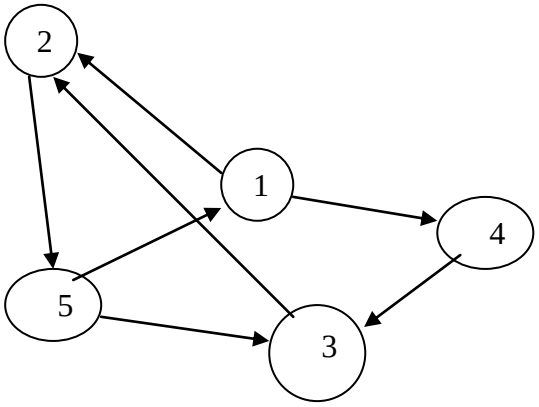
בס"ד

איחזור מידע תש"ע – 372.1.4406
סמסטר חורף מועד א' 31.01.10
ד"ר ברכה שפירא, איליה פרידמן

משך המבחן : שעתיים וחצי
חומר עזר: מותר (**לא מחשב נייד**)
יש להחזיר את השאלון – נא לענות על שאלות 5-8 על השאלון

1. 25%



א. (5%) אם נוסיף קישור מצומת 2 לצומת 4, האם וכיצד יושפעו ערכי HUB ו – AUTHORITY (על פי אלגוריתם HITS) של צומת 5 (כלומר , האם הערכים יהיו גבוהים, או נמוכים יותר, או ללא שינוי ומדוע) .

ב. (5%) הראה את מטריצת המעברים (לצורך חישוב pagerank) של הרשת מסעיף א (לאחר ההוספה של הקישור מסעיף א).

ג. (5%) חשב שתי איטרציות של pagerank של הגרף שבציור לאחר ההוספה מסעיף א d=0.8. (ללא נרמול) (לא בשיטת המטריצות). אין צורך לחלק את (1-d) במספר הצמתים ברשת. את האיטרציה השנייה אין צורך לחשב מספרית, אלא רק להציב ערכים מתאימים בנוסחה.

ד. (5%) האם אפשר להעריך בגף הנתון בשאלה (לאחר ההוספה ב -א') איזה צומת תהיה בעלת ערך pagerank גבוה ביותר לאחר ההתכנסות. אם כן הערך והסבר את הערכתך, או הסבר מדוע אי אפשר להעריך.

ה. (5%) האם אלגוריתמים של link-analysis כדוגמת pagerank מודדים גם את פופולאריות הדף אצל הגולשים. אם כן, כיצד? ואם לא איך אפשר למדוד פופולאריות של דפים אצל גולשים?

2. 8% - רשומה בקובץ posting הופכי כוללת נתונים על פי terms. על כל term הקובץ כולל את רשימת כל המסמכים שבהם הוא מופיע, ואינפורמציה נוספת (למשל, מיקום המופעים של ה term במסמכים). רשימת המסמכים לכל term בדרך כלל ממוינת באחת משתי הדרכים הבאות:

   א. מיון על פי זיהוי המסמך.
   ב. מיון על פי שכיחות (חשיבות, משקל), המילים במסמכים, כלומר מסמכים שבהן ה-term יותר משמעותי יופיעו ברשימה לפני מסמכים שבהם ה term פחות משמעותי.

   ציין יתרון אחד לכל אחת מהשיטות.

3. 10% מטה-מנועים ממזגים תוצאות של כמה מנועי חיפוש לתוצאה ממוינת אחת שמוצגת למשתמש.
א.5% מיזוג תוצאות נחשבת לבעיה שמצריכה אלגוריתם מורכב. הסבר ממה נובע הקושי במיזוג התוצאות.
ב. 5% אם כל מנועי החיפוש שמשתתפים במטה-מנוע ישתמשו באותו אלגוריתם דרוג. האם הקושי שתיארת בסעיף א יימנע? הסבר מדוע כן או לא.

4. 18% הנח שאתה בונה מנוע המבוסס על המודל הווקטורי. בנוסף הדרישה היא שפונקציית הדרוג תתבסס על שלושה סוגי מידע: 1) הטקסט שבגוף המסמך, 2) כותרת המסמך, 3) טקסט של קישורים (anchor text).
   א. 8% הסבר כיצד אפשר לשלב את המודל הווקטורי כך ששלושת סוגי המידע יבואו לידי ביטוי בחישוב הדרוג של המסמך וחישוב הדמיון יהיה על סמך התאמה בין ווקטור השאילתא לווקטור המסמך. סוגי המידע יכולים להשפיע באופן שונה על הדירוג, (יתכן שסוג מידע אחד יהיה משמעותי יותר מאחר).
   ב. 10% על פי תשובתך ב-א, תאר פונקציית דירוג (תיאור בנוסחה) על פי הכללים הבאים. (אפשר לתאר פונקציה נפרדת לכל אחד מהכללים) :
      - אם מילת חיפוש מופיעה בכותרת, לא מתחשבים במספר ההופעות שלה בגוף הטקסט או בטקסט של קישורים.
      - אם מילת חיפוש לא מופיעה בכותרת, אז מתחשבים בתדירות הופעתה במסמך ובטקסט של קישורים, כאשר טקסט של קישורים משפיע פחות מטקסט בגוף המסמך.

5. 10% מנוע E1 החזיר 10 תוצאות נכונות לשאילתא Q1. מתוך 20 תוצאות שהחזיר. סה"כ יש במאגר 30 תוצאות מתאימות לשאילתא. ובכלל במאגר יש 1000 מסמכים.
הגדילו את המאגר ל10,000 מסמכים. למאגר לא התווספו עוד תוצאות רלוונטיות לשאילתא Q1 (כלומר, נשארו 30 תוצאות מתאימות). הריצו מנוע אחר E2 על המאגר החדש והוא החזיר בדיוק את אותם 20 מסמכים שהחזיר המנוע הראשון על המאגר הקטן יותר: סמן "נכון" או "לא נכון" על כל אחד מהמשפטים הבאים:

   א. 2% ה precision ו ה recall של שני המנועים לשאילתא Q1 זהים. נכון/ לא נכון
   ב. 2% כדאי לחשב precision באלף מסמכים כדי לבטא את ההבדל באיכות המנועים. נכון/ לא נכון.
   ג. 2% Precision מנורמל למספר המסמכים במאגר יבטא את ההבדל בין איכות המנועים. נכון/ לא נכון.
   ד. 2% יתכן שמנוע E1 אם יופעל על המאגר הגדול (10,000) על שאילתא Q1 יחזיר 15 תוצאות רלוונטיות. נכון/לא נכון
   ה. 2% יתכן שמנוע E2 , אם יופעל על המאגר הקטן (1000) עם שאילתא Q1 יחזיר 15 תוצאות רלוונטיות. נכון/לא נכון.

6. 5% משתמש במערכת סינון תוכנית קיבל כהמלצה 3 מסמכים: d1,d2,d3, המשתמש כתגובה טען שמסמך d1 לא רלוונטי, ואילי d2 ו-d3 רלוונטים מאוד. המערכת מעדכנת את פרופיל המשתמש בהתאמה לתגובותיו . אפשר להסיק ש: (בחר תשובה נכונה אחת).

   א. D2 ו-d3 דומים ביניהם .
   ב. המסמכים הבאים שיחזרו למשתמש יהיו דומים יותר ל d2 מאשר ל d3.
   ג. הפרופיל של המשתמש אחרי העדכון יהיה דומה יותר ל d2 ו-d3 מאשר לפני העדכון.
   ד. הפרופיל של המשתמש לפני העדכון דומה יותר ל – d1 מאשר ל d2 ו-d3.

7. 14% לשאילתא מסוימת התקבלו הערכים הבאים בגרף P-R (עם אינטרפולציה) כאשר המנוע החזיר 20 מסמכים וישנם סך הכול במאגר 10 מסמכים רלוונטיים לשאילתא. להלן הערכים בגרף.

Recall	Precision
0	1
0.1	1
0.2	0.66
0.3	0.6
0.4	0.5
0.5	0.5
0.6	0.3
0.7	0
0.8	0
0.9	0
1	0

9% מה היה ה precision לאחר שהמערכת החזירה 5 מסמכים, 8 מסמכים ו -20 מסמכים,
5% מה ה precision לאחר שהמערכת החזירה 5 מסמכים **רלוונטיים.**

8. 10% בטבלה הבאה נתונים דירוגים של משתמשים על מוצרים שרכשו (1-5). לצורך ניבוי מוצר למשתמש בשיטה השיתופית לוקחים בחשבון רק שכנים שהדמיון שלהם למשתמש הוא גדול או שווה ל- 0.5 .
לצורך ניבוי ההתאמה של מוצר 5 למשתמש 3 , השכנים שיילקחו בחשבון הם (רק תשובה אחת נכונה):

	משתמש 1	משתמש 2	משתמש 3	משתמש 4	משתמש 5
מוצר 1	5	2	5		4
מוצר 2				5	
מוצר 3				5	
מוצר 4	3	5	2		3
מוצר 5		5		5	5

א. אין שכנים מתאימים
ב. משתמשים 2 ו-5
ג. משתמשים 1–ו- 5
ד. משתמש 5
ה. אף תשובה לא נכונה

בהצלחה

ברכה ואיליה