אחזור מידע – תרגיל בית 2 – חורף התשפ"ה - Crawl Commandos

מגישים:  
צחי בקל 315730176

דניאל ארמגניאן 209146943  
יונתן שרר 318317682  
ליאור ז'ילגו 316109115

**קישור לגיט:** <https://github.com/audiblemaple/Data-acquisition-course>  
  
שאלה 1

**חישוב Precision**Precision = (relevant documents retrieved)/(retrieved documents) = (5/15) = 0.333 = 33.3%

**חישוב recall**Recall = (relevant documents retrieved)/(total relevant documents) = (5/25) = 0.2=20%

שאלה 2  
סעיף א'  
מהטבלה הנתונה בשאלה נבנה טבלת מונים של מסמכים רלוונטים שהחזיר בכל רגע מנוע החיפוש:

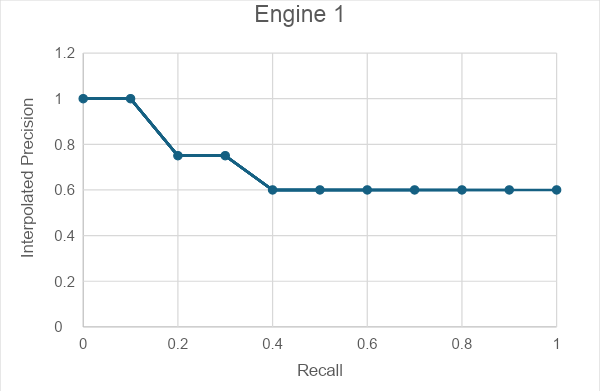
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | **9** | **8** | **7** | **6** | **5** | **4** | **3** | **2** | **1** |  |
| 6 | 5 | 4 | 4 | 3 | 3 | 3 | 2 | 1 | 1 | **Engine 1** |
| 5 | 5 | 5 | 5 | 4 | 3 | 3 | 2 | 1 | 0 | **Engine 2** |

ניעזר בטבלה למעלה כדי לחשב Precision ו Recall כנדרש:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | **9** | **8** | **7** | **6** | **5** | **4** | **3** | **2** | **1** |  |  |
| 6/10=0.6 | 5/9 | 4/8=0.5 | 4/7 | 3/6=0.5 | 3/5=0.6 | 3/4=0.75 | 2/3 | 1/2 | 1/1=1 | **Engine 1** | **Precision** |
| 6/10=0.6 | 5/10=0.5 | 4/10=0.4 | 4/10=0.4 | 3/10=0.3 | 3/10=0.3 | 3/10=0.3 | 2/10=0.2 | 1/10=0.1 | 1/10=0.1 | **Recall** |
| 5/10=0.5 | 5/9 | 5/8 | 5/7 | 4/6 | 3/5=0.6 | 3/4=0.75 | 2/3 | 1/2=0.5 | 0/1=0 | **Engine 2** | **Precision** |
| 5/10=0.5 | 5/10=0.5 | 5/10=0.5 | 5/10=0.5 | 4/10=0.4 | 3/10=0.3 | 3/10=0.3 | 2/10=0.2 | 1/10=0.1 | 0/10=0 | **Recall** |

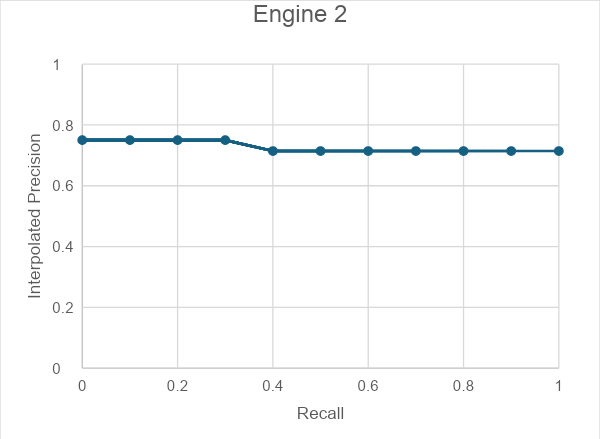
אינטרפולציה ב11 נקודות מלקת את recall לחתום בין 0 ל1 ל11 נקודות קבועות. בכל נקודה יש לקחת את הערך המקסימלי של precision שנמצא עבור ערך recall השווה או גדול לנקודה זו.  
  
  
  
טבלה וגרף של אינטרפולציה עבור Engine 1:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Recall | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| Interpolated Precision | 1 | 1 | 0.75 | 0.75 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 |

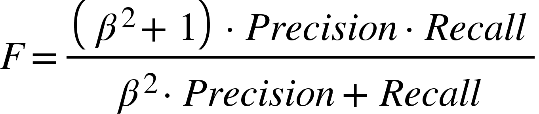


טבלה וגרף של אינטרפולציה עבור Engine 2:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Recall | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| Interpolated Precision | 0.75 | 0.75 | 0.75 | 0.75 | 0.714 | 0.714 | 0.714 | 0.714 | 0.714 | 0.714 | 0.714 |



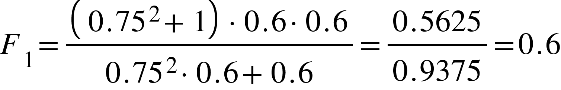
נוסחת חישוב f-measure:



ניקח את הprecision- וה-recall בהחזרה האחרונה בכל מנוע חיפוש, ובנוסף נתון כי **{"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>&#x3B2;</mi><mo>=</mo><mn>0</mn><mo>.</mo><mn>75</mn></mstyle></math>","origin":"MathType for Microsoft Add-in"}**.

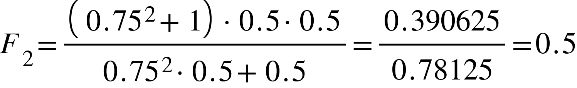
f-measure עבור מנוע 1:

Precision = 0.6 ,Recall = 0.6



f-measure עבור מנוע 2:

Precision = 0.5, Recall = 0.5



לסיכום, ניתן לראות כי מנוע חיפוש 1 טוב יותר ממנוע חיפוש 2.

סעיף ב' חוק ZIPF

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"

תחילה, נבנה טבלה של המילים בכל מסמך לאחר הורדת stop words והורדת סיומות מילה הנתונות:

|  |  |
| --- | --- |
| dog love music listen roll stone | D1 |
| information retrieval course | D2 |
| dog can roll love roll throw stone | D3 |
| help pick up stone road | D4 |

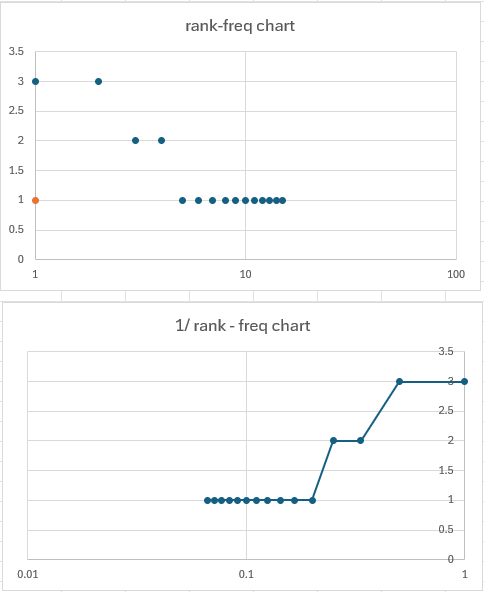
נבנה טבלה המרכזת את הופעות המילים בכל מסמך

|  |  |
| --- | --- |
| 1, 3 | dog |
| 1, 3 | love |
| 1 | music |
| 1 | listen |
| 1, 3 | roll |
| 1, 3, 4 | stone |
| 2 | information |
| 2 | retrieval |
| 2 | course |
| 3 | can |
| 3 | throw |
| 4 | help |
| 4 | pick |
| 4 | up |
| 4 | road |

נבנה טבלה המרכזת את סך התדירויות של כל מונח בארבעת המסמכים

|  |  |
| --- | --- |
| תדירות | מילה |
| D1:1 + D3:2 = 3 | roll |
| D1:1 + D3:1 + D4:1 = 3 | stone |
| D1:1 + D3:1 = 2 | dog |
| D1:1 + D3:1 = 2 | love |
| D1:1 = 1 | music |
| D1:1 = 1 | listen |
| D2:1 = 1 | information |
| D2:1 = 1 | retrieval |
| D2:1 = 1 | course |
| D3:1 = 1 | can |
| D3:1 = 1 | throw |
| D4:1 = 1 | help |
| D4:1 = 1 | pick |
| D4:1 = 1 | up |
| D4:1 = 1 | road |

חוק Zipf מתייחס בעיקר לקורפוסים גדולים ולכן אם נשים את הנתונים בגרפים, עבור קורפוס גדול נצפה לקבל קו ישר, במקרה שלנו קשה לראות עקב העובדה שהקורפוס שלנו הוא קטן.



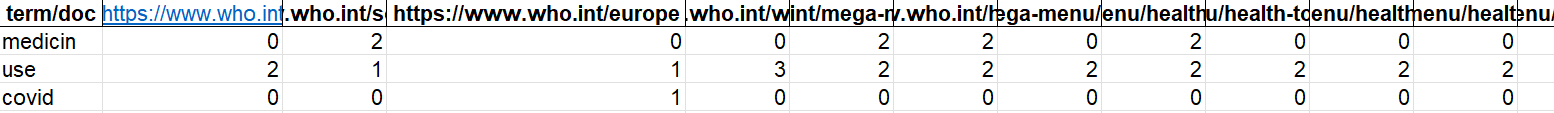
שאלה 3: קדם פרויקט –בניית זחלן

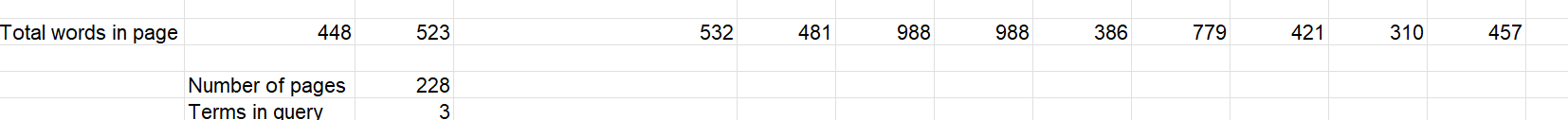
א. נמצא בקובץ inverted\_index. לאחר בניית האינדקס על כל המסמכים שחזרו מהשאילתות, לקחנו את 15 המילים הנפוצות ביותר ועשינו להם inverted index, לכל מילה לקחנו את מקסימום 20 המסמכים הראשונים שבהם היא מופיעה כנדרש.

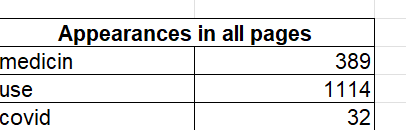
ב. בחרנו בשאילתא "which medicine is used for covid 19".

לאחר הורדת stop words ופעולת stemming קיבלנו את השאילתא המצומצמת: "medicin use covid"

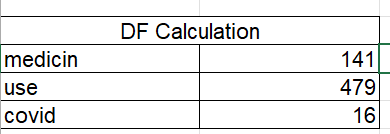
את החישובים אנו מבצעים בקובץ **calculation**.  
תחילה נבנה טבלה בת שלוש שורות כאשר כל שורה היא מילה לאחר הצמצום ועמודה זה עמוד שנסרק ע"י הזחלן. בכל תא מופיע כמות ההופעות של המילה בדף.



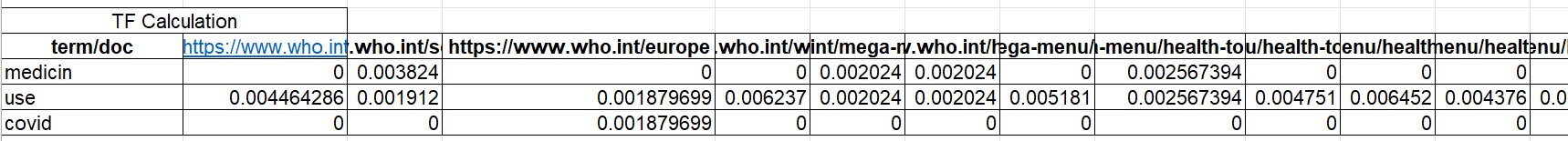
לאחר מכן חישבנו בעזרת SUM בקובץ term\_doc\_appearance\_all\_terms על כל עמודה את כמות המילים בסה"כ המופיעות בדף.  
כמות הדפים בסה"כ חישבנו ע"י מספר העמודות.  


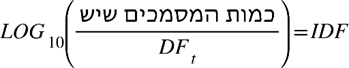
בשלב הבא חישבנו את כמות המופעים של כל אחד משלושת המילים בכל הדפים, גם כאן בעזרת SUM על כל שורה של כל אחד מהמילים  


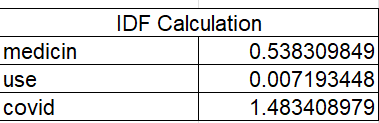
נחשב DF - מספר המסמכים שבהם מופיע כל מונח בעזרת פקודת COUNTIF



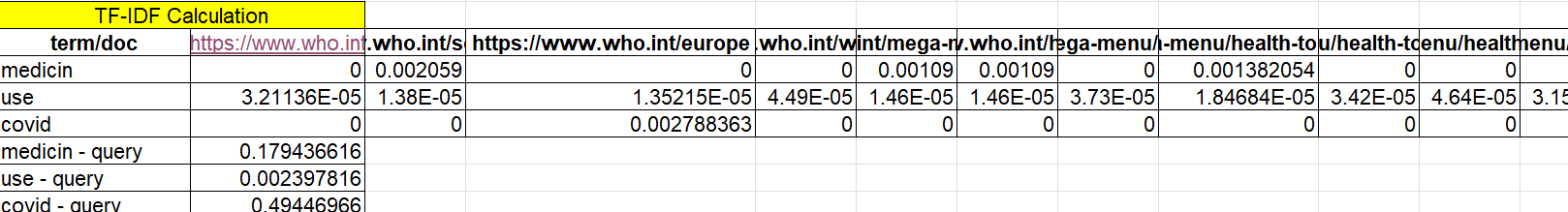
נחשב TF על ידי הנוסחא: 



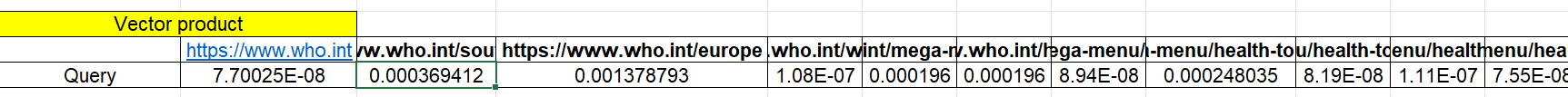
נחשב IDF על ידי הנוסחא: 

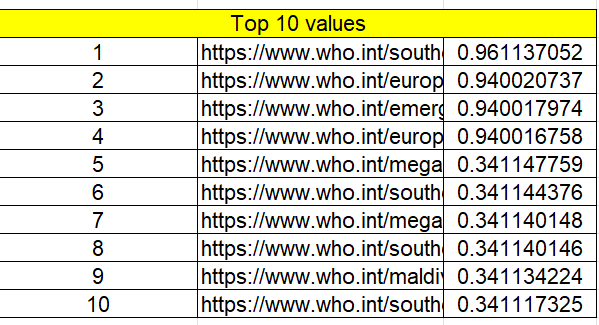


נחשב tf-idf ע"י הנוסחא:  לכל דף וגם של השאילתא עצמה



3.

לצורך החזרת 10 דפים רלוונטים נחשב Cosine similarity. תחילה נחשב מכפלה סקלרית בין וקטורי השאילתא וקטור המסמך:  


לאחר מכן חישבנו את גדלי הוקטורים והצבנו בנוסחא:  
  
  
10 הערכים הגדולים ביותר מהווים את הדמיון הגדול ביותר בין הוקטורים.  


שאלנו שני אנשים האם 10 התוצאות הראשונות רלוונטיות לשאלה:

1. סטודנט לרפואה באוניברסיטת חיפה שנה חמישית.
2. שכנה של דניאל שקוראת חדשות רפואה מידי יום.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |  |
| NR | R | NR | NR | R | R | R | R | R | R | J1 |
| R | R | R | NR | R | NR | R | NR | R | R | J2 |

כעת נחשב את ה-Recall וה-Precision עבור כל אחד מהנשאלים.

כאשר מחפשים את המונח covid 19 באתר WHO מקבלים 20 דפים של תוצאות כאשר בכל דף 20 כתבות שונות, כלומר בסה"כ 400 כתבות הקשורות לcovid 19. נניח כי סקרנו כ10% מתוכם, כלומר 40 כתבות העונות על על הקריטריונים.

|  |  |  |
| --- | --- | --- |
| Recall | Precision |  |
| 7/40 | 7/10 | J1 |
| 7/40 | 7/10 | J2 |