בס"ד

מספר נבחן 18/8

איחזור מידע תשע"ד – 372.1.4406

סמסטר חורף מועד ב' 16.02.2014

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. המבחן כולל 3 דפים. יש לענות על כל השאלות

שאלה 1 - 16% - יש לענות על הגיליון

1.1 2% SVD מניח אי תלות בין ה terms נכון/לא נכון

1.2 2% המודל הווקטורי מצטיין במנשק מובן למשתמש הקצה נכון/לא נכון

1.3 2% כדי לאפשר דינמיות של אינדקס ניתן להחזיק אינדקסים קטנים שיתעדכנו באופן שוטף ויתמזגו עם האינדקס המרכזי פעם בתקופה כדי למנוע עדכונים תכופים מדי של האינדקס המרכזי. נכון/לא נכון

1.4 כדי להגיע לתוצאות טובות יותר מבחינת הfreshness של דפים עדיף לcrawler לנקוט בגישה פרופורציאנלית לתכיפות עדכון הדפים מאשר בגישה אחידה נכון/לא נכון

1.5 2% אחד ההבדלים בין מערכת המלצה למנוע חיפוש הוא שמערכת המלצה ממליצה על רלוונטיות של מוצרים (או מסמכים) למשתמש לעומת מנוע חיפוש שקובע את הרלוונטיות של המסמכים למשתמש נכון/לא נכון

1.6 4% נתונים שני מנועי חיפוש s1, s2 שונים המופעלים על שני מאגרים נפרדים, שחלק מהמסמכים בהם משותפים. שאילתא נשלחת לשני המנועים, כל מנוע מחזיר רשימה מדורגת של מסמכים. כדי למזג את הרשימה המוחזרת לרשימה מדורגת אחת מאוחדת מתוצאות שני המנועים מספיק ש:

א. יודעים שהאלגוריתמים של המנועים שווים,
ב. המנועים יחזירו את רמת הדמיון של כל מסמך לשאילתא
ג. את כמות המסמכים המשותפת בין המאגרים
ד. א+ג
ה. א+ב
ו. א+ב+ג
ז. אין דרך למזג את התוצאות

1.7 4% אסטרטגיה של crawler כוללת החלטות על :
א. תגובה להוראות בקבצי robots.txt
ב. תעדוף ובחירה של אתרים שונים בתור
ג. שיטת אינדוקס דפים שה crawler מוריד
ד. שיטת הביזור של הדפים ב repository שהcrawler מכין
ה. תכיפות עדכון של דפים באינדקס
ו. א+ב+ד
ז. א+ב+ה

2. 17% קיימים שני מסמכים d1, d2. ב-d1 2500 מילים ואילו ב-d2 3000 מילים. נתונה השאילתה "who shot Lincoln". בטבלה המופיעה למטה מוצגות מספר ההופעות של חלק מהמילים במסמכים:

	D1	D2
who	4	7
theatre	3	15
show	6	9
Lincoln	1	3
play	2	2
shot	8	1

א) 6% חשב את דירוגי המסמכים ביחס לשאילתה, תוך שימוש במודל השפה הבסיסי שהוצג בכיתה,

ב) 6% חשב את דירוגי המסמכים ביחס לשאילתה ,תוך שימוש ב-KL divergence, , קבע מי המסמך הטוב יותר על פי שיטה זו.

ג) 5% הסבר מהי החלקה וציין שתי סיבות לשימוש בה בעת דירוג מסמכים באמצעות מודלי שפה

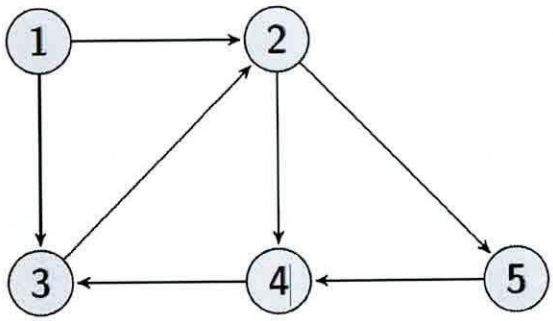
3. 12%

א. 4% מה יהיו ערכי ה hub ו authority של רשת שבה כל הקשתות דו-צדדיות , כלומר :

$$(u,v) \in E \Leftrightarrow (v,u) \in E$$

הצדק את תשובתך.

ב. 4% הצג את מטריצת המעברים עבור אלגוריתם Page rank לרשת הבאה:



ג. 4% מיהם שני הצמתים בעלי ה pagerank הנמוך ביותר (הסבר , ללא חישוב!)

4. 13% נתונה הטבלה הבאה של דירוגי משתמשים למוצרים שונים (items)

	User1	User2	User3	User4	User5
Item1	5		4	4	4
Item2		5	5	3	1
Item3	2		1		2
Item4		1		3	
Item5	5				
Item6		5	4		

א. 6% רוצים להמליץ למשתמש 5 על item מס 6 בגישת Collaborative Filtering באלגוריתם המתחשב בדמיון בין משתמשים - אילו משתמשים ישתתפו בחישוב? מה הניבוי של הציון שיתן משתמש 5 לitem מספר 6?

ב. 7% רוצים להמליץ על למשתמש 2 על item מס 3, אם משתמשים בגישת הדמיון בין המוצרים? הסבר אילו דירוגים של אילו מוצרים משתתפים בחישוב.

5. 20 % מפתח של מנוע חיפוש החליט שבכל ריצה של שאילתא הוא ירחיב את השאילתא עם מילים נוספות מתוך וויקיפדיה, שקשורות למילות השאילתא .

א. 15% הצע שיטה למציאת מילים קשורות למילות השאילתא בוויקיפדיה, במפורט תאר את האלגוריתם למציאת מילים קשורות והערכת מידת הקשר של מילה מסויימת בוויקיפדיה למילות השאילתא.

ב. 5% אילו מבני נתונים נוספים (מעבר לאינדקס ההופכי ולפוסטינג של המאגר של המנוע) כדאי למנוע להחזיק כדי לממש את הרעיון הנ"ל.

6. 10% אלגוריתם דרוג של מנוע חיפוש נותן עדיפות למילים שהופיעו בכותרת, כלומר, אם כל המילים של השאילתא מופיעות בכותרת, המנוע מחליט שהמסמך הוא רלוונטי, ואז מחשב את רמת הרלוונטיות שלו, אחרת המסמך לא רלוונטי. תאר מבנה של אינדקס (מילון ופוסטינג) שייתן פתרון יעיל למנוע כזה.

7. 12% נתון מאגר המסמכים הבא:

	milk	pepper	raisins	sugar	cinnamon	apples	flour	eggs	clove	jelly
d_1	4	0	0	4	0	1	1	0	0	0
d_2	1	1	0	2	0	0	0	0	1	0
d_3	3	1	0	2	0	0	0	2	0	0
d_4	1	2	1	1	2	0	2	1	0	0
d_5	2	0	2	0	1	0	5	2	1	2
d_6	1	0	0	0	0	0	1	1	0	2
d_7	2	1	0	0	1	0	0	0	0	1
d_8	0	0	3	2	0	1	0	4	0	0

א. 6% נתונה השאילתא (cinnamon, clove), חשב את הדרוג של המסמכים שיוחזרו על פי המודל הווקטורי ונוסחת הקוסינוס (יש לנרמל לפי ה term השכיח במסמך)

ב. 6% נניח שהמנוע החזיר את המסמכים d2,d4,d7 והמשתמש סימן את d2, d4 כרלוונטים ואת d7 כלא רלוונטי.

חשב את השאילתא החדשה שתיוצר על פי אלגוריתם Rocchio אם α=2, β=1, γ=0 (יש להראות את הווקטור החדש של השאילתא)

בהצלחה! ברכה אביגיל ויוסי