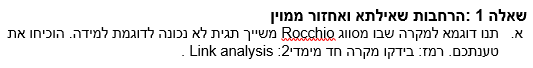
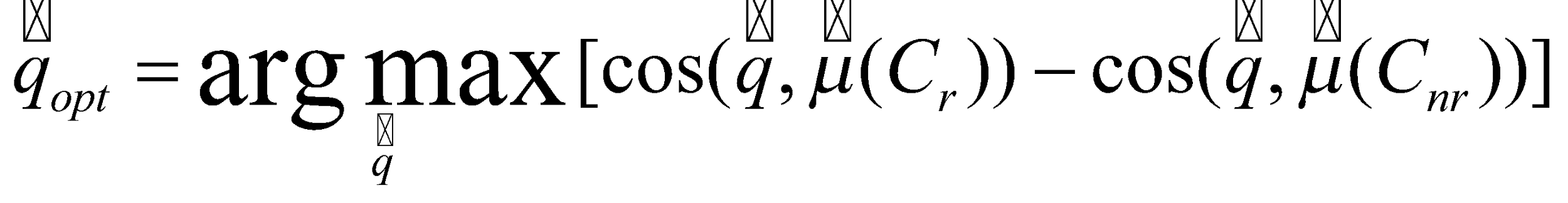
אחזור מידע – תרגיל בית 3 – חורף התשפ"ה  
Leonid Vyushkov 327312526  
Yotam Gilad 315819599  
Tomer Ben Lolo 318318987  
Yaakov Shitrit 304959422

פתרון שאלה 1:

הקדמה:  
באלגוריתם Rocchio, הרעיון המרכזי הוא למדוד את הדמיון בין השאילתה החדשה לבין מרכזי הכובד (centroid) של הקטגוריות השונות. השאילתה משויכת לקטגוריה שמרכז הכובד שלה הוא הקרוב ביותר לשאילתה במרחק קוסינוסי:  
(נציין שיש מרחקים נוספים כמו אוקלידי, מנהטן, jaccard ועוד, תלוי במטרת בסוג הסיווג שמעניין אותנו)  
בחישוב רב מימדי אנו מחשבים את קוסינוס הזווית שבין השאילתה לבין וקטור ה-centroid של המסמכים הרלוונטיים עבור הקטגוריה הנבדקת (הוקטור הזה הוא וקטור התכונות [לפי גודל המימד] כאשר כל תכונה היא ממוצע רכיב אותה התכונה בין כל המסמכים הרלוונטיים) ולאחר מכן אנחנו מחסירים את התוצאה הנ"ל בקוסינוס הזווית שבין השאילתה לוקטור ה-centroid של המסמכים הלא רלוונטיים עבור הקטגוריה הנבדקת. לאחר חישוב זה עבור כל הקטגוריות הקיימות אנחנו נשייך את השאילתה למרכז הכובד הקרוב ביותר לשאילתה.

ניגש לפתרון:  
על מנת להמחיש טעות בסיווג ניעזר ברמז המנחה אותנו להמחיש את הבעיה באמצעות מקרה חד מימדי (כלומר, במקרה שבו יש רק תכונה אחת, כך שאין לנו ווקטור לחישוב זווית הקוסינוס בינו לבין המסמכים הרלוונטיים והלא רלוונטיים ולכן החישוב הוא ממוצע רכיב התכונה עבור כל מסמך ומסמך).  
(לפשטות ההמחשה ניקח מקרה פרטי שבו יש 3 מסמכים רלוונטיים ו-3 מסמכים לא רלוונטיים לכל קטגוריה)  
טעות עשויה להתרחש כאשר מרכז הכובד של קטגוריה מסוימת קרוב באותה המידה למרכז הכובד של קטגוריה אחרת, או במקרה הקיצוני יותר (המקרה הרגיל) שבו ישנם כמה מימדים, יכול להיות אפילו שהמסמך המיועד לסיווג יהיה קרוב יותר למרכז הכובד של הקטגוריה הלא נכונה.  
מקרה כזה קל להמחיש באמצעות בדיקת מקרה חד מימדי (שיש בו רק תכונה אחת).  
נניח מרחב חד-ממדי שבו ישנן שתי קטגוריות:

קטגוריה A מכילה את המסמכים: {1,2,3)  
קטגוריה B מכילה את המסמכים: {7,8,9}

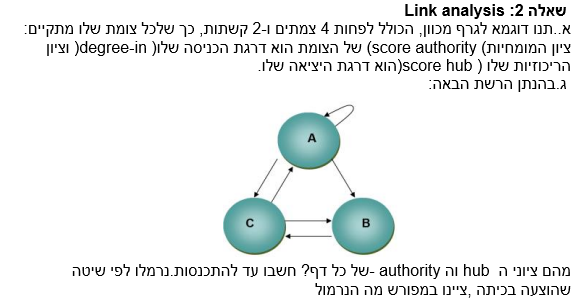
#### **חישוב מרכזי הכובד:**

* מרכז כובד של קטגוריה A:
* מרכז כובד של קטגוריה B:

כעת, נניח שיש לנו מסמך x=5 שאנחנו מעוניינים לסווג לקטגוריה.

* מרחק לקטגוריה A:
* מרחק לקטגוריה B:

במקרה זה, Rocchio עשוי לטעות בתוצאה מכיוון שהמרחקים זהים (ההפרש בין המסמכים הרלוונטיים ללא רלוונטיים בכל קטגוריה שווה ל-0) ולכן הוא יבחר בקטגוריה כלשהי באופן אקראי ומשום כך הוא עלול לבחור בקטגוריה הלא נכונה.  
מהדוגמה ניתן גם להבין שהבעיה רק הולכת וגדלה במקרים רב מימדיים.



פתרון שאלה 2:

**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.

הגרף המכוון



לפי הבקשה ניתן לראות כי יש לפחות 4 צמתים ו2 קשתות

**Initially:**

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

**After 1 iteration:**

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

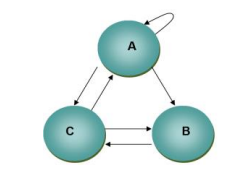
**After 2 iteration**

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

לפי הגרף שבנינו ניתן לראות שכבר מהאיטרציה השנייה אין שינוי ולכן קיבלנו:

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

 ב.בהנתן 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 1 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 2 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 3 iteration:**

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

**After 4 iteration:**

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

**After 5 iteration:**

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

ניתן לראות כי לאחר האיטרציה ב5 יש המערכת מתכנסת.

**שאלה 3**

1. רשימת הטכנולוגיות השונות שהשתמשנו בהם בפרויקט:

**Beautifulsoup** ספריייה בפייתון שמשמשת לחילוץ נתונים מדפי אינטרנט של קבצי HTML ו- XML.

**RAG-**משלב אחזור מידע עם יצירת טקסט כדי לשפר את הביצועים של מודלים לשוניים באמצעות שילוב מקורות מידע חיצוניים.

**PageRank-** כלי לבדיקת קישורים בין דפים ועל פיהם לדרג את חשיבות הדף

2.

זמן ריצת השאילתות :   
Nut Mix discount - 1:52 minute  
Snack sample -1:40 minute  
L T-Shirt - 1:24 minute

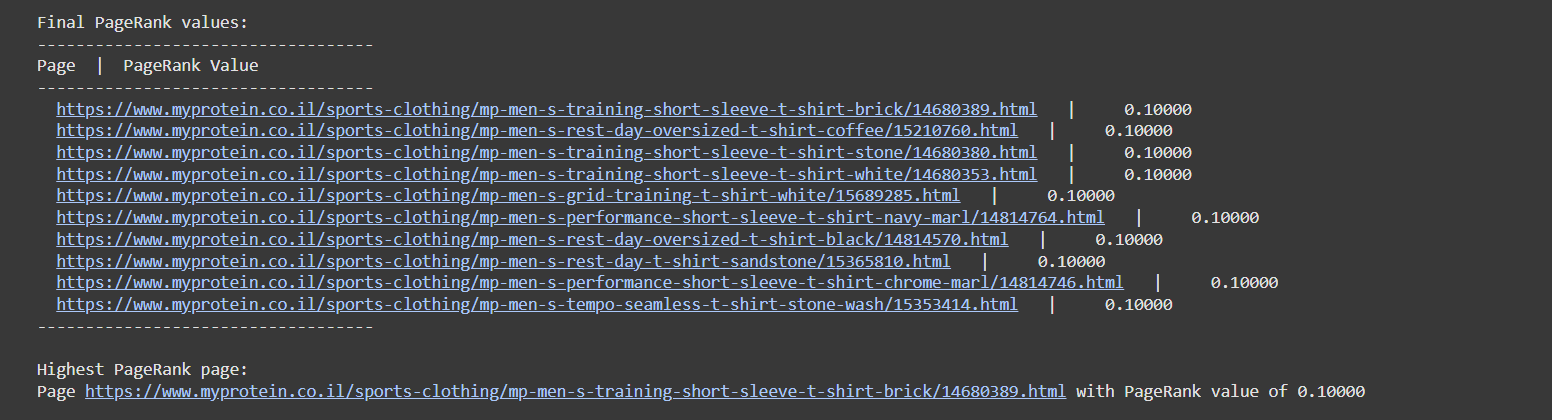
יש את זמן איסוף תוכן האתר שהוא לוקח את רוב הזמן וגם מערכת RAG שמשתמשת ב API לוקחת יחסית הרבה זמן במהלך עיבוד נתונים ומציאת התשובה.

ניתן לשפר את זמן הריצה עם יצירת מילון או קטגוריות האתר ובכך להקטין את זמן שאיבת נתונים .   
מבחינת עיבוד הנתונים אפשר להשתמש במערכת לוקאלית כמו LLAMA 3.

3.

לאחר שלוש השאילתות שהרצנו ובנוסף גם שהסתכלנו באתר , בעקבות הדפים שהוחזרו שיש להם קישור מדפים אחרים גם אליהם וגם הם מוציאים קישורים לדפים אחרים, אז קיימים Hubs ו Authorities לכל אחד מהדפים שהוחזרו.

4.



5.  
  
 משתמש א:   
סימן את כל העמודים כרלוונטיים חוץ מ <https://www.myprotein.co.il/sports-clothing/mp-men-s-rest-day-oversized-t-shirt-black/14814570.html>   
משתמש ב:   
סימן את כל העמודים כרלוונטיים לשאילתה.

בהתבסס על המשוב, ניתן לשפר את השאילתות על ידי הוספת תיאור מורחב של רלוונטיות אישית למשתמשים, כדוגמא במקרה זה עבור חולצה בגדול L ,ניתן יהיה להוסיף צבע חולצה,סוג בד,ובהנחה או שלא . שינוי זה יאפשר התאמה טובה יותר של תוצאות החיפוש להעדפות המשתמשים ויכול לשפר את תוצאות הדפים שחוזרים.

[קישור לריפו](https://github.com/MrLegiongame/Royal-Searchers/tree/main/HW3)  
  
[קישור לקוד](https://colab.research.google.com/drive/1ANLp0XIG15zbQVCFJKI7jdpq7glB8UkK?usp=sharing)

[קישור לאתר](https://hw3-98go.onrender.com)