אחזור מידע – תרגיל בית 3 – אביב התשפד

אופק בן אברהם 313310559

מור שמואל 315040980

מעיין אביטן 313550907

רותם פורת 206607731

<https://github.com/avtn96/Information_Retrieval.git>  
<https://project-information-gamma.vercel.app/final-project>

**שאלה 1 :הרחבות שאילתא ואחזור ממוין**

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

**ניקח את ה-terms הבאים:**

"Engineer"

"Doctor"

"Teacher"

"Pilot"

"Guitar"

Original query vector Q = (2,1,0,1,0)

Relevant document: R1: (1,2,1,1,0), R2: (2,1,1,0,1)

Non-relevant document: N1: (0,1,2,1,1)

**נרצה לחשב את וקטור ה centroid עבור המסמכים הרלוונטיים:**

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

**נפעיל את אלגוריתם Rocchio:**

נניח כי:

לפי הנוסחה:

**נקבל:**

**עכשיו נראה מצב שבו מתקיים שיוך של תג באופן שגוי:**

ניקח מסמך חדש d = (1,1,2,1,1) עם נתונים תדירויות הבאות:

"Engineer" - 1

"Doctor" - 1

"Teacher" - 2

"Pilot" - 1

"Guitar" - 1

נחשב את הדמיון עם וקטור השאילתה המקורי Q: Similarity (Q , D) = 2\*1 + 1\*1 + 0\*2 + 1\*1 + 0\*1 = 4

נחשב את הדמיון עם הווקטור: Similarity ( = 3.125\*1 + 1.875\*1 + 0.25\*2 + 1.125\*1 + 0.125\*1 = 6.75

וקטור השאילתא המשתנה מגביר את ציון הדמיון עם המסמך החדש D מה שמגדיל את הסיכוי שהוא יסווג כרלוונטי, למרות שהוא בעצם קרוב יותר למרכז הלא רלוונטי. לכן, ניתן להבין שמסווג Rocchio עלול לבצע סיווג שגוי של מסמכים.​​

A diagram of a diagram

Description automatically generated**שאלה 2: Link analysis**

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

**קיבלנו שאין שינוי בערכים ולכן המערכת הגיעה למצב יציב כפי שנדרש:**

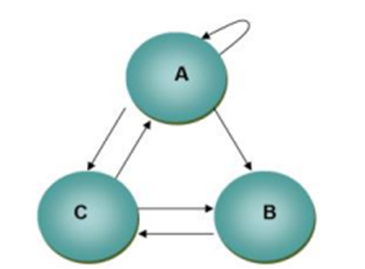
A: Authority=degree in =0, Hub=degree out=1

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

**נתונה הרשת הבאה:**



שלבי החישוב:

* אתחול של 1 לכל צומת גם בhub וגם בauthority
* 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

**קיבלנו שאין שינוי בערכים ולכן המערכת הגיעה למצב יציב.**

**שאלה 3 :קדם פרויקט –בדיקת הזחלן**  
לאחר שבניתם את הזחלן והרצתם את תוצאותיו. עליכם לבצע את הצעדים הבאים:

א. רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.  
**asyncio:** ספריית Python לניהול אסינכרוניות, שאפשרה לנו להריץ מספר פעולות במקביל בצורה יעילה יותר, במיוחד כאשר מדובר בביצוע שאילתות רשת.  
**pyppeteer:** כלי זה משמש כאוטומציה לדפדפן (מבוסס על Puppeteer ל-Node.js), ומאפשר לנו להפעיל את דפדפן Chrome מתוך Python ולבצע פעולות כגון טעינת עמודים ולקיחת תוכן HTML בצורה תוכניתית.  
**BeautifulSoup:** ספרייה לעיבוד וניתוח של תוכן HTML/XML, אשר שימשה אותנו לשליפת מידע מתוך מבנה ה-HTML של הדפים שסרקנו.  
**pandas:** ספריית Python לניתוח ועיבוד נתונים, שאפשרה לנו לבנות DataFrame מכל המידע שנאסף, לבצע פעולות עיבוד נתונים, ולשמור את הנתונים לקובץ CSV. טכנולוגיות אלו אפשרו לנו לבצע סקריפט סריקה מתקדם, לעבד את הנתונים שנאספו, ולהפיק תובנות מפורטות מהאתר Rotten Tomatoes.

ב. כמה זמן רצו השאילתות שלכם? במה זה תלוי? האם לדעתכם ניתן לשפר זמן זה?  
השאילתות רצות בערך כשנייה בגלל שהם ממוקדות והאינדקס מוגדר היטב, ריצת האינדקס תלויה בקובץ csv, ככל שהוא גדול יותר כך יידרש זמן לסרוק ולהחזיר את התוצאות המתאימות, בנוסף זה תלוי בתכנון האינדוקס ובשאילתות, כלומר אם השאילתות מכילות תנאים מורכבים(סינונים למינהם) כך ידרש יותר זמן לסינון התוצאות.   
אנו בטוחים שיש טכנולוגיות מתקדמות יותר שעושות אופטימיזציה לקוד שלנו, שמירה בזכרון מטמון, שימוש במבני נתונים יעילים, חלוקה לתהליכים שיפעלו במקביל, לכן ניתן לשפר זמן זה.

ג.האם בדפים שהוחזרו קיימים hubs? Authorities? נמקו.  
בדפים שהוחזרו קיימים Authorities וגם hubs משום שהדפים מקושרים להרבה דפים חשובים אחרים וגם את הדפים מקשרים הרבה דפים אחרים חשובים ולכן זה גם Authorities וגם hubs.

דוגמה: הדף - <https://www.rottentomatoes.com/tv/tracker_2023/s01>

מוביל להרבה דפים אחרים בתוך rottentomatoes ומצד שני יש כמה דפים שמזכירים את הקישור לדף כמו- <https://www.rottentomatoes.com/browse/tv_series_browse/sort:popular>

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

**השאילתא:** “movies from 2023 with duration > 1.5 h”

<https://www.rottentomatoes.com/m/asteroid_city>: 0.028942919002214944

<https://www.rottentomatoes.com/m/a_man_called_otto>: 0.05354440042450051

<https://www.rottentomatoes.com/m/the_contestant>: 0.028942919002214944

<https://www.rottentomatoes.com/m/the_hatchet_wielding_hitchhiker>: 0.237280124048726

<https://www.rottentomatoes.com/m/to_kill_a_tiger>: 0.041243659713357736

<https://www.rottentomatoes.com/m/this_much_we_know>: 0.23063032926193258

<https://www.rottentomatoes.com/m/boston_strangler_2023>: 0.14224714535737545

<https://www.rottentomatoes.com/m/killers_of_the_flower_moon>: 0.09867594704077756

<https://www.rottentomatoes.com/m/reptile_2023>: 0.06924627807445118

<https://www.rottentomatoes.com/m/pain_hustlers>: 0.06924627807445118

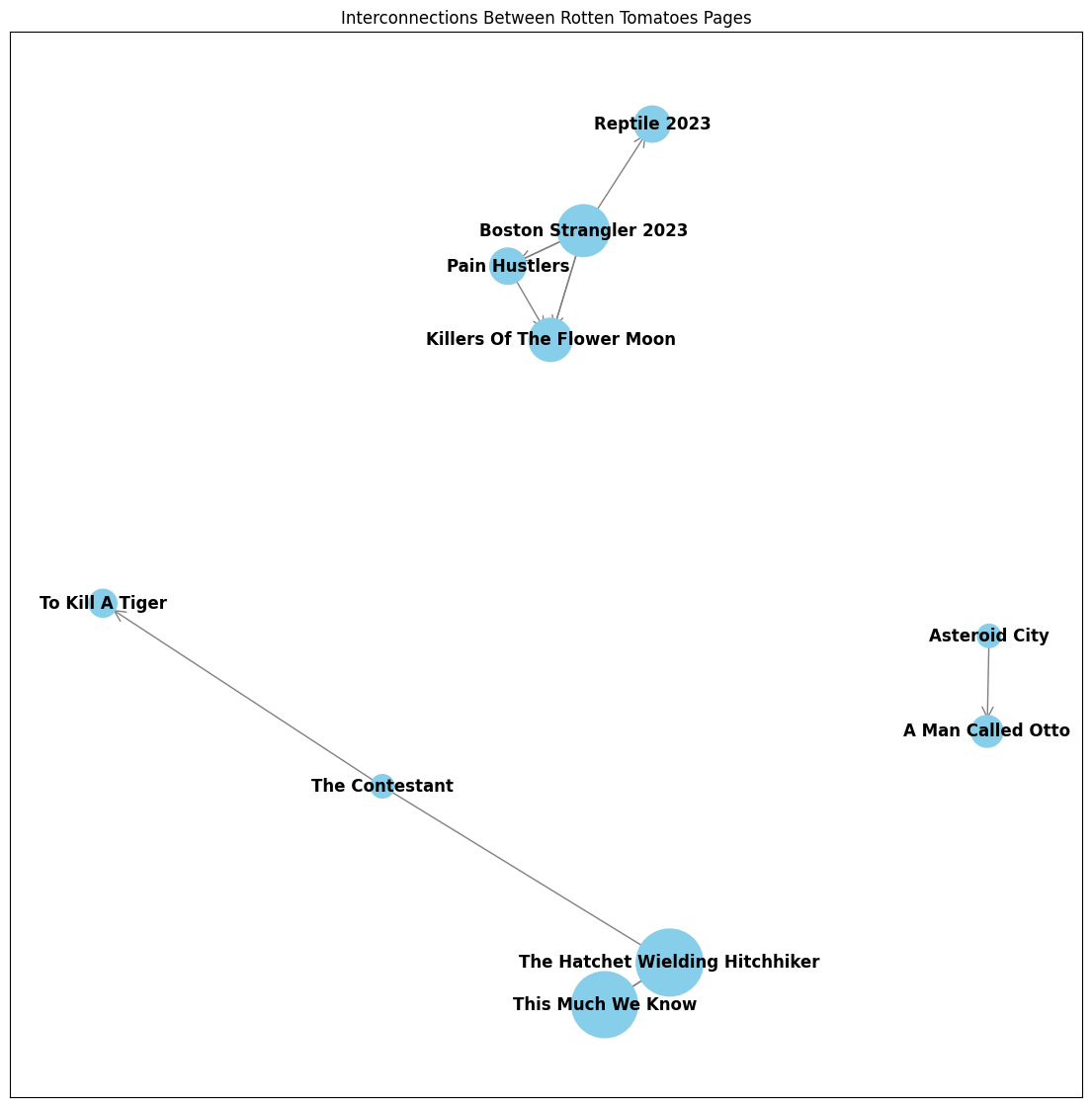
**שלבי החישוב של PageRank:**

יצירת מטריצת סמיכויות (Adjacency Matrix) : הקוד מייצר מטריצה שמייצגת את הקישורים בין הדפים. כל תא במטריצה מציין אם יש קישור מדף מסוים לדף אחר.

חישוב PageRank : משתמשים במטריצת סמיכות מנורמלת לחישוב ערכי PageRank לכל דף. כל דף מקבל ערך שמייצג את מידת החשיבות שלו כאשר ערך זה מחושב לפי הנוסחאות שלמדנו בכיתה, כאשר דפים עם יותר קישורים פנימיים מקבלים ערכים גבוהים יותר. האלגוריתם חוזר על עצמו מספר פעמים (עד למקסימום איטרציות שאצלנו מוגדר כ100) עד שהערכים מתכנסים לשינוי קטן מאוד מוגדר על ידי tol=1e-6.

הצגת תוצאות PageRank : התוצאות מודפסות על המסך כך שכל URL מוצג עם ערך הPageRank שלו.

הדמיית הקשרים בין הדפים: הקוד מייצר גרף שמציג את הקשרים בין הדפים, כאשר גודל הצמתים משקף את ערך ה-PageRank של כל דף.

**התמונה להלן מדגימה את הקישוריות בין הדפים כאשר גודל כל צומת הוא לפי הדירוג.**קישור לקוד של ה pageRank : <https://github.com/avtn96/Information_Retrieval/blob/main/page_rank.py>

ה. הראו לשני משתמשים שונים את הדירוג מהסעיף הקודם, בקשו מהם לסמן relevance ובצעו feedback relevance. האם ניתן להציע שאילתא מותאמת על מנת לשפר תוצאות?

**Ron:**

<https://www.rottentomatoes.com/m/asteroid_city> R

<https://www.rottentomatoes.com/m/a_man_called_otto> R

<https://www.rottentomatoes.com/m/the_contestant> NR

<https://www.rottentomatoes.com/m/the_hatchet_wielding_hitchhiker> R

<https://www.rottentomatoes.com/m/to_kill_a_tiger> R

<https://www.rottentomatoes.com/m/this_much_we_know> NR

<https://www.rottentomatoes.com/m/boston_strangler_2023> R

<https://www.rottentomatoes.com/m/killers_of_the_flower_moon> R

<https://www.rottentomatoes.com/m/reptile_2023> R

<https://www.rottentomatoes.com/m/pain_hustlers> NR

**Amit:**

<https://www.rottentomatoes.com/m/asteroid_city> NR

<https://www.rottentomatoes.com/m/a_man_called_otto> R

<https://www.rottentomatoes.com/m/the_contestant> NR

<https://www.rottentomatoes.com/m/the_hatchet_wielding_hitchhiker> R

<https://www.rottentomatoes.com/m/to_kill_a_tiger> R

<https://www.rottentomatoes.com/m/this_much_we_know> R

<https://www.rottentomatoes.com/m/boston_strangler_2023> NR

<https://www.rottentomatoes.com/m/killers_of_the_flower_moon> R

<https://www.rottentomatoes.com/m/reptile_2023> R

<https://www.rottentomatoes.com/m/pain_hustlers> R

**ניתן להציע שאילתא מותאמת על מנת לשפר תוצאות אלו.**

הצעות לשיפור השאילתה:

* אופטימיזציה לפי פידבק:

ניתן להגדיר שאילתה מותאמת אישית שמביאה בחשבון את הפידבק של המשתמשים, למשל מתן עדיפות לדפים שסומנו כרלוונטיים על ידי שני המשתמשים.

* הוספת מילות מפתח:

הוספת מילות מפתח המבוססות על התכנים שמעניינים את המשתמשים.

* אינטראקציה מתמשכת עם המשתמש:

ניתן להמשיך ולשאול את המשתמשים על תוצאות אחרות כדי לשפר את איכות השאילתה.

**Ron:**

Precision= 7/10 = 0.7

Recall = 7/76 = 0.0921

**Amit:**

Precision = 7/10 = 0.7

Recall = 7/76 = 0.0921

 ו. הציגו את כל התוצרים של סעיפים א-ו בדף HTML יחיד. הדף יכול להיות דינאמי או סטטי (כלומר מציג צילומי מסך של התשובות לשאלות). העלו את הדף לgit pages. צרפו קישור לדף שהעלתם.

**קישור לאתר:** <https://project-information-gamma.vercel.app/final-project>