אחזור מידע – פרויקט – 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:**

האתר של ארגון הבריאות העולמי (WHO - World Health Organization), הוא אתר המתמקד בבריאות הציבור בעולם ומספק מידע על מגוון רב של נושאים בריאותיים.   
ניתן למצוא בו מידע עדכני אודות התפרצויות של מחלות, מדיניות בריאות ויוזמות בריאות עולמיות.   
בנוסף, ניתן לאתר בו נתונים, פרסומים מחקריים והנחיות בריאות.   
מאפשר גישה למידע על מגפות עדכניות, חיפוש מהיר לפי אות ראשונה וקישורים לשותפים לארגון הבריאות העולמי כגון קבוצות מחקר, מימון וגופים האוספים נתונים מרחבי העולם עבורו.

[קישור לאתר](https://www.who.int/)

**שאלה 2:**

**Types of covid-19 treatment.**  
**Needed information:** Approved antiviral medications, supportive therapies, and emerging treatments.

**Guidelines on antibiotic resistance.  
Needed information:** Best practices for antibiotic use, global resistance trends, and alternative treatment options.

**Mental health resources during emergencies.**  
**Needed information:** Available helplines, coping strategies, and professional support services.

**שאלה 3:**

הקוד בגיט.

**שאלה 4:**

בקורס למדנו מגוון טכנולוגיות, מהמעניינות שלמדנו:

* **Porter Stemmer**: למדנו את השימושיות שלו ונעזרנו בו עוד מפרויקטים שלנו בקורס מחשוב ענן.
* **BeautifulSoup**: ספרייה לעיבוד וניתוח תוכן HTML/XML, שבה השתמשנו לשליפת מידע מתוך מבנה ה-HTML של הדפים שסרקנו באתר.
* **pandas**: ספריית Python ששימשה אותנו לניתוח ועיבוד נתונים. באמצעותה בנינו DataFrame מכל המידע שנאסף, ביצענו פעולות עיבוד נתונים, ושמרנו את הנתונים בקובץ CSV. השימוש בה אפשר לנו ליצור סקריפט סריקה מתקדם, לעבד את הנתונים שנאספו, ולהפיק תובנות מעמיקות מהאתר.

**שאלה 5:**

השאילתות רצו במשך 327 שניות**.**זמן הריצה תלוי במגוון דברים. ראשית, במהירות השרת המארח את האתר. שנית, הריצה תלוי בכמות הדפים שרצנו עליהם ועל כמות הדפים שיש בכל עמוד, כלומר ככל שהזחלן יגיע עמוק יותר, כך יגדל הזמן.  
אנו סבורים כי ניתן לשפר את החיפוש ע"י מספר אפשרויות:  
כאשר ישנה שאילתא שמבקשת מידע על מדינה או יבשת מסויימת, אפשר להתחיל את הזחלן מהעמוד בו יש מידע על היבשת ברצויה. לדוגמה בקישור הזה: <https://www.who.int/europe/home?v=welcome>  
יש מידע ספיציפית על מדינות באירופה.  
בנוסף, ניתן לשפר זמן זה ע"י אופטימיזציה של הזחלן, שמירת מידע במטמון ותהליכים שירוצו במקביל.

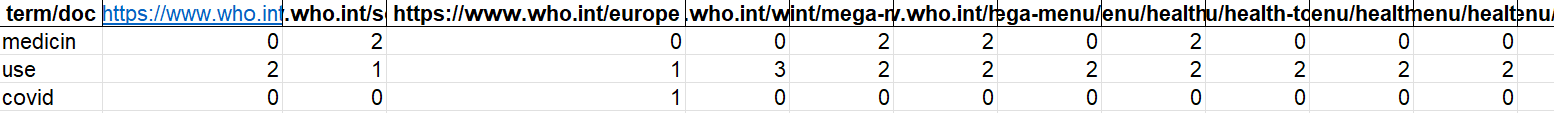
**שאלה 6:**לאחר בניית האינדקס על כל המסמכים שחזרו מהשאילתות, לקחנו את 15 המילים הנפוצות ביותר ועשינו להם inverted index, לכל מילה לקחנו את מקסימום 20 המסמכים הראשונים שבהם היא מופיעה כנדרש. נמצא בקובץ inverted\_index בקישור: <https://github.com/audiblemaple/Data-acquisition-course/tree/main/HW/HW2>

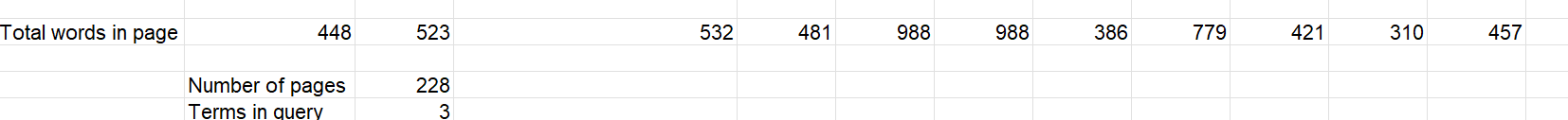
**שאלה 7:**

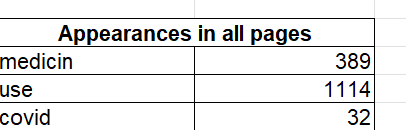
בחרנו בשאילתא "which medicine is used for covid 19".

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

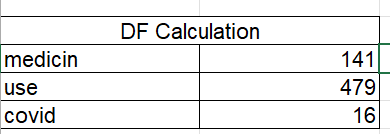
את החישובים אנו מבצעים בקובץ **calculation**.  
תחילה נבנה טבלה בת שלוש שורות כאשר כל שורה היא מילה לאחר הצמצום ועמודה זה עמוד שנסרק ע"י הזחלן. בכל תא מופיע כמות ההופעות של המילה בדף.



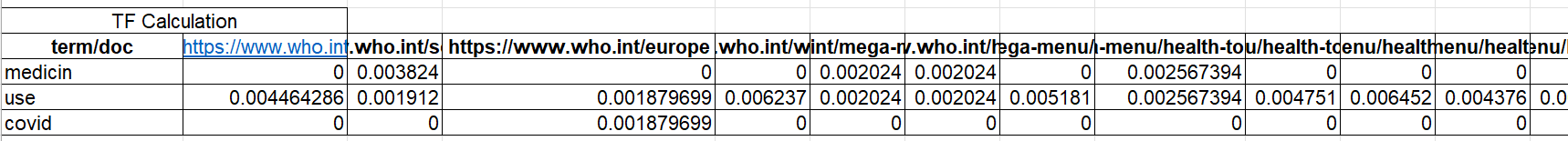
לאחר מכן חישבנו בעזרת SUM בקובץ term\_doc\_appearance\_all\_terms על כל עמודה את כמות המילים בסה"כ המופיעות בדף.  
כמות הדפים בסה"כ חישבנו ע"י מספר העמודות.  


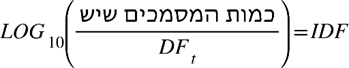
בשלב הבא חישבנו את כמות המופעים של כל אחד משלושת המילים בכל הדפים, גם כאן בעזרת SUM על כל שורה של כל אחד מהמילים  


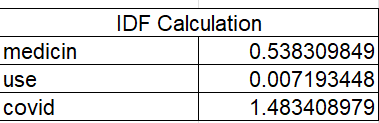
נחשב DF - מספר המסמכים שבהם מופיע כל מונח בעזרת פקודת COUNTIF



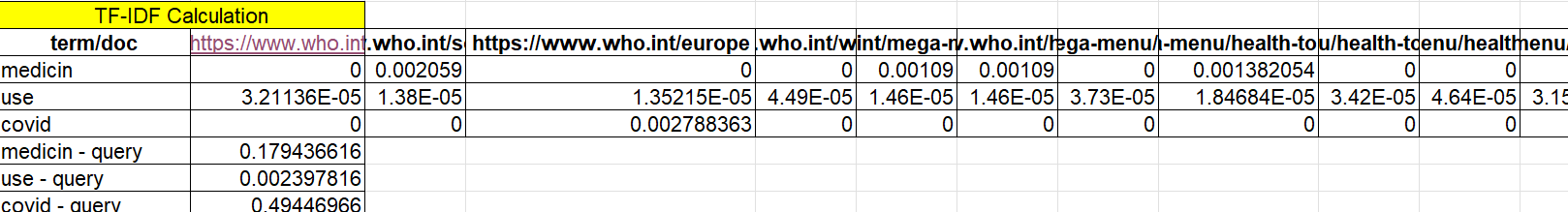
נחשב TF על ידי הנוסחא: 



נחשב IDF על ידי הנוסחא: 



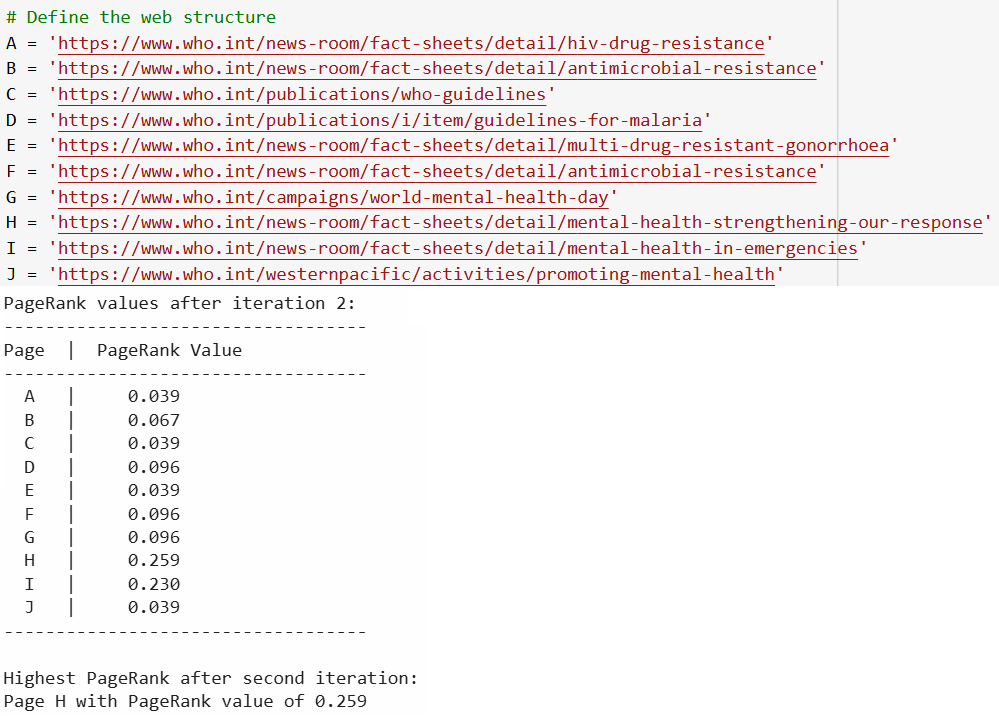
נחשב tf-idf ע"י הנוסחא:  לכל דף וגם של השאילתא עצמה



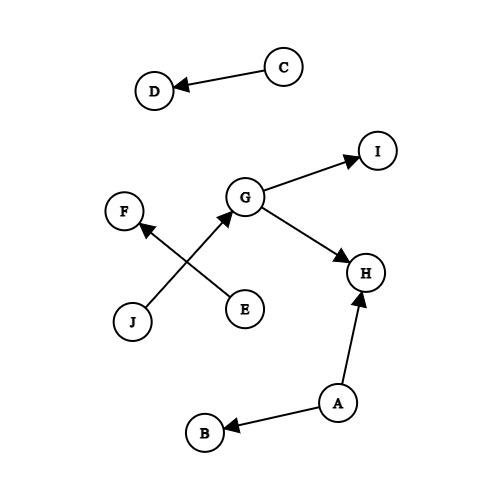
**שאלה 8:**בדפים שהוחזרו קיימיים hubs. ניתן לראות שכל דף של כתבה על מחלה, מחקר וכו' ישנם קישורים המקושרים למחלה או הדברים הקשורים למחקר, לסימטומים קשורים ועוד. משום כך, ניתן לגשת לדפים המקושרים מכל דף של כתבה על מחלה. בנוסף, ניתן לראות כי קיימים גם authorities שהם דפי הקישורים של הכתבות. נראה כי לכל כתבה אפשר לגשת על ידי דפי מחלות או מחקרים אחרים.

**שאלה 9:**לכל שאילתה הוחזרו 20 הדפים הרלוונטיים.בדקנו את הקישוריות בין דפים אלה באמצעות אלגוריתם אשר עובר על הקישורים של כל דף ובודק האם קיימים ביניהם שאר הדפים. אספנו 10 קישורים מתוך התוצאות, השתמשנו באלגוריתם המוצע בקורס ליצירת גרף ולבסוף מחשב את ערך ה-PageRank.

להלן התוצאות:

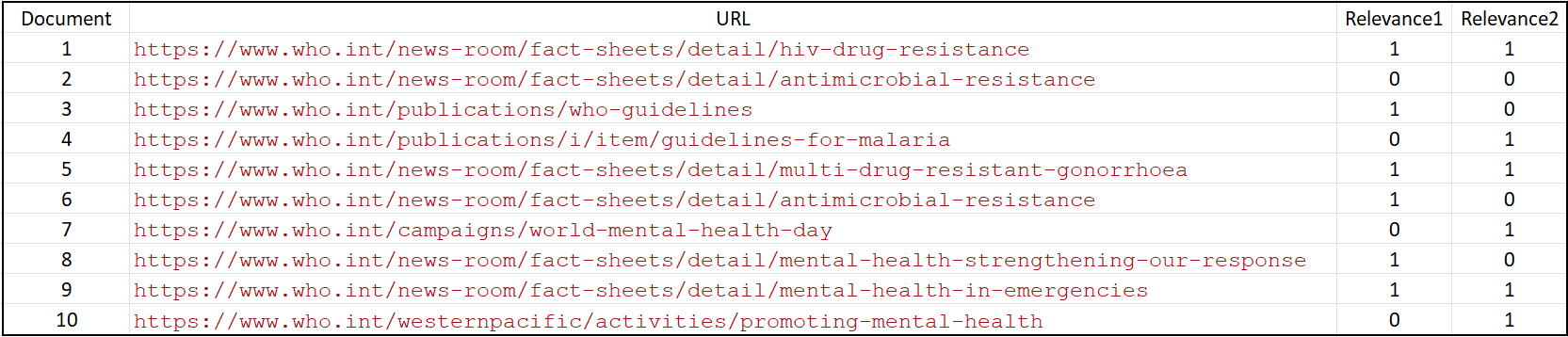


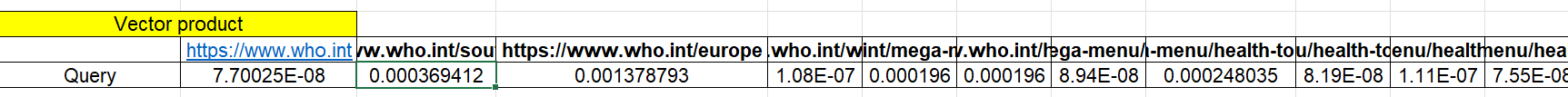
גרף מכוון של הלינקים והקישורים בינהם:



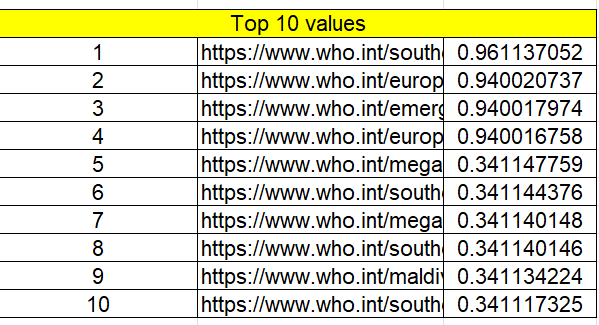
**שאלה 10:**

1. מנתוני הרלוונטיות של המשתמשים עולה כי הם מתעניינים בדפים שבהם מצוי מידע עבור תרופות המונעות מחלות מראש. שאילתה מוצעת עבור משתמשים אלה:  
   Common disease prevention drugs



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:  
תמונה שמכילה שחור, חשיכה

התיאור נוצר באופן אוטומטי  
  
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 |

**שאלה 11:**

<https://audiblemaple.github.io/Data-acquisition-course/>

**שאלה 12:**

|  |  |
| --- | --- |
| איזה שינוי הוצע? | האם לדעתכם יש מקום לשיפור זה? אם כן - הסבירו כיצד ניתן לשפר. אם לא, נמקו |
| להשתמש בקאש | הרבה קבוצות עשו שימוש בקאש באופן שלא בהכרח מצריך זאת. קאש אמור לזרז פעילויות שחוזרות על עצמן, אך התבצע שימוש יחיד ולכן הקאש חסר משמעות. כמו כן, בפרוייקט שלנו הצורך הקאש לא היה מזרז את הזחילה באתר או מציאת תוצאות למיניהם. |
| שימוש בזחילה מקבילית ע"מ להקטין את זמן הריצה | יש משמעות רבה לשימוש במקביליות. ואכן ראינו שקבוצות אחרות שהשתמשו בריצה מקבילית שיפרו רבות את זמן הריצה שלהם. |
| שימוש בגרפים | ויזואליזציה של המידע והממצאים יכולים לעזור בהבנת הפרוייקט שלנו. ואכן, הוספנו גרף באתר שלנו אשר מתאר את הקישורים בין הלינקים בPageRank. |

**שאלה 13:**

עיקר האתגרים עבור חברי הקבוצה היה חוסר זמן עקב עומס בעבודה ובלימודים באופן כללי. בנוסף, לחלק מחברי הקבוצה היה קשה להבין את החומר הנלמד בקורס. בפרט, TF-IDF ,PageRank וזה יצר קושי בעבודה השוטפת שלנו בקורס ובפרויקט.  
עם זאת, הצלחנו לשלב בין הקורס לבין החיים האישיים, התגברנו על הפער ההבנתי בקורס והגענו למטרות הקורס.