**Information retrieval HW 2**

**Group: Result Wizards**

Nevo Gottlieb 206234353

Raz Kessel 206017725

Lior Zucker 316375187

Yoni Azeraf 209459239

<https://result-wizards.onrender.com/>  
<https://github.com/liorZucker11/Result-Wizards.git>

שאלה 1:

מערכת יש מחיפוש 5 מסמכים רלוונטיים ו-10 מסמכים לא רלוונטיים.

ידוע כי באוסף ישנם עוד מסמכים, מתוכם 20 מסמכים רלוונטיים.

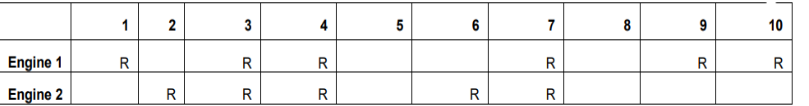
מהו precision ומה recall?

**חישוב Precision:**

**חישוב Recall:**

שאלה 2:

1. נתונה טבלת תוצאות של שני מנועי חיפוש:



בהנחה שבקורפוס יש 500 מסמכים, 10 מהם רלוונטיים לשאילתה הנתונה.

הציגו טבלה עם Precision,recall לאחר כל חזרה.

הציגו טבלה וגרף של 11 נקודות אינטרפולציה.

חשבו f-measure עם β = 0.75, איזה מנוע טוב יותר?

מנוע 1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Recall | Precision | סה״כ רלוונטיים | סה״כ הוחזרו | מסמך |
|  |  | 1 | 1 | 1 |
|  |  | 1 | 2 | 2 |
|  |  | 2 | 3 | 3 |
|  |  | 3 | 4 | 4 |
|  |  | 3 | 5 | 5 |
|  |  | 3 | 6 | 6 |
|  |  | 4 | 7 | 7 |
|  |  | 4 | 8 | 8 |
|  |  | 5 | 9 | 9 |
|  |  | 6 | 10 | 10 |

מנוע 2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Recall | Precision | סה״כ רלוונטיים | סה״כ הוחזרו | מסמך |
|  |  | 0 | 1 | 1 |
|  |  | 1 | 2 | 2 |
|  |  | 2 | 3 | 3 |
|  |  | 3 | 4 | 4 |
|  |  | 3 | 5 | 5 |
|  |  | 4 | 6 | 6 |
|  |  | 5 | 7 | 7 |
|  |  | 5 | 8 | 8 |
|  |  | 5 | 9 | 9 |
|  |  | 5 | 10 | 10 |

**כעת נציג טבלה וגרף של 11 נקודות אינטרפולציה:**

**ניצור נקודות אינטרפולציה ע״י בחירת 11 נקודות Recall באופן קבוע: [0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0]**

**באינטרפולציה לוקחים את ערכי ה Precision הגבוהים ביותר עבור כל נקודת Recall קבועה.**

מנוע 1:

|  |  |
| --- | --- |
| Interpolated Percision | Recall |
| 1.0 | 0.0 |
| 1.0 | 0.1 |
| 0.75 | 0.2 |
| 0.75 | 0.3 |
| 60. | 0.4 |
| 0.6 | 0.5 |
| 0.6 | 0.6 |
| 0 | 0.7 |
| 0 | 0.8 |
| 0 | 0.9 |
| 0 | 1.0 |

מנוע 2:

|  |  |
| --- | --- |
| Interpolated Percision | Recall |
| 0.75 | 0.0 |
| 0.75 | 0.1 |
| 0.75 | 0.2 |
| 0.75 | 0.3 |
| 0.714 | 0.4 |
| 0.714 | 0.5 |
| 0 | 0.6 |
| 0 | 0.7 |
| 0 | 0.8 |
| 0 | 0.9 |
| 0 | 1.0 |

A graph with blue and orange lines

Description automatically generated

נחשב את ה f-measure עם ערך β = 0.75:

על מנת לקבוע איזה מנוע טוב יותר נחשב את הממוצע של ערך ה f-measure בכל אחד המנועים ונבחר במנוע עם הממוצע הגבוה יותר:

עבור מנוע 1 :

Avg precision:

Avg recall:

f-measure:

עבור מנוע 2 :

Avg precision:

Avg recall:

f-measure:

מאחר וה- f-measure של מנוע 2 גבוה יותר נוכל לקבוע גם כי הוא המנוע הטוב יותר.

ב. חוק ZIPF - נתונים המסמכים הבאים:

D1: "My dogs love music a lot, and often listen to the Rolling Stones"

D2: "Information Retrieval course"

D3: "The dog can roll. He loves rolling and throwing stones"

D4: "They also often help me pick up stones from the road"

מילים אלו הן stop words:

a, also, and, are, from, he, how, lot, me, my, of, often, some, the, they, to, very

א. בנו את ה-inverted index תוך השמטת s, ing, ed, ly.

שימו לב - בנו inverted index עבור כלל terms (החליטו מהם ה -terms וכיצד לשמור אותם).

ג. בנו טבלת תדירות לכלל המילים. האם ההתפלגות מתנהגת לפי חוק zipf?

1. inverted index :

|  |  |
| --- | --- |
| Frequency | Term |
| D1,D3 | Dog |
| D1,D3 | Love |
| D1 | Music |
| D1 | Listen |
| D1,D3 | Roll |
| D1,D3,D4 | Stone |
| D2 | Information |
| D2 | Retrieval |
| D2 | Course |
| D3 | Can |
| D3 | Throw |
| D4 | Help |
| D4 | Pick |
| D4 | Up |
| D4 | Road |

inverted index עבור כל term:

|  |  |
| --- | --- |
| Frequency | Term |
| D1 | Dogs |
| D1 | Love |
| D1 | Music |
| D1 | Listen |
| D1,D3 | Rolling |
| D1,D3,D4 | Stones |
| D2 | Information |
| D2 | Retrieval |
| D2 | Course |
| D3 | Dog |
| D3 | Can |
| D3 | Roll |
| D3 | Loves |
| D3 | Throwing |
| D4 | Help |
| D4 | Pick |
| D4 | up |
| D4 | Road |

1. טבלת תדירות לinverted index:

|  |  |
| --- | --- |
| Frequency | Term |
| 3 | Stone |
| 3 | Roll |
| 2 | Dog |
| 2 | Love |
| 1 | Music |
| 1 | Listen |
| 1 | Information |
| 1 | Retrieval |
| 1 | Course |
| 1 | Can |
| 1 | Throw |
| 1 | Help |
| 1 | Pick |
| 1 | Up |
| 1 | Road |

טבלת תדירות לinverted index עבור כל term :

|  |  |
| --- | --- |
| Frequency | Term |
| 3 | Stones |
| 2 | Rolling |
| 1 | Help |
| 1 | Pick |
| 1 | up |
| 1 | Road |
| 1 | Dog |
| 1 | Can |
| 1 | Roll |
| 1 | Loves |
| 1 | Throwing |
| 1 | Information |
| 1 | Retrieval |
| 1 | Course |
| 1 | Dogs |
| 1 | Love |
| 1 | Music |
| 1 | Listen |

נראה שההתפלגויות אינן עומדות באופן מלא בחוק זיפ, מאחר שהקווים אינם יוצרים דפוס לינארי רציף בלוג-לוג. בשני המקרים יש סטייה חדה שמפרה את חוק זיפ ברמות הנמוכות.

שאלה 3: קדם פרויקט –בניית זחלן

הריצו את החיפוש והחזירו את התוצאות בקובץ אקסל, עם כל השדות הרלוונטיים.

כעת, עליכם לבנות אינדקס לתוצאות שחזרו.

1. בנו inverted index ל-15 מילים הנפוצות ביותר שחזרו. במידה והמילים הנפוצות מופיעות ביותר מ-20 דפים, התייחסו רק ל- 20 הדפים הראשונים שחזרו.
2. בחרו את אחת השאילתות בה עשיתם שימוש, וחשבו tf-idf של המושגים בשאילתא, על פני כל הדפים שחזרו. פרטו את כל שלבי החישוב.
3. הציגו את 10 התוצאות הראשונות שחזרו עבור השאילתה ל-2 משתמשים שונים. בקשו מכל משתמש להחליט עבור כל תוצאה האם היא רלוונטית. חשבו precision ו-recall.

הנחיות הגשה:

יש להגיש קישור לריפו שלכם ב -GIT, הכולל את תיקייה בשם HW2 שבה קובץ וורד ובו מענה על השאלות.

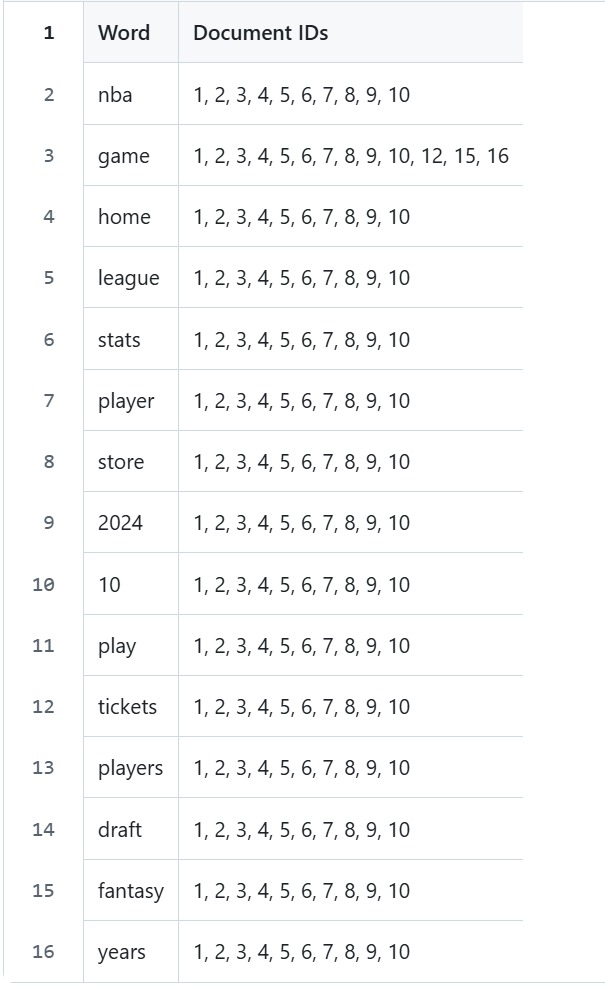
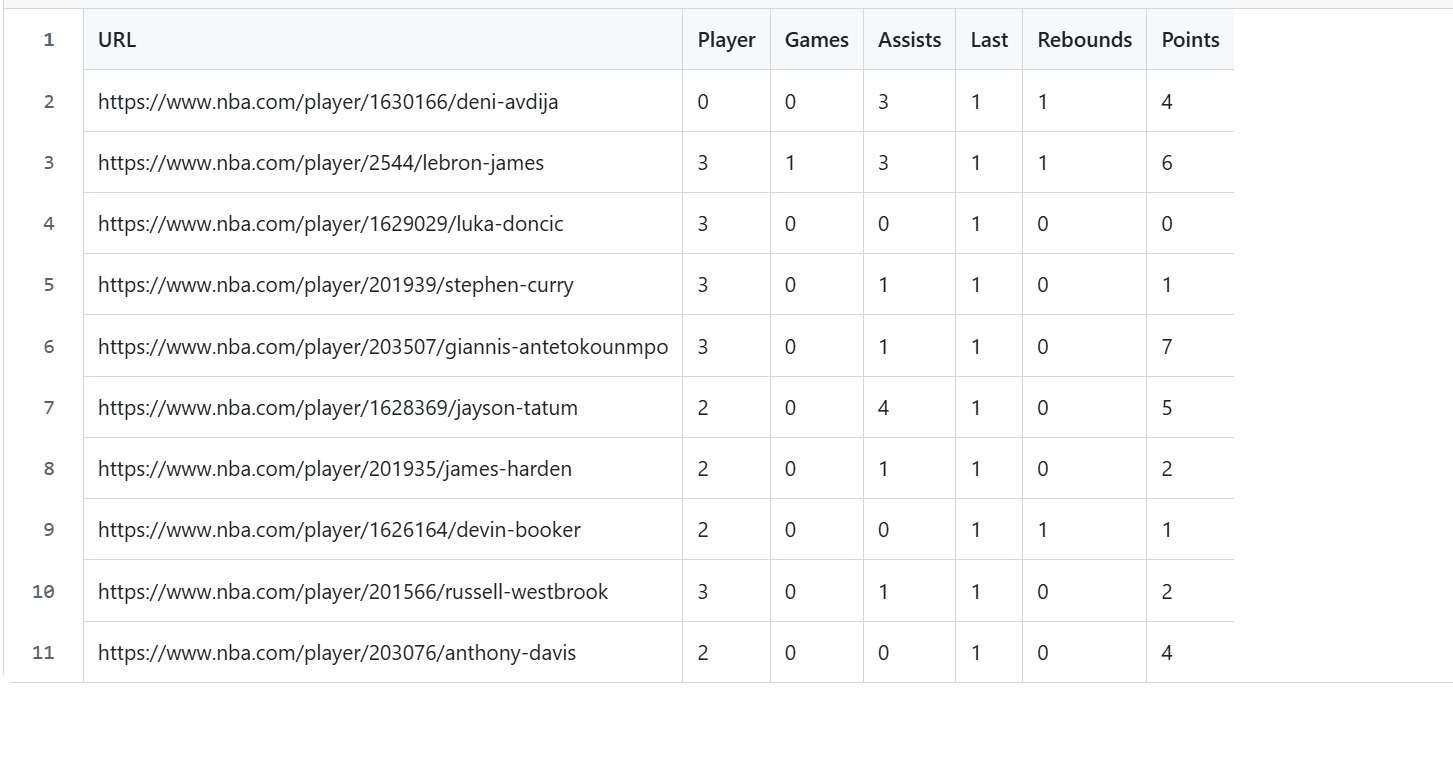
כמו כן התיקייה תכיל כל קבצי הזחלן. את החישובים והתוצאות יש לבצע בקובץ excel

ולשמור בתיקייה זו.

ניתן להעלות שאלות הקשורות לתרגיל בפורום הקורס.

קישור לריפו:<https://github.com/liorZucker11/Result-Wizards.git>

קישור לדפלוי: <https://result-wizards.onrender.com/>

1. 
2. 

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IDF | D/Dfi | Dfi | P10 | P9 | P8 | P7 | P6 | P5 | P4 | P3 | P2 | P1 | Word |
|  | *10/9* | 9 | 2 | 3 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 0 | Player |
|  | 10/1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | Games |
|  | 10/7 | 7 | 0 | 1 | 0 | 1 | 4 | 1 | 1 | 0 | 3 | 3 | Assists |
|  | 10/10 | 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Last |
|  | 10/3 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | Rebounds |
|  | 10/9 | 9 | 4 | 2 | 1 | 2 | 5 | 7 | 1 | 0 | 6 | 4 | Points |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TF-10 | TF-9 | TF-8 | TF-7 | TF-6 | TF-5 | TF-4 | TF-3 | TF-2 | TF-1 | Word |
| 1.301 | 1.301 | 1.477 | 1.301 | 1.301 | 1.301 | 1.477 | 1.477 | 1.477 | 0 | Player |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | Games |
| 0 | 1 | 0 | 1 | 1.602 | 1 | 1 | 0 | 1.477 | 1.477 | Assists |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Last |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | Rebounds |
| 1.602 | 1.301 | 1 | 1.301 | 1.698 | 1.845 | 1 | 1.778 | 1.778 | 1.602 | Points |

תמונה שמכילה טקסט, צילום מסך, שחור, מקלדת

התיאור נוצר באופן אוטומטי



התוצאות שחזרו מהשאילתה:

נקבע כי page1 , page 2, page 5 , page 6, page 8   
הינם דפים רלוונטיים עבור השאילתה (לצורך חישוב precision & recall)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Recall | Precision | Page 10 | Page 9 | Page 8 | Page 7 | Page 6 | Page 5 | Page 4 | Page 3 | Page 2 | Page 1 | User |
| 5/5=1 | 7/10 |  |  | R | R | R | R |  | R | R | R | a |
| 4/5 | 5/10 | R |  |  |  | R | R |  |  | R | R | b |