספים
ה. בקרה על תכיפות הגישה לאותו דף
ו. כל התשובות נכונות
ז. א+ב+ג נכונים
ח. ב+ג+ה נכונים

1.3 4% בעת ביצוע ההחלקה לצורך חישוב דמיון של שאילתא למסמך באמצעות מודל שפה, משתמשים במודל השפה של כלל המסמכים במאגר של המנוע ולא בקבוע כלשהו או במאגר מסמכים חיצוני כלשהו. משום ש:
א. לא כל מונחי השאילתה בהכרח מופיעים במסמך
ב. רוצים לייצג באופן טוב את סביבת הנתונים בה אנחנו פועלים – כלומר את המאגר של המנוע
ג. אמנם כל מאגר מסמכים גדול יתאים, אבל המאגר של המנוע זמין ללא צורך בעיבוד נוסף
ד. השימוש במודל השפה של המאגר של המנוע חוסך את הצורך בהגדרת פרמטר חופשי נוסף
ה. השימוש במודל השפה של המאגר של המנוע מונע הסתברות 0 למונח כלשהו
ו. כל התשובות נכונות

1.4 4% כאלטרנטיבה ל IDF)(inverse Document frequency) כמרכיב למשקול term הוצע להשתמש ב ICF (inverse collection frequency) – שהוא המספר ההופכי של כמות מופעים במאגר.

א. הסיבה להעדיף IDF על פני ICF היא כי אפשר לחשב IDF מראש

ב. אין שום עדיפות ל IDF על פני ICF

ג. IDF עדיף משום ש ICF לא מדגיש ומבדיל את יחודיות ה term למסמך ויכול להיות מוטה מ terms שמופיעים הרבה פעמים במעט מסמכים.

ד. ICF עדיף משום שהוא מודד כמה ה term ייחודי במאגר על פי מספר הופעותיו הכולל, ככל שהוא מופיע פחות, הוא יותר ייחודי למסמכים שבהם הוא מופיע.

ה. א+ג נכונים

2. 16% על שני מנועי חיפוש: e1,e2 הורצו שתי שאילתות: Q1,Q2 . ידוע שלשאילתא Q1 20 מסמכים רלוונטים במאגר, לא ידוע מספר המסמכים הרלוונטים במאגר לשאילתא Q2.
להלן תוצאות ההרצה של השאילתא לשני המנועים (+ מסמן מסמך רלוונטי שהוחזר, - מסמן מסמך לא רלוונטי שהוחזר):

מנוע E1		מנוע E2		
Q1	Q2	Q1	Q2	
-	+	-	+	1
-	+	+	-	2
+	+	-	-	3
+	+	-	+	4
+	-	+	-	5
-	+	+	+	6
-	-	-	+	7
-	-	+	+	8
-	-	+	+	9
-	-	+	-	10

א. 5% הסבר – האם (ואם כן, כיצד) ניתן להעריך את מספר המסמכים הרלוונטים הנמצאים במאגר לשאילתא Q2 (בהנחה שאפשר להריץ עוד שאילתות על כל אחד מהמנועים)?

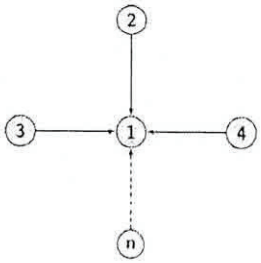
ב. **11% חשב, או הסבר איזה נתונים חסרים לחישוב**, לערכים הבאים:

1. 2% mean average precision על פני שתי השאילתות למנוע E1.
2. 2% DCG לשאילתא Q1 למנוע E2.
3. 2% E-measure לשאילתא Q1 למנוע E1 כאשר ניתן משקל כפול ל precision לעומת ה recall.
4. 2% R-precision למנוע E1.
5. 3% Interpolated average precision למנוע E2 ושאילתא Q1.

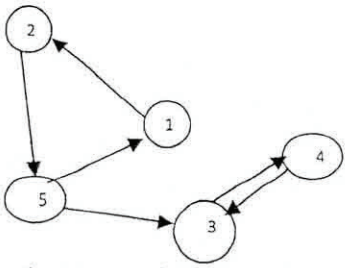
3. 15% חברה רוצה לפתח מנוע חיפוש שיחזיר תוצאות רלוונטיות אבל גם מגוונות. הכוונה בגיוון היא שהמסמכים החוזרים לא יהיו דומים יותר מדי אחד לשני, שהמשתמש ירגיש שהוא מקבל את כלנקודות המבט על השאילתא ששאל. הגדר מדד חדש שיוכל לשמש להערכת המנוע הנ"ל (יש להסביר כיצד המדד שהוגדר משלב בין רלוונטיות וגיוון, וכיצד הוא מודד גיוון). יש להגדיר נוסחא מדוייקת למדד.

4. 20% ניתוח הרשת :

4.1 5% הנח רשת במבנה של כוכב (לפי הציור, יכולים להיות עד n צמתים לרשת). חשב את ה Hub ו ה authority של הצמתים כפונקציה של n.



4.2 5% הסבר מה יהיו ערכי ה pagerank של הרשת הבאה- הנח d=0.85. אין צורך לחשב , אלא להעריך מה יהיו הערכים.



4.3 10% הסבר האם אלגוריתם HITS מצריך טיפול מיוחד בdead ends וב leaks (spider trap) בדומה לאלגוריתם pagerank (תן הסבר גם אם התשובה חיובית או שלילית).

5. 13% Spelling

5.1 6% מה המרחק (levenstein distance) בין זוגות המילים (חשבו את המעבר של המילה הכתובה משמאל למילה הכתובה מימין):

apfel, apple

apfel, apples

פרט את הפעולות שבצעת על כל זוג מילים כדי לחשב את המרחק.

5.2 7% מצא מילה נוספת עם מרחק זהה למרחק שבין apfel ו apple. הנח שמשתמש טעה (שגיאות spell)שלח את apfel כשאילתא למנוע. בהנחה שרק שתי המילים (apple והמילה השנייה שמצאת) נמצאות במילון של מנוע. הסבר **בדיוק** איך תדע לאיזו מילה מבין שתי המילים האלה התכוון המשתמש לשלוח כשאילתא.

6. 8% נניח שמאגר מסמכים כולל רק 4 מילים שכיחות : a, b, c, d (ואף לא מילה אחרת). סדר השכיחויות : a > b > c > d. מספר המילים הכולל במאגר הוא 5000. הנח שהמאגר מתנהג בדיוק לפי חוק Zipf. מה השכיחות של כל אחת מהמילים במאגר?

7. 12% כתוב אלגוריתם לחישוב הדמיון בין terms במאגר (בין ה terms לעצמם) (יש לכתוב פסאודו-קוד מדוייק), תן דוגמא לשימוש באלגוריתם כזה.

בהצלחה – ברכה אביגייל ויוסי