**תרגיל בית 3:**

**מגישים:**

ישראל שושן 318740834

עמית פרץ 209054303

דוד אסולין 205908569

יובל הילאי 318609237

**שאלה 1:**

**א. דוגמא למקרה שבו מסווג Rocchio משייך תגית לא נכונה לדוגמת למידה:**

נבחן מקרה חד ממדי פשוט:

נניח שיש לנו דוגמאות למידה במרחב חד-ממדי (ציר אחד) ושתי תגיות: 'חיובי' ו'שלילי'.

* דוגמאות חיוביות: 1, 2, 3
* דוגמאות שליליות: 8, 9, 10

מסווג Rocchio מחושב על ידי מציאת מרכז המסה (centroid) של כל קטגוריה:

* מרכז המסה של הדוגמאות החיוביות: (1 + 2 + 3) / 3 = 2
* מרכז המסה של הדוגמאות השליליות: (8 + 9 + 10) / 3 = 9

עכשיו, נבחן דוגמה חדשה 4:

* המרחק ממרכז המסה החיובי: |4 - 2| = 2
* המרחק ממרכז המסה השלילי: |4 - 9| = 5

לפי מסווג Rocchio, דוגמה זו תסווג כ'חיובית' מכיוון שהמרחק למרכז המסה החיובי קטן יותר מהמרחק למרכז המסה השלילי.

אבל, אם נניח שבעצם הדוגמה החדשה היא 'שלילית', מסווג Rocchio יטעה בסיווגה. הוכחנו מקרה שבו מסווג Rocchio משייך תגית לא נכונה לדוגמת למידה.

**שאלה 2:**

**א. דוגמא:**

נבחן גרף מכוון עם ארבעה צמתים A, B, C, D ושתי קשתות :A → B C → D

* צומת A: דרגת כניסה 0, דרגת יציאה 1
* צומת B: דרגת כניסה 1, דרגת יציאה 0
* צומת C: דרגת כניסה 0, דרגת יציאה 1
* צומת D: דרגת כניסה 1, דרגת יציאה 0

בגרף זה ניתן לראות כי ציון הauthority של כל צומת שווה לדרגת הכניסה שלו, וציון הhub של כל צומת שווה לדרגת היציאה שלו, כפי שהתבקש.

**ג. חישוב ציוני ה HubוהAuthority**

בהנתן הרשת הבאה, נחשב את ציוני הhub והauthority של כל דף עד להתכנסות.

A diagram of a diagram of a diagram

Description automatically generatedנציב את הגרף באופן הבא כפי שבוצע בהרצאה:

|  |  |  |  |
| --- | --- | --- | --- |
| C | B | A |  |
| 1 | 1 | 1 | A |
| 1 | 0 | 0 | B |
| 0 | 1 | 1 | C |

X =

**חישוב ציוני ה Hub והAuthority:**

1. נתחיל עם הציונים התחלתיים, ה- Hub ו Authority של כל צומת הם 1.
2. נחשב מחדש את ציוני ה Authority וה Hub של הצמתים המקושרים:

נקבל Authority (כמה נכנסים אליו)

* + a(A) = h(A) + h(C) = 1 + 1 = 2
  + a(B) = h(A) + h(C) = 1 + 1 = 2
  + a(C) = h(A) + h(B) = 1 + 1 = 2

נקבל Hub (כמה יוצאים ממנו)

* + h(A) = a(A) + a(B) + a(C) = 6
  + h(B) = a(C) = 2
  + h(C) = a(B) + a(A) = 4

ננרמל:

סה"כ Authority = 2 + 2 + 2 = 6

* a(A) = 2 / 6
* a (B) = 2 / 6
* a(C) = 2 / 6

סה"כ Hub = 6 + 2 + 4 = 12

* a(A) = 6 / 12 = 0.5
* a (B) = 2 / 12 = 0.1667
* a(C) = 4 / 12 = 1/3

שלב שני:

Authority:

a(A) = h(A) + h(C) = 0.5 + 0.3333 = 0.8333

a(B) = h(A) + h(C) = 0.5 + 0.3333 = 0.8333

a(C) = h(A) + h(B) = 0.5 + 0.1667 = 0.6667

Hub:

h(A) = a(A) + a(B) + a(C) = 0.8333 + 0.8333 + 0.6667 = 2.3333

h(B) = a(C) = 0.6667

h(C) = a(B) + a(A) = 0.8333 + 0.8333 = 1.6666

נרמול:

סה"כ Authority = 0.8333 + 0.8333 + 0.6667 = 2.3333

a(A) = 0.8333 / 2.3333 = 0.3571

a(B) = 0.8333 / 2.3333 = 0.3571

a(C) = 0.6667 / 2.3333 = 0.2858

סה"כ Hub = 2.3333 + 0.6667 + 1.6666 = 4.6666

h(A) = 2.3333 / 4.6666 = 0.5

h(B) = 0.6667 / 4.6666 = 0.1429

h(C) = 1.6666 / 4.6666 = 0.3571

שלב שלישי:

Authority:

a(A) = h(A) + h(C) = 0.5 + 0.3571 = 0.8571

a(B) = h(A) + h(C) = 0.5 + 0.3571 = 0.8571

a(C) = h(A) + h(B) = 0.5 + 0.1429 = 0.6429

Hub:

h(A) = a(A) + a(B) + a(C) = 0.8571 + 0.8571 + 0.6429 = 2.3571

h(B) = a(C) = 0.6429

h(C) = a(B) + a(A) = 0.8571 + 0.8571 = 1.7142

נרמול:

סה"כ Authority = 0.8571 + 0.8571 + 0.6429 = 2.3571

a(A) = 0.8571 / 2.3571 = 0.3638

a(B) = 0.8571 / 2.3571 = 0.3638

a(C) = 0.6429 / 2.3571 = 0.2724

סה"כ Hub = 2.3571 + 0.6429 + 1.7142 = 4.7142

h(A) = 2.3571 / 4.7142 = 0.5

h(B) = 0.6429 / 4.7142 = 0.1364

h(C) = 1.7142 / 4.7142 = 0.3636

שלב רביעי:

Authority:

a(A) = h(A) + h(C) = 0.5 + 0.3636 = 0.8636

a(B) = h(A) + h(C) = 0.5 + 0.3636 = 0.8636

a(C) = h(A) + h(B) = 0.5 + 0.1364 = 0.6364

Hub:

h(A) = a(A) + a(B) + a(C) = 0.8636 + 0.8636 + 0.6364 = 2.3636

h(B) = a(C) = 0.6364

h(C) = a(B) + a(A) = 0.8636 + 0.8636 = 1.7272

נרמול:

סה"כ Authority = 0.8636 + 0.8636 + 0.6364 = 2.3636

a(A) = 0.8636 / 2.3636 = 0.3653

a(B) = 0.8636 / 2.3636 = 0.3653

a(C) = 0.6364 / 2.3636 = 0.2694

סה"כ Hub = 2.3636 + 0.6364 + 1.7272 = 4.7272

h(A) = 2.3636 / 4.7272 = 0.5

h(B) = 0.6364 / 4.7272 = 0.1346

h(C) = 1.7272 / 4.7272 = 0.3654

שלב חמישי:

Authority:

a(A) = h(A) + h(C) = 0.5 + 0.3654 = 0.8654

a(B) = h(A) + h(C) = 0.5 + 0.3654 = 0.8654

a(C) = h(A) + h(B) = 0.5 + 0.1346 = 0.6346

Hub:

h(A) = a(A) + a(B) + a(C) = 0.8654 + 0.8654 + 0.6346 = 2.3654

h(B) = a(C) = 0.6346

h(C) = a(B) + a(A) = 0.8654 + 0.8654 = 1.7308

נרמול:

סה"כ Authority = 0.8654 + 0.8654 + 0.6346 = 2.3654

a(A) = 0.8654 / 2.3654 ≈ 0.3660

a(B) = 0.8654 / 2.3654 ≈ 0.3660

a(C) = 0.6346 / 2.3654 ≈ 0.2680

סה"כ Hub = 2.3654 + 0.6346 + 1.7308 = 4.7308

h(A) = 2.3654 / 4.7308 = 0.5000

h(B) = 0.6346 / 4.7308 ≈ 0.1341

h(C) = 1.7308 / 4.7308 ≈ 0.3660

ניתן כבר לראות שהערכים באיטרציות כבר קרובים לכן אפשר לעצור את החישובים.

**שאלה 3:**

* 1. השתמשנו בטכנולוגיות הבאות:

1. ספריות Python:
   * - requests: לשליחת בקשות HTTP כדי למשוך דפי אינטרנט.
   * - BeautifulSoup: לניתוח HTML וחילוץ מידע.
   * - pandas: לניהול נתונים ולייצוא תוצאות לExcel .
   * openpyxl: - לעבודה עם קבצי Excel ולעיצוב תאים.
   * nltk: - לעיבוד שפה טבעית, כולל הסרת מילים לא חשובות.
2. ניהול נתונים:
   * - Counter לספירת תדירות מילים.
   * defaultdict - לבניית אינדקס הפוך.
3. חישובים מתמטיים:
   * - mathלחישוב ציוני IDF ו TF – IDF.
4. טכנולוגיות אינטרנט:
   * urljoin - ליצירת כתובת URL מלאה ממסלולים יחסיים.
   1. זמן ריצה של השאילתות:



הוא תלוי ב:

* + **גודל הדף**: דפים ארוכים לוקחים יותר זמן לעיבוד.
  + **מספר הדפים**: מספר הדפים המתקבלים משפיע על הזמן הכולל.
  + **תשובות מהשרת**: זמן תגובה של השרת.
  + **קוד**: יעילות הקוד ואופטימיזציה.

שיפורים:

* + **שימוש בטכניקות -cashing**לשמור תוצאות של דפים שנבדקו.
  + **הגבלת מספר השאילתות**: כדי למנוע עיכובים מהשרת.
  + **אופטימיזציה של קוד**: שיפור אלגוריתמים לניקוי טקסט והפקת מילות מפתח
  1. לכל דף שהוחזר חושבו כל הלינקים שיש בדף הזה:

A blue and white text

Description automatically generated

עם קביעת Nodes ו Edges



ועבור כל דף שהוחזר חושבו ה hubsו –Authorities:

A screenshot of a computer

Description automatically generated

* אם יש ערכים גבוהים בשדה ה hubs זה מצביע על כך שישנם דפים שמרכזים קישורים רבים, מה שהופך אותם ל-hubs חשובים.
* אם יש ערכים גבוהים בשדה ה authorities זה מצביע על כך שישנם דפים שמקבלים קישורים רבים, מה שהופך אותם ל-authorities חשובים.

ד.

10 דפים שהזחלן החזיר:

1. 'https://www.ea.com/news/hiring-our-heroes',
2. 'https://www.ea.com/ea-play',
3. 'https://www.ea.com/games',
4. 'https://www.ea.com/sports',
5. 'https://www.ea.com/careers',
6. 'https://www.ea.com/news/2024-ea-sports-latest-tech-innovations',
7. 'https://www.ea.com/news/how-the-community-helps-shape-the-sims',
8. 'https://www.ea.com/legal',
9. 'https://www.ea.com/playtesting',
10. 'https://www.ea.com/store'

קיבלנו:

A screenshot of a computer

Description automatically generated

A diagram of a network

Description automatically generated

ה.

|  |  |  |  |
| --- | --- | --- | --- |
| page | PageRank | User1 | User2 |
| 1 | 0.1763 | רלוונטי | רלוונטי |
| 2 | 0.1601 | רלוונטי | רלוונטי |
| 3 | 0.145 | רלוונטי | רלוונטי |
| 4 | 0.145 | רלוונטי | רלוונטי |
| 5 | 0.1130 | לא רלוונטי | רלוונטי |
| 6 | 0.1130 | רלוונטי | לא רלוונטי |
| 7 | 0.0369 | לא רלוונטי | לא רלוונטי |
| 8 | 0.0369 | לא רלוונטי | לא רלוונטי |
| 9 | 0.0369 | לא רלוונטי | לא רלוונטי |
| 10 | 0.0369 | לא רלוונטי | לא רלוונטי |

ניתן לראות כי ארבעת הדפים הראשונים רלוונטיים עבור שני המשתמשים, הדף החמישי והשישי חלוקים בעיניי שניהם וארבעת הדפים האחרונים הוגדרו בתור לא רלוונטיים.

**Git:**

[AmitPerets/Index\_Invincibles\_HW\_3 (github.com)](https://github.com/AmitPerets/Index_Invincibles_HW_3)