**Information retrieval HW 3.**

**Group: 404 Brain Not Found.**

**Submmiters:**

Tomer Lazarovitch 316377100  
Itamar Kraus 318304763  
Guy Pariente 206132250  
Chay Fadida 318302783

**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 2:

Link to analysis .

**פתרון:**

נניח שיש לנו את ה- terms הבאים:

1. “apple”
2. “banana”
3. “cherry”
4. “date”
5. “elderberry”

בנוסף:

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)

ראשית, נחשב את וקטור ה centroid עבור המסמכים הרלוונטיים:

שנית, נחשב את וקטור ה centroid עבור המסמכים הלא הרלוונטיים:

כעת, נפעיל את אלגוריתם Rocchio:

נניח כי

ונחשב לפי הנוסחה:

נקבל:

כעת נראה דוגמא עבורה מתקיימת שיוך תג שגוי:

נניח כי נתון מסמך חדש D = (1,1,2,1,1) עם התדירויות הבאות –

1. “apple” : 1
2. “banana” : 1
3. “cherry” : 2
4. “date” : 1
5. “elderberry” : 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 השאילתה המשתנה מגדיל את ציון הדמיון עם המסמך החדש D, מה שהופך אותו לסביר יותר להיות מסווג כרלוונטי למרות שהוא קרוב יותר למרכז הלא רלוונטי.  
**לכן ניתן לראות כי מסווג Rocchio עלול לסווג מסמכים בצורה לא נכונה**.

**שאלה 2: Link analysis**

א.

תנו דוגמא לגרף מכוון, הכולל לפחות 4 צמתים ו-2 קשתות, כך שלכל צומת שלו מתקיים: ציון המומחיות (score authority) של הצומת הוא דרגת הכניסה שלו (degree-in)וציון הריכוזיות שלו (score hub)הוא דרגת היציאה שלו.

**פתרון:**

הגרף המכוון:

ניתן לראות כי ישנם 4 צמתים ו-2 קשתות.

B

A

D

C

נחשב Hub scores ו- Authority score:

**Initially:**

|  |  |
| --- | --- |
| Hub scores | Authority score |
| A -> 1 | A -> 1 |
| B -> 1 | B -> 1 |
| C -> 1 | C -> 1 |
| D -> 1 | D -> 1 |

**After 1st iteration:**

|  |  |
| --- | --- |
| Hub scores | Authority score |
| A -> 1 | A -> 0 |
| B -> 0 | B -> 1 |
| C -> 1 | C -> 0 |
| D -> 0 | D -> 1 |

**After 2nd iteration:**

|  |  |
| --- | --- |
| Hub scores | Authority score |
| A -> 1 | A -> 0 |
| B -> 0 | B -> 1 |
| C -> 1 | C -> 0 |
| D -> 0 | D -> 1 |

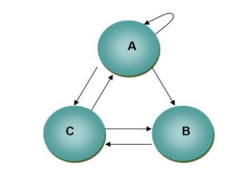
ניתן לראות כי כבר באיטרציה השנייה אין שינוי והתוצאה יציבה.

קיבלנו:

* Node A : degree in = 0 , authority score = 0 , degree out = 1 , hub score = 1.
* Node B : degree in = 1 , authority score = 1 , degree out = 0 , hub score = 0.
* Node C : degree in = 0 , authority score = 0 , degree out = 1 , hub score = 1.
* Node D : degree in = 1 , authority score = 1 , degree out = 0 , hub score = 0.

ב.

בהינתן הרשת הבאה:



מהם ציוני ה hub וה- authority של כל דף? חשבו עד להתכנסות. נרמלו לפי שיטה שהוצעה בכיתה, ציינו במפורש מה הנרמול.

**פתרון:**

ננרמל כפי שלמדנו כלומר :

**Initially:**

|  |  |
| --- | --- |
| Hub scores | Authority score |
| A -> 1 | A -> 1 |
| B -> 1 | B -> 1 |
| C -> 1 | C -> 1 |

**After 1st iteration:**

|  |  |  |  |
| --- | --- | --- | --- |
| Hub scores | Authority score | H-norm | A-norm |
| A -> 3 | A -> 2 | A -> | A -> |
| B -> 1 | B -> 2 | B -> | B -> |
| C -> 2 | C -> 2 | C -> | C -> |

**After 2nd iteration:**

|  |  |  |  |
| --- | --- | --- | --- |
| Hub scores | Authority score | H-norm | A-norm |
| A -> 1 | A -> | A -> | A -> |
| B -> | B -> | B -> | B -> |
| C -> | C -> | C -> | C-> |

**H(a) = A(a,b,c) , H(b) = A(c) , H(c) = A(a,b) | A(a) = H(a,c) , A(b) = H(a,c) , A(c) = H(a,b)**

**After 3rd iteration:**

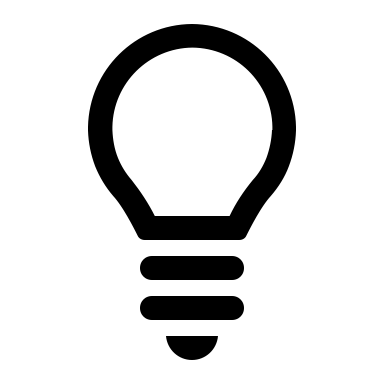
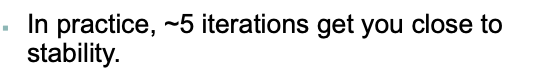
|  |  |  |  |
| --- | --- | --- | --- |
| Hub scores | Authority score | H-norm | A-norm |
| A -> | A -> | A -> | A -> |
| B -> | B -> | B -> | B -> |
| C -> | C -> | C -> | C-> |

**After 4th iteration:**

|  |  |  |  |
| --- | --- | --- | --- |
| Hub scores | Authority score | H-norm | A-norm |
| A -> | A -> | A -> | A -> |
| B -> | B -> | B -> | B -> |
| C -> | C -> | C -> | C-> |

**After 5th iteration:**

|  |  |  |  |
| --- | --- | --- | --- |
| Hub scores | Authority score | H-norm | A-norm |
| A -> | A -> | A -> | A -> |
| B -> | B -> | B -> | B -> |
| C -> | C -> | C -> | C-> |



**שאלה 3 :קדם פרויקט –בדיקת הזחלן**

לאחר שבניתם את הזחלן והרצתם את תוצאותיו. עליכם לבצע את הצעדים הבאים:

1. רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.
2. כמה זמן רצו השאילתות שלכם?במה זה תלוי? האם לדעתכם ניתן לשפר זמן זה?
3. האם בדפים שהוחזרו קיימים hubs? Authorities? נמקו.
4. בחרו 10 דפים שהזחלן החזיר, ואשר יש ביניהם קישורים .חשבו pageRank לכל דף, מומלץ להעזר בסקריפט מוכן. הציגו את החישובים ואת הדירוג הסופי.
5. הראו לשני משתמשים שונים את הדירוג מהסעיף הקודם, בקשן מהם לסמן relevance ובצעו feedback relevance .האם ניתן להציע שאילתא מותאמת על מנת לשפר תוצאות?
6. הציגו את כל התוצרים של סעיפים א-ו בדף HTML יחיד. הדף יכול להיות דינמי או סטטי (כלומר מציג צילומי מסך של התשובות לשאלות). העלו את הדף לgit pages. צרפו קישור לדף שהעליתם.

**פתרון:**

1. רשימת טכנולוגיות שונות ומעניינות שהשתמשנו בהם בפרויקט:

* **Selenium:** כלי אוטומציה שמשמש לתכנות זחלנים (crawlers) על מנת לגשת ולשלוט על דפדפני אינטרנט בצורה אוטומטית לצורך איסוף מידע מאתרים.
* **BeautifulSoup:** ספרייה ב-Python המשמשת לניתוח ועיבוד קוד HTML ו-XML ומקלה על חילוץ נתונים מתוך עמודי אינטרנט בקלות וביעילות.
* **PrettyTable:** היא ספריית Python המשמשת ליצירה והצגה של טבלאות בפורמט טקסטואלי נוח לקריאה, בין אם זה בקונסול, בקובץ או בכל פורמט טקסט אחר. היא מאפשרת עיצוב וארגון נתונים בטבלאות בצורה קלה ונוחה.
* **Networkx:** היא ספריית Python המאפשרת ליצור, לנתח ולתפעל גרפים, והיא שימושית במיוחד לצורכי אחזור מידע בזכות יכולת ניתוח קשרים וסטטיסטיקות בין צמתים.
* **Matplotlib:** היא ספריית Python המאפשרת ליצור גרפים וויזואליזציות של נתונים בצורה גמישה ומותאמת אישית.

1. 3 שאילתות:
   1. טופ 5 בפסים בכל הזמנים (כל העונות) לשחקנים ממוצא אנגלי שמשחקים עדיין בקבוצות מלונדון.
   2. טופ 10 שוערים שהמוצא שלהם הוא מחוץ לאירופה ועם הכי הרבה רשת נקייה בכל הזמנים(משחקים ללא ספיגות גולים).
   3. כל כתבות החדשות עבור השחקנים מ-Ipswich Town שסטטוס ההעברה שלהם הוא: Transfer In או Loan Out
2. ⁠query run time: 95.85 seconds First
3. ⁠query run time: 94.50 seconds Second
4. ⁠ Third query run time: 20.06 seconds

נשים לב כי שתי השאילתות הראשונות מחפשות בעמוד הStats, עמוד בעל פילטרים, צריכות לדפדף בטבלה שנוצרה מהסינונים שנבחרו על מנת לסנן תוצאות עבור השאילתא שהגדרנו לעומת השאילתא השלישית שמחפשת בעמוד אחר ובטבלה קטנה בהרבה.  
יתכן וניתן לקצר את זמנים אלו ע״י שימוש בטכנולוגיות אחרות.

1. שתי השאילתות הראשונות מחזירות לינקים לעמודי השחקנים, כל עמוד הוא עצמאי ואינו מכיל קישור לעמוד שחקן אחר, לכן בשתי שאילתות אלו לא יהיה לנו Hubs & Authorities.  
   בשאילתא האחרונה אנו מחזירים עמוד חדשות על ההעברה של שחקן, בעמוד זה יש קישורים לחדשות אחרות לכן ייתכן ויהיו קיימים Hubs & Authorities.
2. PageRank for each page(alpha=0.85):

<https://www.itfc.co.uk/news/2024/june/30/omari-is-back> : 0.0823

<https://www.itfc.co.uk/news/2024/july/01/town-sign-ben-johnson> : 0.0823

<https://www.itfc.co.uk/news/2024/july/12/greaves-becomes-third-signing> : 0.0823

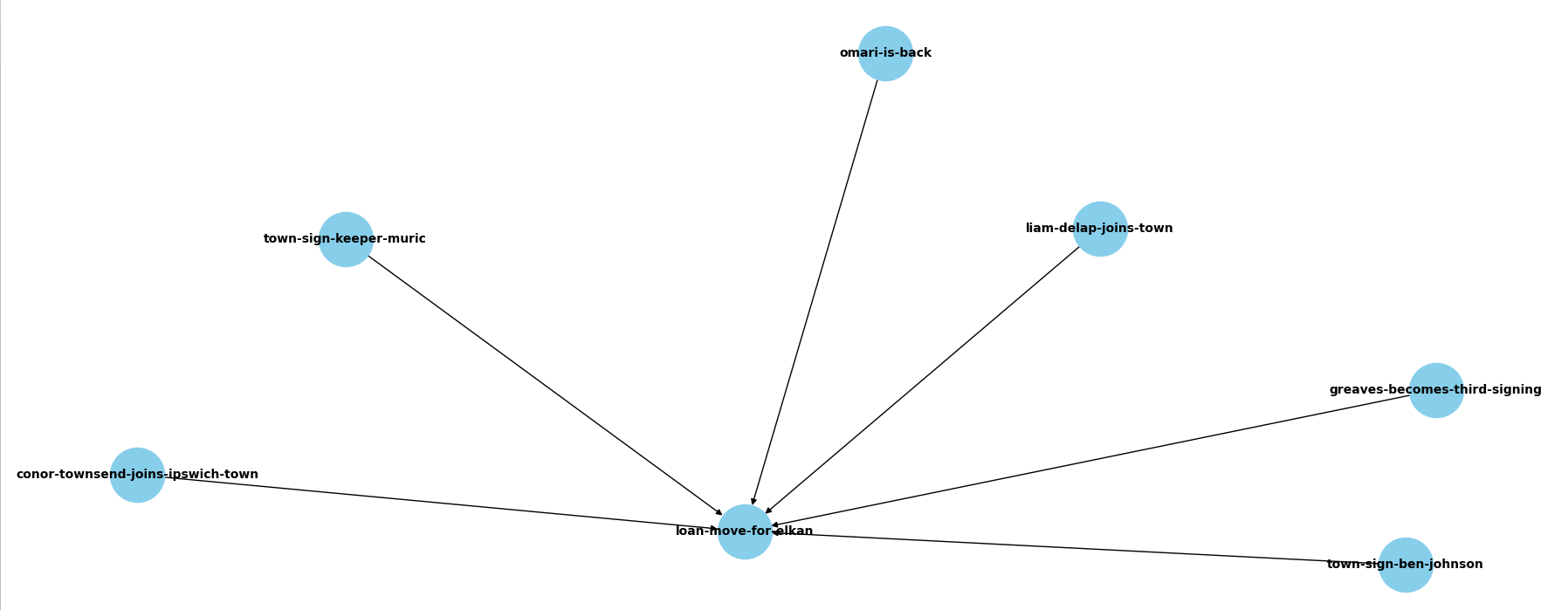
<https://www.itfc.co.uk/news/2024/july/13/liam-delap-joins-town> : 0.0823

<https://www.itfc.co.uk/news/2024/july/17/town-sign-keeper-muric> : 0.0823

<https://www.itfc.co.uk/news/2024/august/01/conor-townsend-joins-ipswich-town> : 0.0823

<https://www.itfc.co.uk/news/2024/august/05/loan-move-for-elkan> : 0.504

החישובים בוצעו באמצעות ספריית networkx, באמצעות ספרייה זו יצרנו גרף מכוון לכל עם צומת עבור כל לינק, וסקריפט לחיפוש href מתוך אותה רשימת צמתים בתוך כל דף. ברגע שמצאנו לינק לדף אחר מוסיפים קשת בהתאם.  
בספריית network אפשר לבצע גם PageRank עם אלפא לבחירה, במקרה שלנו בחרנו 0.85.

הצגת הגרף שהתקבל באמצעות matplotlib:  


1. על פי תוצאות PageRank, אנו רואים שעמוד אחד בלבד מקבל דירוג גבוה באופן משמעותי לעומת השאר. כדי להעריך את רלוונטיות התוצאות, ביקשנו משני משתמשים לסמן את התוצאות לפי הרלוונטיות שלהם.  
   משתמש א: סימן את כל העמודים כרלוונטיים חוץ מ loan-move-for-elkan  
   משתמש ב: סימן את כל העמודים כרלוונטיים לשאילתה

בהתבסס על המשוב, ניתן לשפר את השאילתות על ידי התאמת קריטריונים נוספים לדירוג כגון הוספת פרמטרים של רלוונטיות אישית למשתמשים, כדוגמא במקרה זה פרמטר עבור האם השחקן בסגל הבכיר או בקבוצת הנוער. שינוי זה יאפשר התאמה טובה יותר של תוצאות החיפוש להעדפות המשתמשים ויכול לשפר את איכות החיפושים.

1. <https://scraping-pl.vercel.app/homework/>