בס"ד

איחזור מידע תש"ע – 372.1.4406
סמסטר חורף מועד ב' 21.02.10
ד"ר ברכה שפירא, איליה פרידמן

משך המבחן : שעתיים וחצי
חומר עזר: מותר (**לא מחשב נייד**)
יש להחזיר את השאלון – **יש לענות על שאלה 5 על גבי הגליון.**

1. 8% מהו ה IDF של TERM שמופיע בכל מסמך במאגר ? האם להכללה של term כזה ברשימת stopwords יש אותה תוצאה מכל הבחינות?

2. 12% מנועי חיפוש רבים נוהגים ליצור שתי שכבות של קבצי posting – בשכבה הראשונה נמצאים מסמכים יותר איכותיים (עם pagerank יותר גבוה, יותר מופעים של ה term בכניסה של ה 'Term, עם סיכוי נמוך יותר ל spam וכו'). המנועים נוהגים לחפש מסמכים מתאימים לשאילתא קודם כל בשכבה הראשונה, אם הם לא מוצאים מספיק מסמכים הם עוברים לשכבה השניה.
   א. 6% האם מודל שתי השכבות משפר את הדיוק של החיפוש? הסבר.
   6% תאר לפחות שתי דרכים שבהן מודל שתי השכבות יכול לשפר את היעילות של החיפוש מבחינת זמן?

3. 20%

נתונה רשימת 20 המסמכים הראשונים (משמאל לימין) שמנוע החזיר לשתי שאילתות q1,q2 כאשר R מסמן מסמך רלוונטי ו-N מסמן מסמך לא רלוונטי. במאגר 10,000 מסמכים, לשאילתא q1 12 מסמכים רלוונטים ולשאילתא q2 8 מסמכים רלוונטים.

Q1: R R N N N N N N R N R N N N R N N N N R
Q2: N N R R N R R N R N N N N R R R N N N N

א. 4% מהו ה Precision ו ה f-meausre של שאילתא q1 ?
ב. 2% מהו ה precision ב 25% recall ללא אינטרפולציה עבור שאילתא q1?
ג. 8% מהו ה precision עם אינטרפולציה **על פי שתי השאילתות** בנקודות 0.1, 0.2, 0.4, recall 0.9
ד. 2% מהו ה r-precision לשאילתא q2?
ה. 4% חשב BPref לכל אחת מהשאילתות.

4. 10% שאילתא תחילית של משתמש היא:
cheap CDs cheap DVDs extremely cheap CDs
המשתמש קיבל שני מסמכים מהמנוע וסימן אחד כרלוונטי ואחד כלא רלוונטי:

מסמך רלוונטי	CDs cheap software cheap CDs
מסמך לא רלוונטי	cheap thrills DVDs

הנח שלצורך חישוב הווקטור של השאילתא והמסמכים משתמשים רק ב tf מנורמל לאורך המסמך – ללא idf. מה תהיה השאילתא המעודכנת לאחר הפעלת אלגוריתם rocchio אם משקל של השאילתא הנוכחית הוא 1, משקל של מסמך רלוונטי 0.75, משקל מסמך לא רלוונטי 0.25.

5. 14%

א. (6%) נתונה שאילתא q1 : A and B and C
נתון מסמך d1 המכיל את ה - A Terms ו- B
ומסמך d2 שלא מכיל אף אחד מ ה -Terms הנ"ל.

סמן נכון/לא נכון ליד כל משפט

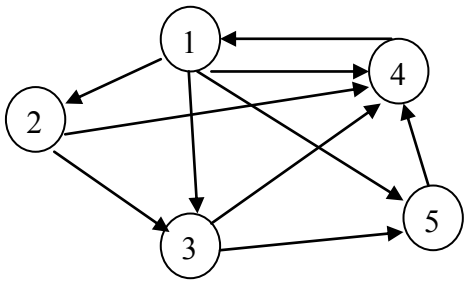
1. 2% מודל בוליאני טהור לא יחזיר אף אחד משני המסמכים, d1, d2 לשאילתא q1. נכון/לא נכון
2. 2% כל מודל בוליאני מורחב יחזיר רק את d1. נכון /לא נכון
3. 2% המודל הווקטורי לא תומך בשאילתות כמו q1. נכון/ לא נכון

ב. 8% נתונה שאילתא q1 : C,B,A
נתון המאגר הבא: המספרים מסמלים תדירות של ה Terms במסמך.
מסמך d1 הכולל את ה terms A (3),B (5),C (10)
מסמך d2 הכולל את ,B (1),C(5)
מסמך d3 הכולל את A(30)
מסמך d4 הכולל את B(2), C(5)
מסמך d5 הכולל את B(2), C(5)
מסמך d6 שלא כולל אף אחד מה terms הנ"ל

סמן נכון/לא נכון ליד כל משפט, בהנחה שמשתמשים בשיטת שקלול של tf*idf

1. 2% ייתכן שמסמך d6 יוחזר לשאילתא q1 אם יופעל אלגוריתם lsi והמסמך קשור לנושא השאילתא. נכון / לא נכון
2. 2%, המודל הווקטורי שמשתמש ב inner-product ידרג את d1 במקום הראשון. נכון/לא נכון
3. 2% הוספה של ה A Term למסמך d2 תשפיע רק על רמת הדמיון בין d2 לשאילתא. נכון/לא נכון
4. 2% הוספה של ה A term למסמך d4 תשפיע בין היתר על רמת הדמיון בין מסמך d1 לשאילתא. נכון/לא נכון

6. 26% נתונה הרשת הבאה:



א. 10% מיהו הצומת/צמתים בעל ה pagerank הגבוה ביותר, הצומת בעל ה pagerank הנמוך ביותר

ב. 8% מיהו הצומת בעל ה hub הגבוה ביותר וה authority הנמוך ביותר.

ג. 8% מה יקרה ל pagerank של צומת 1 אם נבטל את הקישור בין צומת 5 לצומת 4.

7. 10% כדי לעשות pagerank מותאם למשתמש אפשר להתאים את הפרמטר d בנוסחא בהתאם לתחומי העניין של המשתמש, כלומר לאתרים שהם בתחום העניין של המשתמש אפשר להגדיל את ההסתברות של המשתמש להגיע אליהם,

א. 4% מה הבעיה היישומית של pagerank מותאם אישית?

ב. 6% הצע פתרון יישומי (באופן כללי, אין צורך לכתוב אלגוריתם או נוסחאות, רק הסבר).

בהצלחה
ברכה ואיליה