קישור לגיט:  
<https://github.com/shahaf5641/IR---Project>

**מגישים:**

**שרה סאלח 318107448**

**יובל סבאטו 318474269**

**שחף ישראל 209288414**

**שי אוסמו 318370285**

שאלה 1

תנו דוגמא למקרה שבו מסווג Rocchio משייך תגית לא נכונה לדוגמת למידה. הוכיחו את טענתכם. רמז: בידקו מקרה חד מימדי.

**נבחר את ה-terms:**

1. "Jet Lag"
2. "Symptoms"
3. "Treatment"
4. "Causes"
5. "Jaundice"
6. "Pain"

**וקטור שאילתה**  *"What are the symptoms of jaundice?"* **Q**: (0,1,0,0,1,0)   
מסמכים רלוונטיים על התסמינים של צהבת:

**מסמך רלוונטי 1 R1​**: (0,1,0,0,1,0)

**מסמך רלוונטי 2 R2​**: (0,0.8,0.2,0,1,0)

**מסמך לא רלוונטי NR:** (0,0,0,1,0,1)

**וקטור ה centroid של המסמכים הרלוונטיים:**

**וקטור ה centroid של המסמכים הלא הרלוונטיים יהיה:**

**הפעלת נוסחת Rocchio:**

**נגדיר-**α=1 , β=0.8 ,γ=0.3

**שלב החישוב-**

**נראה מצב שבו מתקיים שיוך של תג באופן שגוי:**נבחר מסמך חדש עבור המונחים עם התדירויות הבאות

DN=(1,1,2,1,1,1)

|  |  |
| --- | --- |
| "Jet Lag" | 1 |
| "Symptoms" | 1 |
| "Treatment" | 2 |
| "Causes" | 1 |
| "Jaundice" | 1 |
| "Pain" | 1 |

חישוב הדמיון עם וקטור השאילתה המקורי Q:

**Q**: (0,1,0,0,1,0)  
**DN**=(1,1,2,1,1,1)

Similarity (Q , DN) =0 \*1 + 1\*1 + 0\*2 + 0\*1 + 1\*1+0\*1 =2

חישוב הדמיון עם הווקטור:

Similarity = 0\*1 + 1.72\*1 + 0.08\*2 -0.3\*1 + 1.8\*1-0.3\*1 =3.68

הווקטור המעודכן של השאילתה מעלה את ערך הדמיון עם המסמך החדש . מה שמוביל להסתברות גבוהה יותר שיסווג כרלוונטי, אף על פי שבפועל הוא קרוב יותר למרכז המסמכים הלא רלוונטיים. מכאן ניתן להסיק שאלגוריתם Rocchio עשוי לשגות בסיווג המסמכים.

**שאלה 2**

**סעיף א:**

**D** **A**

**C** **B**

**הסבר:**

* אתחול:

ערכי Hub:

A=B=C=D=1

ערכי Authority:

A=B=C=D=1

* איטרציה ראשונה:

ערכי Hub:

A=C=1, B=D=0

ערכי Authority:

A=C=0, B=D=1

* איטרציה שנייה:  
  ערכי Hub:

A=C=1, B=D=0

ערכי Authority:

A=C=0, B=D=1

ניתן לראות שאין שינוי בערכים. על כן – המערכת הגיעה למצב לפי דרישות השאלה:

**A:** Authority=degree in =0, Hub=degree out=1

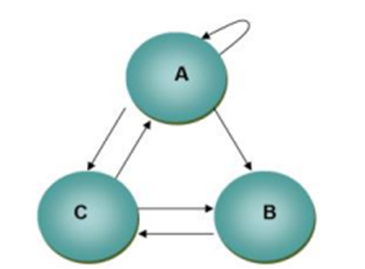
**B:** Authority=degree in =1, Hub=degree out=1

**C:** Authority=degree in =0, Hub=degree out=1

**D:** Authority=degree in =1, Hub=degree out=0

**סעיף ב:**

**נתונה הרשת הבאה:**



שלבי החישוב:

נאתחל כל צומת עם הערך 1 בhub ובauthority.

* אתחול:  
  חישובי ציון authority:

A=B=C=1

חישובי ציון hub:

A=B=C=1

* איטרציה 1:   
  חישובי ציון authority:

Authority(A) = 2

Authority(B) = 2

Authority(C) = 2

חישובי ציון hub:

Hub(A) = 3

Hub(B) = 1

Hub(C) = 2

ננרמל את ציוני הauthority:

Authority(A) = 0.87

Authority(B) = 0.87

Authority(C) = 0.87

ננרמל את ציוני הhub:

Hub(A) = 0.8

Hub(B) = 0.27

Hub(C) = 0.53

* איטרציה 2:   
  חישובי ציון authority:

Authority(A) = 1.33

Authority(B) = 1.33

Authority(C) = 1.07

חישובי ציון hub:

Hub(A) = 2.56

Hub(B) = 0.87

Hub(C) =1.73

ננרמל את ציוני הauthority:

Authority(A) = 0.62

Authority(B) = 0.62

Authority(C) = 0.5

ננרמל את ציוני הhub:

Hub(A) = 0.8

Hub(B) = 0.27

Hub(C) = 0.53

* איטרציה 3:   
  חישובי ציון authority:

Authority(A) = 1.33

Authority(B) = 1.33

Authority(C) = 1.07

חישובי ציון hub:

Hub(A) = 1.72

Hub(B) = 0.49

Hub(C) =1.23

ננרמל את ציוני הauthority:

Authority(A) = 0.62

Authority(B) = 0.62

Authority(C) = 0.5

ננרמל את ציוני הhub:

Hub(A) = 0.8

Hub(B) = 0.27

Hub(C) = 0.53

ניתן לראות שאין שינויים בערכים. על כן – המערכת במצב יציב.

שאלה 3

סעיף א:

**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:**

1. **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.
2. **BeautifulSoup:**
   * ספרייה לניתוח מבנה HTML ולהפקת נתונים מאתרי אינטרנט.
   * בפרויקט נעשה שימוש כדי לאסוף תוצאות חיפוש (כותרות וקישורים) ולנתח אותן.
3. **Pandas:**
   * ספרייה לעיבוד וניתוח נתונים בצורה יעילה באמצעות טבלאות.
   * שימשה ליצירת קובץ Excel עם התוצאות, הדירוג, ומאגר האינדקס ההפוך (Inverted Index).
4. **NLTK:**
   * ספרייה לעיבוד שפה טבעית.
   * שימשה להסרת מילים נפוצות (Stop Words), חיתוך מילים (Stemming), ולניתוח טקסטים על מנת לחלץ משמעות ממילות השאילתה.

סעיף ב

**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.

סעיף ג

**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 "About Chikungunya | Chikungunya Virus | CDC" 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).

סעיף ד

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

בתהליך חישוב ה-PageRank, פונקציה זו מחשבת מחדש את ערכי הדירוג לכל עמוד על בסיס הקישורים המובילים אליו. עבור כל עמוד, היא סורקת אילו עמודים אחרים מקשרים אליו, מחשבת את התרומה שלהם לדירוג על ידי חלוקת הדירוג הנוכחי שלהם במספר הקישורים היוצאים שלהם, ומוסיפה את התרומה הזו לדירוג הכולל של העמוד. תהליך זה חוזר על עצמו מספר איטרציות עד שהדירוגים מתייצבים. התוצאות הראו את הדירוגים הבאים:

A screenshot of a computer program

Description automatically generated

סעיף ה

Bob:

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

Alice:

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?**

**פילטר על פי תוכן דף ומילות מפתח:**

דפים שמזכירים במפורש את הביטוי "jet lag treatment" או "remedies" יקבלו משקל גבוה יותר.

**שימוש במשוב המשתמשים:**

שילוב משוב המשתמשים בקביעת דירוג: דפים שסומנו כ"רלוונטיים" על ידי שניהם יקבלו עדיפות גבוהה יותר בתוצאות.



לתשומת לב:

יש שלושה לשוניות בקובץ האקסל