מרצה: דרור מוגהץ

תאריך הבחינה: יום שישי כ"ט תמוז תשפ"א (09/7/2021)

**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.**

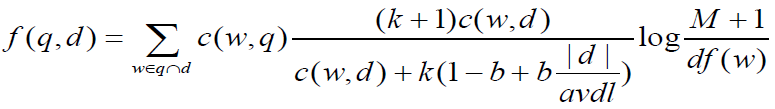
***בהצלחה רבה !***

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

**חלק א'**

שאלה 1

בהנתן פונקצית הדירוג הבאה:



1. הסבירו את הפונקציה ואת מרכיבי הפונקציה.

תשובה:

הפונקציה הזאת היא TF-IDF

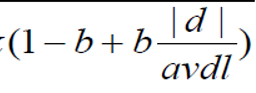
נסביר כל חלק בנפרד:

**k**- קבוע שנותן לנו חסם עליון ותחתון לערכים.

זה בעצם טרנספורמציית BM25

**b-** קבוע שאומר כמה נעניש את המסמך בגלל אורכו**.**

נרמול אורך המסמך-



d

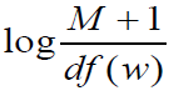
הוא המסמך שלנו, נחלק את האורך שלו(מספר המילים בו) באורך הממוצע של מסמך כדי לתת חשיבות לאורך המסמך.

**TF**-term frequency

כמה פעמים הופיע המילה במסמך: c(w,d)

כמה פעמים הופיעה המילה בשאילתה: c(w,q)

**IDF**- inverse document frequency



M—מספר המסמכים סהכ

Df(w)- מספר המופעים של המילה במסמך d

הרעיון האינטואיטיבי ב- IDF הוא "להעניש" מונחים/מילים בעלות שכיחות כללית גבוהה

)באנגלית למשל המילה the

סוף תשובה

1. צרו **Inverted Index table** למאגר נתונים בהנחה כי המסמכים הבאים מהווים את מאגר הנתונים כך שהטבלה תכיל את כל הנתונים של מנת שנוכל להשתמש בפונקציה שהסברתם בסעיף הקודם

|  |  |
| --- | --- |
| D1 | ראשונים הוא כינוי בתולדות עם ישראל לרבנים שפעלו בין המאה ה-11 והמאה ה-15 לערך, במיוחד בהקשר של השתלשלות ההלכה ופרשנות התלמוד. |
| D2 | חיבורנו על הראשונים מקיף תקופה של ... ועד לגירוש ספרד , המאורע הנחשב כסיומה של תקופת הראשונים . חילקנו את הספר לפרקים לפי הזמנים והארצות . והחשיבות של ידיעת המקומות פעולתם של הראשונים גדולה כמעט כידעת הזמנים |
| D3 | בתקופת הראשונים אין למסגרת הספרותית של השו"ת אותה השפעה מרכזית שהייתה לה בימי הגאונים. בתקופת הראשונים נקבעו הדפוסים של השו"ת כפסק דין מנומק מבורר ומפורט |

תשובה:

נדגים על כמה מילים: (דרור ציין שאם תופיע שאלה דומה במבחן יתנו מספר מילים עליהם נתבקש לעשות את הטבלה)

|  |  |  |
| --- | --- | --- |
| **Doc list** | **freq** | **term** |
| D1->d2->d3 | 6 | ראשונים |
| D1 | 1 | כינוי |
| D2->d3 | 4 | תקופה |
|  |  |  |
|  |  |  |

שאלה 2

בהנתן מטריצת שכנויות הבאה:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| DOC | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
| WORD |
| Antony | 1 | 1 | 0 | 0 | 0 | 1 |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 |
| Cleopatra | 1 | 0 | 0 | 0 | 0 | 0 |
| mercy | 1 | 0 | 1 | 1 | 1 | 1 |
| worser | 1 | 0 | 1 | 1 | 1 | 0 |

מהי התשובה לשאילתה: ***Brutus AND Caesar)*** AND !***(Calpurnia and Brutus))?*** ציינו את דרך

תשובה:

Brutus- 110100

Caesar-110111

Calpurina-01000

*Brutus AND Caesar*- 110100

*Calpurnia and Brutus*- 010000

*Brutus AND Caesar)* AND*Calpurnia and Brutus* – 010000

!-not – 101111

**תשובה סופית: 10111**

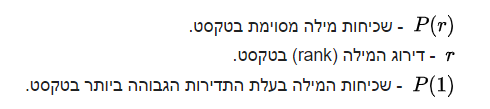
שאלה 3

מה אומר לנו חוק זיפ (Zipf’s Law)?

תשובה:

חוק זה נותן לכל מילה ציון\רמה ביחס לתדירות הופעתה בקורפוס





**חלק ב', שאלות רב ברירתיות, לכל שאלה - רק תשובה אחת נכונה.**

1. הקוד האונרי (Unary code) את המספר 5:
2. **11110**
3. 101
4. 11111
5. אין כאן תשובה נכונה
6. חסרון של התייחסות למסמך כ- BOW בהקשר של חיפוש
7. **ב- BOW ההתייחסות כאילו אין קשר בין המילים**
8. BOW לא תומך בביטויים באורך 3 מילים
9. BOW מתעדף מילים שכיחות
10. כל התשובות נכונות
11. במודל pull:
12. היוזמה בידי המשתמש
13. היוזמה בידי מערכת המחשב
14. היוזמה היא שילוב של המשתמש והמערכת
15. אין כאן תשובה נכונה
16. ב- **Q**uery **L**ikelihood
17. בהינתן מסמך מסויים (שאותו המשתמש צריך) מה ההסתברות שהמשתמש יזין שאילתה מסויימת.
18. מנוע החיפוש "מנחש" מה השאילתה שהמשתמש יזין
19. מנוע החיפוש משלים אוטומאטית את השאילתה של המשתמש
20. אין כאן תשובה נכונה

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | הטיעון |  |  |
| 1 | כל מילה בשאילתה חייבת להופיע במסמכים המאוחזרים | נכון | **לא נכון** |
| 2 | קישורים המופיעים באתרי אינטרנט יכולים לעזור לנו בשיוך אתר לנושא ספציפי- לא למדנו | נכון | לא נכון |
| 3 | מתודולוגיית Cranfield משנה את השאילתה של המשתמש | נכון | **לא נכון** |
| 4 | הרעיון של IDF זה "להעניש" מילים המופיעות בהרבה מאוד מסמכים | נכון | לא נכון |
| 5 | מכפלה פנימית היא אחת מפונקציות הדירוג הטובות ביותר כיום | נכון | **לא נכון** |

**בהצלחה!!!**