מרצה: דרור מוגהץ

תאריך הבחינה:

**מס' תלמיד:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**מכון:** לב

משך הבחינה (בדקות): **90**

חומר עזר מותר לשימוש: **אסור**

מחשבון: **מותר**

המבחן כולל סה"כ **3** חלקים, יש לענות על כל **3** החלקים, ובכל חלק על כל השאלות.

הציון המקסימאלי הוא 100

פירוט ניקוד:

**חלק א': 61 נקודות, 3 שאלות פתוחות.**

**חלק ב': 4 שאלות רב ברירתיות, כאשר בכל שאלה רק תשובה אחת נכונה (יש להקיף את התשובה הנכונה ביותר), לכל שאלה 6 נקודות: סה"כ 24 נקודות.**

**חלק ג': 5 טיעונים, על כל טיעון יש לסמן נכון/לא-נכון (להקיף את התשובה הנכונה ביותר). לכל טיעון 3 נקודות: סה"כ 15 נקודות**.

**תלמיד יקר,**

1. **אם אינך מבין את כוונת המרצה בשאלה כלשהי, עליך לכתוב בראש התשובה כיצד הינך מבין את השאלה ולפתור בהתאם. לשיקול דעתו של המרצה אם יש מקום להבנה זו ואז ינקד בהתאם.**
2. **חובה להחזיר את השאלון אחרת הבחינה לא תיבדק!**
3. **נוהל הבחינות של המכון מחייב אותך ובאחריותך לקוראו ולהכירו - בחינה עלולה להיפסל על כל חריגה מהנוהל.**
4. **לידיעתך, תורדנה נקודות לא רק על שגיאות, אלא גם לתוספות לא רלוונטיות, העדר נימוק הולם לתשובה, חוסר סדר ותשובה דו-משמעית, כאשר נדרשת תשובה חד משמעית.**

***בהצלחה רבה !***

**יש לענות על כל השאלות בטופס הבחינה עצמו!**

**חלק א' (61 נקודות)**

שאלה 1

1. חשבו את ה- IDF של המילים news ו- money (נניח שהחישוב שאנו נעשה הוא על בסיס 2) לפי הנוסחה שלמדנו בהרצאה (IDF הוא חלק מהנוסחה שעליה הוספנו את BM25 ונרמול באורך המשפט), חשוב לדייק, לחשב את ה- IDF בלבד. את החישוב של ה- IDF תעשו על שלושת המסמכים המופיעים בטבלה המופיעה באמצע העמוד.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **news** | about | presidential | campaign | **money** |
| IDF(w) | **1** | 1 | 2.5 | 3.1 | **2** |

הראו את החישובים

לא הבנתי לפי מה דרור הגיע לערכי הidf של שאר המילים, אשמח אם מישהו יבין

ה- IDF של המילה news הוא m=3

= Df(news)2

Log(4\2)=1

ה- IDF של המילה money הוא m=3

=Df(money)1

Log(4\1)=2

כעת, תנו ציונים לפי נוסחת TF-IDF **הפשוטה** (ללא שום נרמולים וכדומה) לכל אחד מהמסמכים הבאים ודרגו את המסמכים (1, 2, 3) הבאים בהנתן השאילתה: presidential campaign news

|  |  |  |  |
| --- | --- | --- | --- |
| דירוג DC | דירוג TF-IDF | מסמך | תוכן המסמך |
| 3 | 2.5+3.1=5.6 | d1 | Putin influence Trump presidential campaign |
| 2 | 1+2.5+3.1=6.6 | d2 | … the news about the presidential campaign … |
| 1 | 4\*(3.1)+1=13.4 | d3 | … the food campaign news, to continue the campaign we gave the campaign money,  the campaign … |

1. נלמד כעת נוסחה חדשה לדירוג מסמכים, הנוסחה נקראת Dice coefficient.

הגדרת Dice coefficient

דרגו את המסמכים הנ"ל על אותה השאילתה.

הראו את החישובים

|d1| =5

|d2|=6

|d3|=15

|q|=3

D1

2(presidential campaign)\8

**1\2**

D2

2(presidential campaign news)\9

**2\3**

D3

2(campaign, news)\18

**2\9**

**הדירוג הסופי הוא: d2-d1-d3**

שאלה 2

1. בהנתן לנו אינדוקס מיקומים, על אילו שאילתות נוכל לענות?

תשובה:

אינדקס מיקומים נותן לנו אפשרות לדעת את המיקומים של המילים ואז גם להבין ולקשר בין ביטויים וזוגות מילים.

עכשיו אנחנו נוכל לבצע שאילתות של ביטויים\זוגות מילים (אפשר גם לשחק עם האורך של המילים ולבקש שכנויות לא בהכרח קרובות ממש וכו)

1. האם "התשלום" על אינדוקס מיקומים נמוך, בינוני או משמעותי? האם משתמשים באינדוקס כזה?

תשובה:

מגדיל את כמות הזיכרון הנצרך באופן משמעותי

למרות זאת, שימוש באינדוקס מיקום הפך להיות סטנדרטי בגלל העוצמה והתועלת של שאילתות ביטוי וקרבה

שאלה 3

בהנתן הטבלה הבאה נא למלא אותה בהתאם [Variable byte (VB) encoding]

|  |  |  |
| --- | --- | --- |
| docIDs | 824 | 829 |
| gaps | 824 | 5 |
| VB code | 0111000**1** 0000110 | 1010000**1** |

הראו את החישובים

VB-code

824

110 0111000

0111000**1** 0000110

829

המרחק 5

101

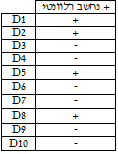
101000**1**

**חלק ב' (24 נק'), שאלות רב בררתיות, לכל שאלה - רק תשובה אחת נכונה.**

1. חוקר הציע ליצור מאגר של אוסף בדיקות IR הוא טען שאח"כ יריצו על האוספים האלו הרבה פעמים את אותן שאילתות

מתלבטת בין א ל ד. א נכון לפי השיטה השנייה.

1. טענת חוקר עמית 1: מיותר כי כל פעם צריך להזין שאילתות שונות המותאמות למסמכים
2. טענת חוקר עמית 2: לא ניתן ליצור אוסף כזה שמייצג את האינטרנט ולכן על ההצעה להדחות
3. אין כאן תשובה נכונה
4. כי כך ניתן יהיה לבדוק מספר אלגוריתמים שונים ל- IR
5. מהי תוצאת ה- DCG (Discounted Cumulative Gain) שלמדנו בכיתה (כאשר מדובר על בסיס 2) בהנתן הנתונים הבאים:
6. 6.13
7. 7
8. 2.807
9. אין כאן תשובה נכונה
10. ה- Mean Average Precision (MAP) של הנתונים הבאים הוא:



1. 16
2. 0.31
3. 0.775
4. 16
5. אין כאן תשובה נכונה

**חלק ג (15 נק'), רשימת טיעונים. עבור כל טיעון – סמן: האם הטיעון נכון, או לא נכון.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | הטיעון |  |  |
| 1 | ~~הרעיון של משוב Rocchio מתבסס על מבט של השאילתה כעל ווקטור~~ | נכון | לא נכון |
| 2 | שימוש ב- "קליקים" על קישורים של תוצאות מנוע החיפוש אחד הכלים החדישים והכי מדוייקים בהערכת איכות תוצאות מנוע חיפוש בדכ ילחצו על הראשונים יותר והתוצאות מוטות | נכון | לא נכון |
| 3 | למדנו מספר פונקציות להערכת תוצאות של ניסויים. שתי פונקציות חשובות הן Accuracy ו- F1. ישנם מצבים בהם הפונקציה Accuracy מייצגת נאמנה את התוצאות במידה דומה לפונקציה F1 accuacy בדרך כלל לא אמין ויתן לנו 99 אחוזים כי הרוב אצלינו negative לעומת F1 שהוא שילוב של recall precision | נכון | לא נכון |
| 4 | בעוד השאילתות שהמשתמש מזין הן חד משמעיות המידע במסמכים הוא, לעיתים, רב משמעי הפוך-השאילתות הם רב משמעותיות | נכון | לא נכון |
| 5 | הפעולה הראשונה שמתבצעת לאחר הזנת השאילתה של המשתמש לחיפוש באינטרנט היא זחילה (crawling) לא מצאתי בחומר את המושג הזה | נכון | לא נכון |

**בהצלחה!!!**