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

מגישים:
צחי בקל 315730176
דניאל ארמגניאן 209146943
יונתן שרר 318317682
ליאור ז'ילגו 316109115

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:

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

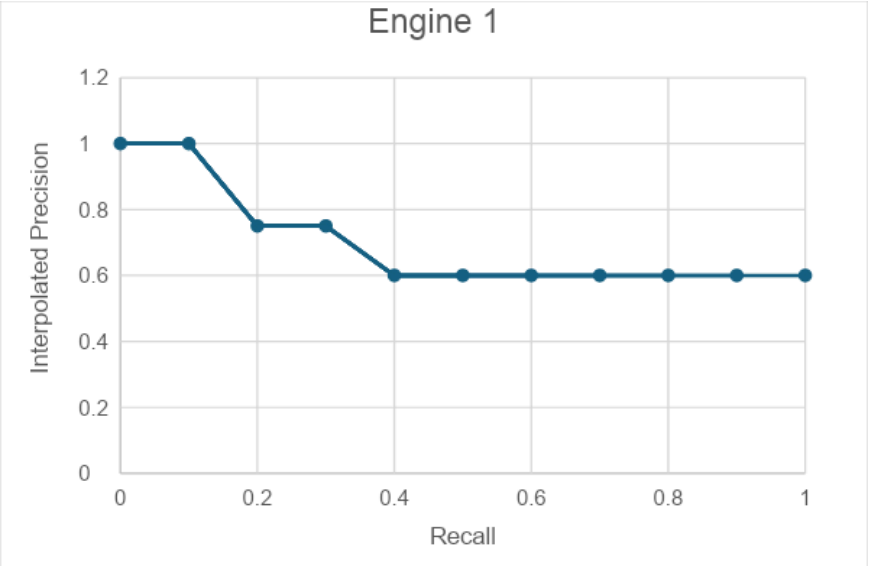
ניעזר בטבלה למעלה כדי לחשב Precision ו Recall כנדרש:

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

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.

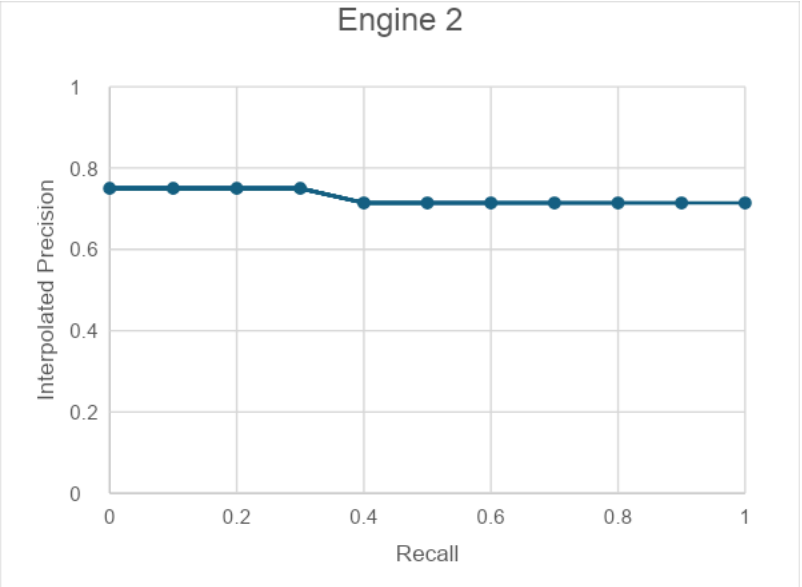
טבלה וגרף של אינטרפולציה עבור Engine 1:

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:

$$F=\frac{\left(\beta^2+1\right)\cdot Precision\cdot Recall}{\beta^2\cdot Precision+Recall}$$

ניקח את הprecision- וה-recall בהחזרה האחרונה בכל מנוע חיפוש, ובנוסף נתון כי $\beta=0.75$.

<u>f-measure עבור מנוע 1</u>:

Precision = 0.6 ,Recall = 0.6

$$F_1=\frac{\left(0.75^2+1\right)\cdot 0.6\cdot 0.6}{0.75^2\cdot 0.6+0.6}=\frac{0.5625}{0.9375}=0.6$$

<u>f-measure עבור מנוע 2</u>:

Precision = 0.5, Recall = 0.5

$$F_2=\frac{\left(0.75^2+1\right)\cdot 0.5\cdot 0.5}{0.75^2\cdot 0.5+0.5}=\frac{0.390625}{0.78125}=0.5$$

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

<u>סעיף ב' חוק ZIPF</u>

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:

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

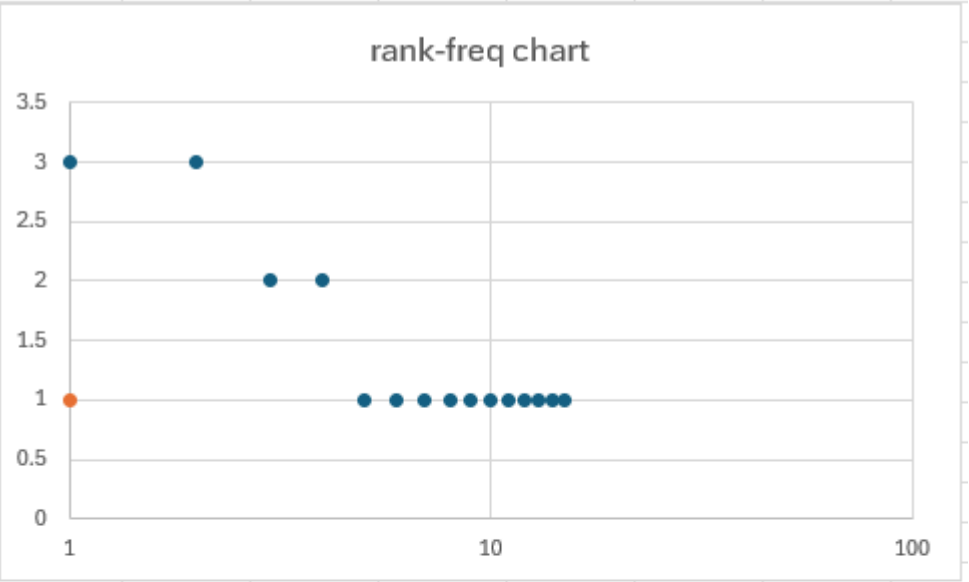
נבנה טבלה המרכזת את הופעות המילים בכל מסמך

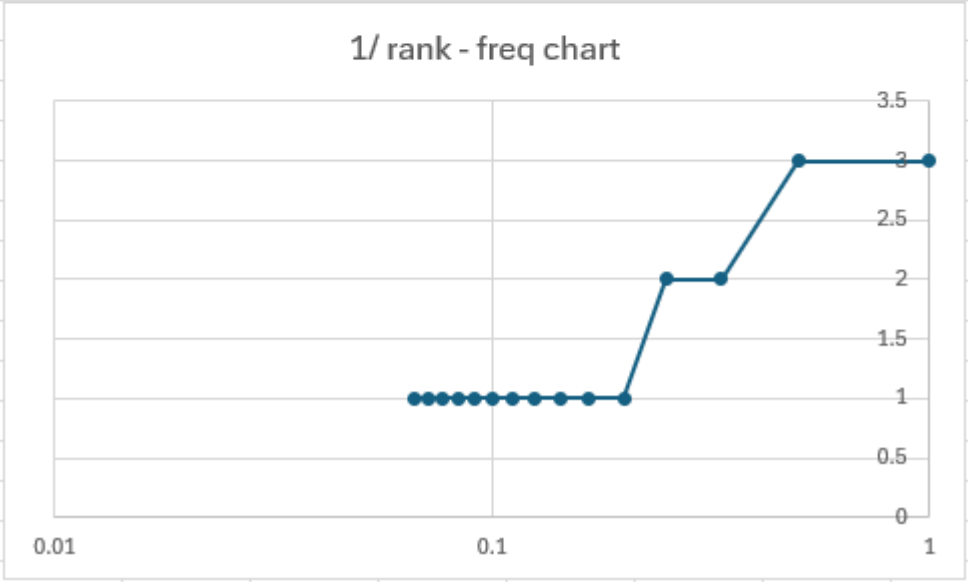
dog	3 ,1
love	3 ,1
music	1
listen	1
roll	3 ,1
stone	4 ,3 ,1
information	2
retrieval	2
course	2
can	3
throw	3
help	4
pick	4
up	4
road	4

נבנה טבלה המרכזת את סך התדירויות של כל מונח בארבעת המסמכים

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

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.

term/doc	https://www.who.int	.who.int/s	https://www.who.int/europe	.who.int/w	int/mega-r	/.who.int/h	ega-menu/	enu/health	u/health-t	enu/health	menu/healt	enu/
medicin	0	2	0	0	2	2	0	2	0	0	0	
use	2	1	1	3	2	2	2	2	2	2	2	
covid	0	0	1	0	0	0	0	0	0	0	0	

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.

Total words in page	448	523	532	481	988	988	386	779	421	310	457
	Number of pages	228									
	Terms in query	3									

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

Appearances in all pages	
medicin	389
use	1114
covid	32

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

DF Calculation	
medicin	141
use	479
covid	16

נחשב TF על ידי הנוסחא: $TF_{t,d} = \frac{\text{מספר ההופעות של מונח במסמך}}{\text{מספר המילים הכולל במסמך}}$

TF Calculation												
term/doc	https://www.who.int	.who.int/s	https://www.who.int/europe	.who.int/w	int/mega-r	/.who.int/h	ega-menu/	-menu/health-to	u/health-t	enu/health	menu/healt	enu/
medicin	0	0.003824	0	0	0.002024	0.002024	0	0.002567394	0	0	0	
use	0.004464286	0.001912	0.001879699	0.006237	0.002024	0.002024	0.005181	0.002567394	0.004751	0.006452	0.004376	0.0
covid	0	0	0.001879699	0	0	0	0	0	0	0	0	

נחשב IDF על ידי הנוסחא:

$$LOG_{10}\left(\frac{\text{כמות המסמכים שיש}}{DF_t}\right)=IDF$$

IDF Calculation	
medicin	0.538309849
use	0.007193448
covid	1.483408979

נחשב tf-idf ע"י הנוסחא: $TF-IDF_{t,d}=TF_{t,d}\cdot IDF_t$ לכל דף וגם של השאילתא עצמה

TF-IDF Calculation											
term/doc	https://www.who.int	.who.int/s	https://www.who.int/europe	.who.int/w	int/mega-r	/.who.int/h	ega-menu/	-menu/health-to	u/health-t	enu/health	menu/
medicin	0	0.002059	0	0	0.00109	0.00109	0	0.001382054	0	0	
use	3.21136E-05	1.38E-05	1.35215E-05	4.49E-05	1.46E-05	1.46E-05	3.73E-05	1.84684E-05	3.42E-05	4.64E-05	3.15
covid	0	0	0.002788363	0	0	0	0	0	0	0	
medicin - query	0.179436616										
use - query	0.002397816										
covid - query	0.49446966										

.3

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

Vector product											
	https://www.who.int	/w.who.int/sou	https://www.who.int/europe	.who.int/w	int/mega-r	/.who.int/h	ega-menu/	-menu/health-to	u/health-t	enu/health	menu/hea
Query	7.70025E-08	0.000369412	0.001378793	1.08E-07	0.000196	0.000196	8.94E-08	0.000248035	8.19E-08	1.11E-07	7.55E-08

לאחר מכן חישבנו את גדלי הוקטורים והצבנו בנוסחא:

$$Cosine\ similarity=\frac{\vec{B}\cdot\vec{A}}{\|\vec{B}\|\cdot\|\vec{A}\|}$$

10 הערכים הגדולים ביותר מהווים את הדמיון הגדול ביותר בין הוקטורים.

Top 10 values		
1	https://www.who.int/south	0.961137052
2	https://www.who.int/europ	0.940020737
3	https://www.who.int/emerg	0.940017974
4	https://www.who.int/europ	0.940016758
5	https://www.who.int/mega	0.341147759
6	https://www.who.int/south	0.341144376
7	https://www.who.int/mega	0.341140148
8	https://www.who.int/south	0.341140146
9	https://www.who.int/maldi	0.341134224
10	https://www.who.int/south	0.341117325

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.

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

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.

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