**אחזור מידע – פרויקט :**

**שחף ישראל 209288414**

**שי אוסמו 318370285**

**שרה סאלח 318107448**

**יובל סבטו 318474269**

**קישור לגיט:** [**https://github.com/shahaf5641/IR---Project**](https://github.com/shahaf5641/IR---Project)

1. **מהו תחום העיסוק המרכזי של האתר? מהו המידע הזמין למשתמשי האתר? ענו בפסקה אחת. צרפו את הקישור לאתר.**

קישור לאתר: <https://search.cdc.gov/search/?query=>

האתר של ה-CDC (Centers for Disease Control and Prevention) מתמקד במתן מידע ומחקר על בריאות הציבור ומניעת מחלות בארצות הברית ובעולם. תחום העיסוק המרכזי שלו הוא הנגשת מידע עדכני ואמין על מחלות, מגפות, חיסונים, בריאות סביבתית, מניעת פציעות ועוד. האתר מציע למשתמשים נתונים סטטיסטיים, הנחיות בריאותיות, פרסומים מדעיים, כלי חינוך ומידע על תוכניות למניעת מחלות ושיפור הבריאות. הוא פונה לקהל רחב, כולל אנשי מקצוע בתחום הבריאות, חוקרים, מחנכים והציבור הכללי, ומספק גם עדכונים חירומיים על מגפות או אירועי בריאות ציבוריים.

1. **רשמו שלוש שאילתות מעניינות שהייתם רוצים לקבל עליהן תשובה באתר, והאתר אינו עונה עליהן כעת. עבור כל שאילתא כזו, ציינו מהם פרטי המידע הנדרשים לצורך מענה על השאילתא.**

* How is jet lag treated?

פרטי מידע נדרשים:

טיפולים מוכחים לשיפור התאקלמות לאחר טיסות (כגון שינוי תזונה, תוספי תזונה, או תרופות).

הנחיות מותאמות לגילאים שונים (ילדים, מבוגרים, קשישים).

המלצות להתמודדות עם ג'ט לג בנסיעות חוזרות או בקרב עובדים בתפקידים בינלאומיים.

* What are the symptoms of jaundice?

פרטי מידע נדרשים:

רשימת הסימפטומים המרכזיים (כגון גוון עור צהוב, שתן כהה).

מידע על הבדלים בסימפטומים בין קבוצות אוכלוסייה (תינוקות, מבוגרים, נשים בהריון).

גורמי סיכון להחמרת התסמינים ומתי לפנות לטיפול רפואי.

* What causes joint pain?

פרטי מידע נדרשים:

פירוט הגורמים המרכזיים לכאבי מפרקים (דלקות, מחלות אוטואימוניות, פציעות).

קשר בין גיל, פעילות גופנית ומחלות כרוניות לכאבים במפרקים.

דרכי אבחון וטיפול על בסיס סוג הכאב ומיקומו בגוף.

1. **בנו זחלן המחזיר את התוצאות לשאילתות שהגדרתם. מומלץ (אך לא חובה) לבנות את הזחלן בפייתון. הציגו את קוד הזחלן.**

קישור לקוד בגיט:

<https://github.com/shahaf5641/IR---Project/blob/8ef85fee334eb219ee0af48dd59dd1e9b296bfcc/HW2/crawler%20tar3.py>

1. **רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.**

בפרויקט זה נעשה שימוש במגוון טכנולוגיות מעניינות שמדגימות את השימוש בכמה כלי עבודה מתקדמים באוטומציה, עיבוד נתונים, ותהליכי ניתוח טקסט:

**טכנולוגיות שנעשה בהן שימוש:**

* **Selenium:**

ספרייה פופולרית לאוטומציה של דפדפנים לצורך גישה לאתרי אינטרנט וביצוע פעולות כמו מילוי טפסים, ניווט, ולחיצה על כפתורים.

בפרויקט נעשה שימוש ב-Selenium כדי לבצע חיפוש באתר ה-CDC ולנווט בין דפי תוצאות.

* **BeautifulSoup:**

ספרייה לניתוח מבנה HTML ולהפקת נתונים מאתרי אינטרנט.

בפרויקט נעשה שימוש כדי לאסוף תוצאות חיפוש (כותרות וקישורים) ולנתח אותן.

* **Pandas:**

ספרייה לעיבוד וניתוח נתונים בצורה יעילה באמצעות טבלאות.

שימשה ליצירת קובץ Excel עם התוצאות, הדירוג, ומאגר האינדקס ההפוך (Inverted Index).

* **NLTK:**

ספרייה לעיבוד שפה טבעית.

שימשה להסרת מילים נפוצות (Stop Words), חיתוך מילים (Stemming), ולניתוח טקסטים על מנת לחלץ משמעות ממילות השאילתה.

1. **כמה זמן רצו השאילתות שלכם?במה זה תלוי? האם לדעתכם ניתן לשפר זמן זה?**

זמן הריצה הממוצע של השאילתות הוא כ-20 שניות.

זמן הריצה של הקוד תלוי בעיקר בביצועי הרשת, מהירות טעינת הדפים, יעילות העיבוד של Selenium ו-BS4, ומשאבי המחשב.

ניתן לשפר את זמן הריצה על ידי שימוש בדפדפן במצב Headless, הפחתת זמני המתנה עם WebDriverWait במקום time.sleep, עיבוד טקסט מקבילי באמצעות ספריות כמו concurrent.futures, ושימוש בכלים מהירים יותר כמו SpaCy לעיבוד שפה טבעית. בנוסף, מעבר לשימוש ב-API רשמי של CDC (אם קיים) במקום Web Scraping עם Selenium יקטין משמעותית את זמן הריצה ויהיה יעיל יותר.

1. **בנו inverted index  ל- 15 המילים הנפוצות שחזרו (מספיק להתיחס ל -20 הדפים הראשונים שחזרו)**



1. **בחרו את אחת השאילתות, וחשבו tfIdf  של המושגים בשאילתא.**

|  |  |
| --- | --- |
| **Title** | **TF-IDF Score** |
| Aircrew and Jet Lag | 1.399901693 |
| Jet Lag | CDC Yellow Book 2024 | 0.839941016 |
| Flu Vaccination by Jet Injector | Influenza (Flu) | 0.316186664 |
| Complementary & Integrative Health Approaches to Travel Wellness | CDC Yellow Book 2024 | 0 |
| Mental Health | CDC Yellow Book 2024 | 0 |
| Aircrew and Workplace Exposures | 0 |
| Deep Vein Thrombosis & Pulmonary Embolism | CDC Yellow Book 2024 | 0 |
| Traveling Safely with Infants & Children | CDC Yellow Book 2024 | 0 |
| Road & Traffic Safety | CDC Yellow Book 2024 | 0 |
| About Work Schedules and Reproductive Health | Reproductive Health and The Workplace | 0 |
| The Pretravel Consultation | CDC Yellow Book 2024 | 0 |
| Aircrew and Reproductive Health | 0 |
| Module 9. Coping with the Night and Evening Shifts, Sleep (Continued) | 0 |
| Module 3. Injuries/Errors (Continued) Groups at Risk for Crashes | 0 |
| Module 9. Coping with the Night and Evening Shifts, References | 0 |
| Other Sources of Information about Work Schedules and Sleep | 0 |
| Examples of Jobs and Reproductive Health | Reproductive Health | 0 |
| Table of Contents | CDC Yellow Book 2024 | 0 |
| Site Index | 0 |
| Substance Use & Substance Use Disorders | CDC Yellow Book 2024 | 0 |

1. **האם בדפים שהוחזרו קיימים hubs? Authorities? נמקו.**

בדפים שהוחזרו קיימים גם Hubs וגם Authorities.

Authorities: דפים אלה מספקים מידע מעמיק ואמין בנושאים ספציפיים ומקבלים הפניות ממקורות רבים. לדוגמה, הדף "Gout | Arthritis | CDC" נחשב ל-Authority, מכיוון שהוא מספק מידע מקיף על מחלת "הגאוט", כולל תסמינים, גורמי סיכון ואפשרויות טיפול. דף זה מהווה מקור סמכותי בתחום הבריאות והרפואה.

Hubs: דפים אלה מתפקדים כמרכזי קישורים ומפנים למקורות מידע נוספים בנושאים רלוונטיים. לדוגמה, הדף "<https://www.cdc.gov/parasites/causes/>" נחשב ל-Hub, מאחר שהוא מרכז קישורים לדפים נוספים בנושאים כמו מניעה, נתונים אפידמיולוגיים ומידע למחנכים בתחום הבריאות.

**הסבר:**

דפי Authority ממוקדים בהצגת תוכן אמין שמבוסס על מומחיות, בעוד ש-Hubs משמשים כאינדקסים שמרכזים קישורים לדפים רלוונטיים אחרים. בדפים שהוחזרו מה-CDC ניתן לזהות את שני הסוגים הללו, כאשר ה-CDC עצמו הוא אתר סמכותי הכולל הן תוכן מומחה (Authorities) והן מרכזי קישורים (Hubs).

**Authorities pages: 12**

1. **בחרו 10 דפים שהזחלן החזיר, ואשר יש ביניהם קישורים .חשבו pageRank  לכל דף, מומלץ להעזר בסקריפט מוכן. הציגו את החישובים ואת הדירוג הסופי.**

קישור לקוד בגיט:

https://github.com/shahaf5641/IR---Project/blob/8ef85fee334eb219ee0af48dd59dd1e9b296bfcc/HW3/pageRank.py

A screenshot of a computer program

Description automatically generated

1. **הראו לשני משתמשים שונים את הדירוג מהסעיף הקודם, בקשו מהם לסמן relevance  feedback .**

Idan:

https://www.cdc.gov/niosh/aviation/prevention/ - R

https://www.cdc.gov/niosh/aviation/prevention/aircrew-jetlag.html - NR

https://www.cdc.gov/niosh/aviation/prevention/aircrew-reproductive-health.html - NR

https://wwwnc.cdc.gov/travel/yellowbook/2024/additional-considerations/substance-use - R

https://wwwnc.cdc.gov/travel/yellowbook/2024/air-land-sea/deep-vein-thrombosis-and-pulmonary-embolism - R

https://wwwnc.cdc.gov/travel/yellowbook/2024/air-land-sea/jet-lag - R

https://wwwnc.cdc.gov/travel/yellowbook/2024/family/infants-and-children - NR

https://wwwnc.cdc.gov/travel/yellowbook/2024/preparing/complementary-and-integrative - R

https://wwwnc.cdc.gov/travel/yellowbook/2024/preparing/mental-health - R

https://wwwnc.cdc.gov/travel/yellowbook/2024/preparing/pretravel-consultation - NR

Yehuda:

https://www.cdc.gov/niosh/aviation/prevention/ - NR

https://www.cdc.gov/niosh/aviation/prevention/aircrew-jetlag.html - R

https://www.cdc.gov/niosh/aviation/prevention/aircrew-reproductive-health.html - NR

https://wwwnc.cdc.gov/travel/yellowbook/2024/additional-considerations/substance-use - NR

https://wwwnc.cdc.gov/travel/yellowbook/2024/air-land-sea/deep-vein-thrombosis-and-pulmonary-embolism - NR

https://wwwnc.cdc.gov/travel/yellowbook/2024/air-land-sea/jet-lag - R

https://wwwnc.cdc.gov/travel/yellowbook/2024/family/infants-and-children - R

https://wwwnc.cdc.gov/travel/yellowbook/2024/preparing/complementary-and-integrative - R

https://wwwnc.cdc.gov/travel/yellowbook/2024/preparing/mental-health - NR

https://wwwnc.cdc.gov/travel/yellowbook/2024/preparing/pretravel-consultation - NR

האם ניתן להציע שאילתא מותאמת על מנת לשפר תוצאות?

שימוש במשוב המשתמשים: אם אנשים דירגו דפים מסוימים כטובים בעבר – אפשר לתת להם עדיפות גבוהה יותר

. Total Relevant Documents In Dataset: 20

|  |  |  |
| --- | --- | --- |
| **User** | **Precision** | **Recall** |
| Idan | 0.5 | 0.25 |
| Yehuda | 0.4 | 0.2 |

1. קישור לדף HTML - <https://shahaf5641.github.io/IR---Project/>

|  |  |
| --- | --- |
| איזה שינוי הוצע? | האם לדעתכם יש מקום לשיפור זה? אם כן - הסבירו כיצד ניתן לשפר. אם לא, נמקו |
| תאמת שאין משהוא שעולה בראש אבל הזמן ריצה 20 שניות זה די המון אני חושב | כן, במקום שימוש בפונקציית time.sleep אפשר להחליף אותה ולהשתמש בפונקציה דינמית שנמצאת בתוך ספריית Selenium שנקראת webDriverWait שתפקידה לחכות שהדף נטען במקום לחכות פרק זמן קבוע גם אם הדף כבר נטען |
| כדאי באינדקס להציג את המונחים ללא סטמינג | ניתן להחליף את הporter stemmer ב Lemmatization |
| לשלב תכנות מקבילי לשיפר זמני הריצה. | לכל שאילתא ניתן לטעון את הדפים בצורה מקבילית ולא אחד אחרי השני באמצעות ספרייה concurrent.futures |

1. מהו היו האתגרים בקורס / פרויקט כצוות?

* קושי לעמוד בלוחות זמנים, במיוחד כשיש עוד קורסים או מחויבויות אחרות.
* פגישות ותיאומים שלקחו זמן רב כי היה קשה למצוא שעה שכולם פנויים.
* שינויי האתר בפרויקט שגרם לביצוע שינויים.