אחזור מידע – תרגיל בית 3 – חורף התשפ"ה - Crawl Commandos

מגישים:  
צחי בקל 315730176

דניאל ארמגניאן 209146943  
יונתן שרר 318317682  
ליאור ז'ילגו 316109115

**קישור לגיט:** <https://github.com/audiblemaple/Data-acquisition-course>

שאלה 1

אלגוריתם Rocchio (המקושר במקור למערכת אחזור מידע) מבוסס על חישוב "וקטור מרכז" (centroid) עבור כל תווית/class, בדרך־כלל על ידי ממוצע (או ממוצע משוקלל) של הווקטורים המתאימים לדוגמאות באותה התווית.  
  
נסתכל על דוגמה של מקרה חד מימדי:  
נניח שיש לנו שני סוגי תוויות (מחלקות):

* תווית חיובית (חיובי)
* תווית שלילית (שלילי)

וסט אימון חד מימדי שכל ערך מייצג נקודה בציר X:

* דוגמאות חיוביות 4,5 (+=Y)
* דוגמאות שליליות 0,1 (-=Y)

נחשב סנטרוידים עבור המחלקה החיובית והשלילית:

סיווג של דוגמה חדשה:

נבחן עבור 2 = X חדש

* נחשב מרחקים לסנטרויד החיובי והשלילי:

נשים לב שהמרחק למרכז השלילי (1.5) קטן יותר מהמרחק למרכז החיובי (2.5) ולכן ROCCHIO יסווג את הנקודה החדשה (2) כשלילית.

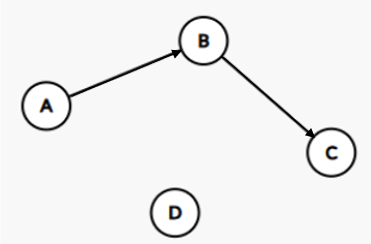
נניח שהנקודה 2 החדשה היא דוגמה חיובית, קיבלנו ש ROCCHIO סיווג אותה לא נכון.

ובאופן כללי, נקודה חדשה ש"נופלת באמצע" (אך נוטה טיפה קרוב יותר למרכז המחלקה השנייה) עלולה להיות מסווגת שגוי, גם אם היא באמת שייכת למחלקה אחרת.

הצגנו דוגמה נגדית וכך הוכחנו את הטענה שלנו.

שאלה 2

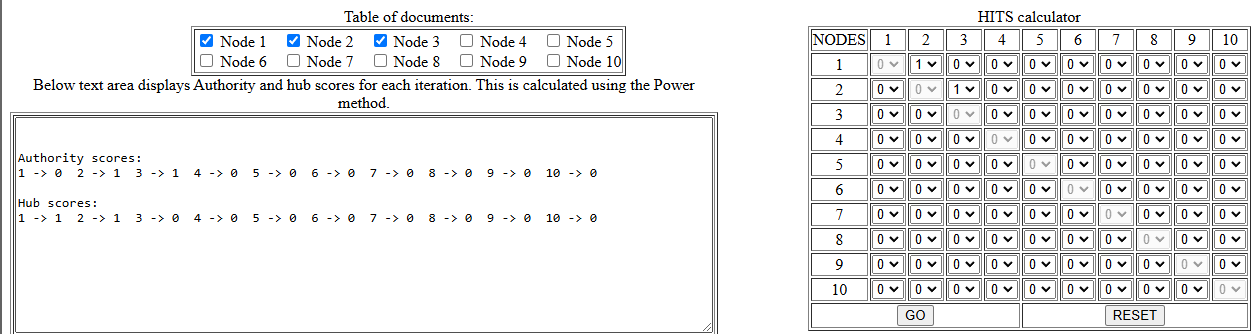
א. דוגמה לגרף:



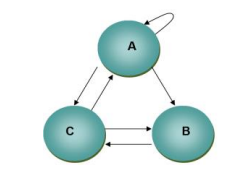
נבצע ניתוח של הגרף:

* **Node A**
  + in-degree = 0 no incoming edges
  + out-degree = 1 edge from A to B
* **Node B**
  + in-degree = 1 incoming edge from A
  + out-degree = 1 edge from B to C
* **Node C**
  + in-degree = 1 incoming edge from B
  + out-degree = 0
* **Node D**
  + in-degree = 0
  + out-degree = 0

|  |  |
| --- | --- |
| ​in-deg(A)=0 | out-deg(A)=1 |
| in-deg(B)=1 | out-deg(B)=1 |
| in-deg(C)=1 | out-deg(C)=0 |
| in-deg(D)=0 | out-deg(D)=0 |

נבדוק זאת ע"י: <https://www.cs.sjsu.edu/faculty/pollett/masters/Semesters/Spring10/amith/HITS.html>  


ג.



**איטרציה 1**

|  |  |  |
| --- | --- | --- |
| Authority | Hub | initialization |
| 0 | 1 | A |
| 0 | 1 | B |
| 0 | 1 | C |

נעדכן את ציוני ה HUB

|  |  |
| --- | --- |
| Authority | Update HUB |
| 2 | A |
| 2 | B |
| 2 | C |

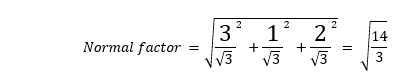
ננרמל את שתי הציונים

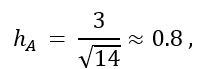


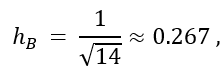


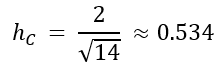
נעכן ציוני HUB וננרמל

|  |  |
| --- | --- |
| Hub | Update HUB |
|  | A |
| 2 | B |
| 2 | C |









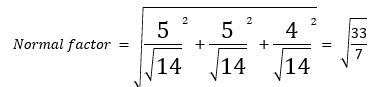
**איטרציה 2**

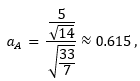
|  |  |  |
| --- | --- | --- |
| Authority | Hub | initialization |
|  |  | A |
|  |  | B |
|  |  | C |

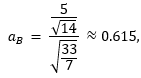
נעדכן authority

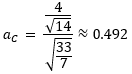
|  |  |
| --- | --- |
| Authority | initialization |
|  | A |
|  | B |
|  | C |

ננרמל





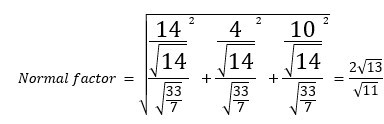


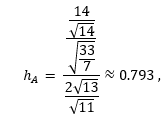


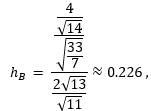
נעדכן ציוני HUB

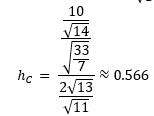
|  |  |
| --- | --- |
| Hub | initialization |
|  | A |
|  | B |
|  | C |

ננרמל:









**איטרציה 3**

|  |  |  |
| --- | --- | --- |
| Authority | Hub | initialization |
| 0.615 | 0.793 | A |
| 0.615 | 0.226 | B |
| 0.492 | 0.566 | C |

|  |  |
| --- | --- |
| Hub | initialization |
| 1.539 | A |
| 1.539 | B |
| 1.019 | C |









נעדכן ציוני HUB

|  |  |
| --- | --- |
| Hub | initialization |
| 1.704 | A |
| 0.424 | B |
| 1.280 | C |









התכנסנו לאחר 3 איטרציות ל AUTHORITY ו- HUB וקיבלנו את הציונים הסופיים.

|  |  |  |
| --- | --- | --- |
| Authority | Hub | Final |
| 0.640 | 0.784 | A |
| 0.640 | 0.195 | B |
| 0.424 | 0.0589 | C |

שאלה 3

1. בקורס למדנו מגוון טכנולוגיות, מהמעניינות שלמדנו:

**Porter Stemmer:** למדנו את השימושיות שלו ונעזרנו בו עוד מפרויקטים שלנו בקורס מחשוב ענן.

**BeautifulSoup**: ספרייה לעיבוד וניתוח תוכן HTML/XML, שבה השתמשנו לשליפת מידע מתוך מבנה ה-HTML של הדפים שסרקנו באתר.  
**pandas**: ספריית Python ששימשה אותנו לניתוח ועיבוד נתונים. באמצעותה בנינו DataFrame מכל המידע שנאסף, ביצענו פעולות עיבוד נתונים, ושמרנו את הנתונים בקובץ CSV. השימוש בה אפשר לנו ליצור סקריפט סריקה מתקדם, לעבד את הנתונים שנאספו, ולהפיק תובנות מעמיקות מהאתר.

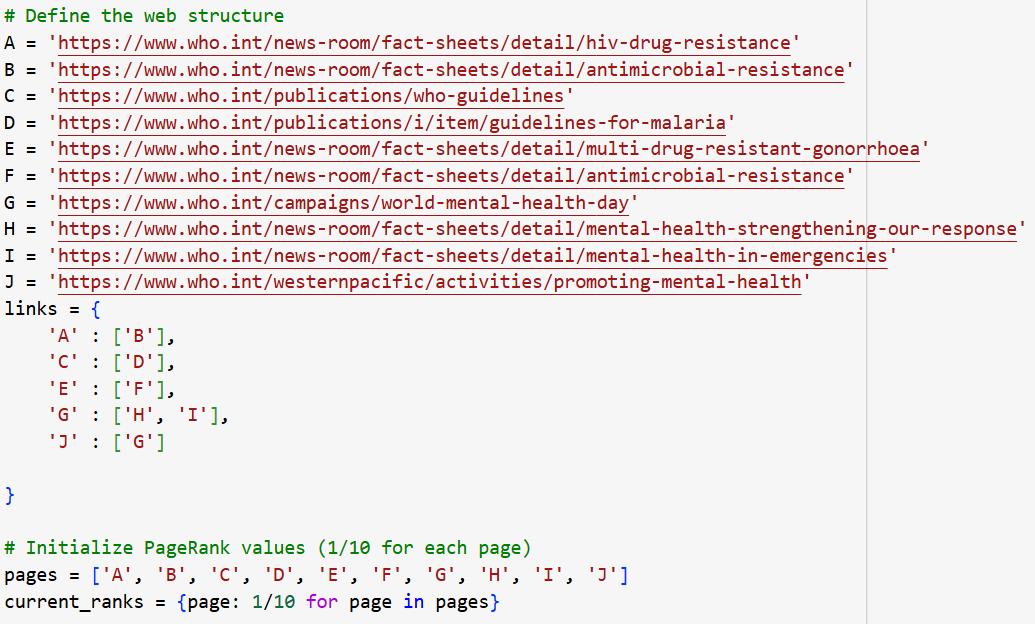
1. השאילתות רצו במשך 327 שניות

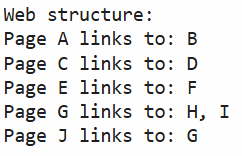
זמן הריצה תלוי במגוון דברים. ראשית, במהירות השרת המארח את האתר. שנית, הריצה תלוי בכמות הדפים שרצנו עליהם ועל כמות הדפים שיש בכל עמוד, כלומר ככל שהזחלן יגיע עמוק יותר, כך יגדל הזמן.  
אנו סבורים כי ניתן לשפר את החיפוש ע"י מספר אפשרויות:  
כאשר ישנה שאילתא שמבקשת מידע על מדינה או יבשת מסויימת, אפשר להתחיל את הזחלן מהעמוד בו יש מידע על היבשת ברצויה. לדוגמה בקישור הזה: <https://www.who.int/europe/home?v=welcome>  
יש מידע ספיציפית על מדינות באירופה.  
בנוסף, ניתן לשפר זמן זה ע"י אופטימיזציה של הזחלן, שמירת מידע במטמון ותהליכים שירוצו במקביל.

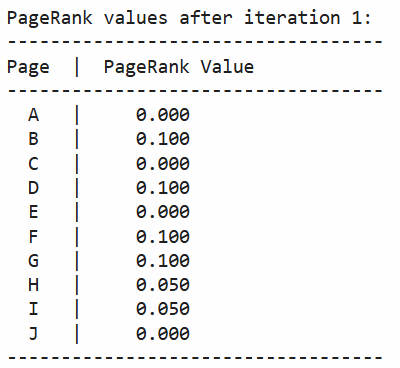
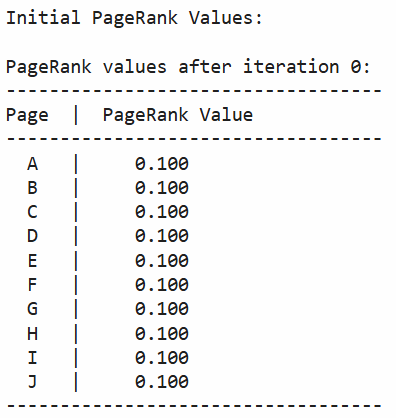
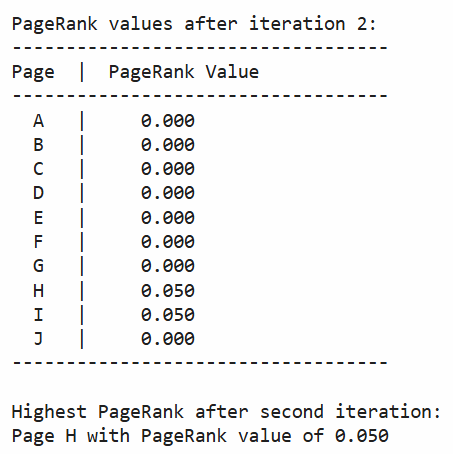
1. בדפים שהוחזרו קיימיים hubs. ניתן לראות שכל דף של כתבה על מחלה, מחקר וכו' ישנם קישורים המקושרים למחלה או הדברים הקשורים למחקר, לסימטומים קשורים ועוד. משום כך, ניתן לגשת לדפים המקושרים מכל דף של כתבה על מחלה. בנוסף, ניתן לראות כי קיימים גם authorities שהם דפי הקישורים של הכתבות. נראה כי לכל כתבה אפשר לגשת על ידי דפי מחלות או מחקרים אחרים.
2. לכל שאילתה הוחזרו 20 הדפים הרלוונטיים.

בדקנו את הקישוריות בין דפים אלה באמצעות אלגוריתם אשר עובר על הקישורים של כל דף ובודק האם קיימים ביניהם שאר הדפים. אספנו 10 קישורים מתוך התוצאות, השתמשנו באלגוריתם המוצע בקורס ליצירת גרף ולבסוף מחשב את ערך ה-PageRank.

להלן התוצאות:







1. מנתוני הרלוונטיות של המשתמשים עולה כי הם מתעניינים בדפים שבהם מצוי מידע עבור תרופות המונעות מחלות מראש. שאילתה מוצעת עבור משתמשים אלה:  
   Common disease prevention drugs

