**אחזור מידע – פרויקט להגשה**

**מגישים:**

**גל פולק 208638015**

**שרון מור 207470329**

**מיכל גינת 208111146**

**הקדמה**

בתחילת הקורס כאשר נודע לנו שאנחנו צריכים לבחור אתר עליו נבצע את פרויקט הזחלן, היה ברור לנו שנבחר באתר העוסק בתחום הטניס מכיוון שזה נושא אהוב על ידי כל חברי הקבוצה. האתר הנבחר בשלב הראשון היה <https://www.itftennis.com/en/> . בחרנו בו מכיוון שהיה נראה עדכני, מסודר ובעיקר מכיוון שראינו שחסר בו מידע עליו נוכל לעשות שאילתות בצורה יעילה ולבנוסף לנסיון ביצירת זחלן, נקבל גם ערך מוסף מהאתר עצמו.  
לכן, את החלק הראשון של הפרויקט ביצענו על אתר זה, שלפנו מידע מעניין והצגנו נתונים בצורה נוחה שלא ניתנת באתר, וכך יצרנו את האתר הבא: <https://sharonmor.github.io/Query-Kings---A/>  
עם המשך הקורס, הבנו כי המהות העיקרית של זחלן, חיפוש דפים באופן דינאמי באתר, פוספסה בחלקו הראשון של הפרויקט וכי לא ניתן להמשיך לעבוד על האתר שבחרנו מהסיבה שחיפוש בתוכו מחזיר תוצאות גרועות מאוד, כלכך גרועות שאפילו אינן דפים באתר אלא ממש קבצי pdf שניתן להוריד:  
A screenshot of a computer

Description automatically generated





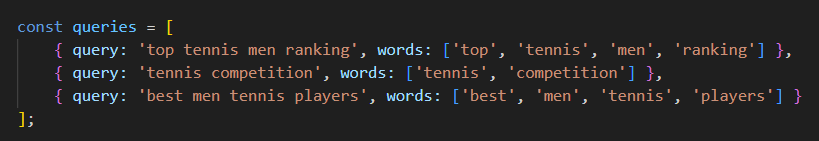
לבסוף הוחלט לשנות את האתר עליו נבצע חיפוש ובחרנו באתר[**www.espn.com**](http://www.espn.com) מכיוון שמכיל את כל הדברים אותם חיפשנו וגם קל לבצע בו חיפוש ולכן מהווה קרקע פורייה עבור זחלן.

1. **מהו תחום העיסוק המרכזי של האתר? מהו המידע הזמין למשתמשי האתר? ענו בפסקה אחת. צרפו את הקישור לאתר.**

האתר שבחרנו הוא ESPN.

קישור לאתר שבחרנו: <https://www.espn.com/>

ESPN היא חברת מדיה מובילה המתמחה באספקת תוכן ספורט מקיף על פני פלטפורמות שונות. האתר מציע מגוון רחב של שירותים למשתמשים בענפי ספורט מרכזיים כמו כדורגל, כדורסל, בייסבול, טניס ועוד. משתמשים יכולים לגשת לעדכוני תוצאות משחקים בשידור חי, לקרוא מאמרים מעמיקים, לצפות במיטב קליפים, לצפות בסטטיסטיקות נתונים עבור קבוצות ושחקנים ועוד.

**2. רשמו שלוש שאילתות מעניינות שהייתם רוצים לקבל עליהן תשובה באתר, והאתר אינו עונה עליהן כעת. עבור כל שאילתא כזו, ציינו מהם פרטי המידע הנדרשים לצורך מענה על השאילתא.**השאילתא העיקרית אותה רצינו לחקור היא לקבל ולהציג את שחקני הטניס הטובים בעולם – ‘top tennis men ranking'. את שאילתא זו חקרנו ממספר כיוונים, ראינו כי האתר לא מציג את המידע בצורה מספקת ועניין אותנו לדעת כמה האתר, espn שם במרכז את שאילתא זו, עניין אותנו האם תוצאות החיפוש של שאילתא זו יהיו מספקות עבורנו והאם עבור אחרים. סביב שאילתא זו, בחרנו בשתי שאילתות נוספות tennis competition, best men tennis players, כדי שוב להבין את התמונה המלאה ואת התתייחסות של האתר לתוכן השונה של השאילתא.  


**3. בנו זחלן המחזיר את התוצאות לשאילתות שהגדרתם. מומלץ (אך לא חובה) לבנות את הזחלן בפייתון. הציגו את קוד הזחלן.**

**קוד זחלן:** [**https://github.com/SharonMor/Query-Kings---A/blob/main/HW2/HW2.js**](https://github.com/SharonMor/Query-Kings---A/blob/main/HW2/HW2.js)

**4. רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.**

הטכנולוגיות בהן השתמשנו הן Puppeteer עבור ביקור בדפים ושאיבת המידע הרלוונטי, Javascript כשפת הריצה וNodeJs כסביבת הרצה עבור JS. בנוסף השתמשנו בספריית ExcelJS עבור יצירת ועריכת קבצי Excel.

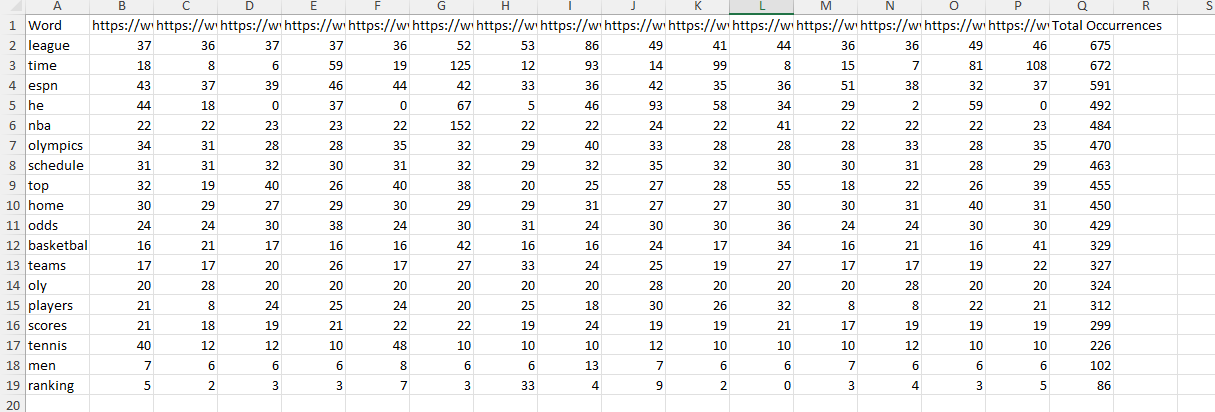
**5. כמה זמן רצו השאילתות שלכם?במה זה תלוי? האם לדעתכם ניתן לשפר זמן זה?**

סך השאילתות לקחו 7 דקות ו40 שניות, כלומר כל שאילתא לקחה לpuppeteer כלומר עבור כל שאילתא כ 2 דקות ו33 שניות. הזמן הארוך הנ"ל נובע מכך שלא שמנו דגש על סיבוכיות וזמן ריצה, נתנו הרבה זמני "אקסטרא" בהם חיכינו מספר שניות לדפים להיטען וערכנו מספר לולאות שאנו סבורים שניתן לשפר.

לדעתנו ניתן לשפר את זמן הריצה עם שימוש בטכנולוגיות אחרות, בחרנו להשתמש בpuppeteer on JS וזהו לא הכלי האידיאלי עבור זמני ריצה. בנוסף, פתיחת העמודים בצורה מקבילית תקצר באופן גדול את זמן הריצה.

**6. בנו inverted index ל- 15 המילים הנפוצות שחזרו (מספיק להתיחס ל -20 הדפים הראשונים שחזרו)**

**Inverted index for the first query, each column is another page.  
words on the first column are 15 most commonly used words + query words**

****

**7. בחרו את אחת השאילתות, וחשבו tfIdf של המושגים בשאילתא.**

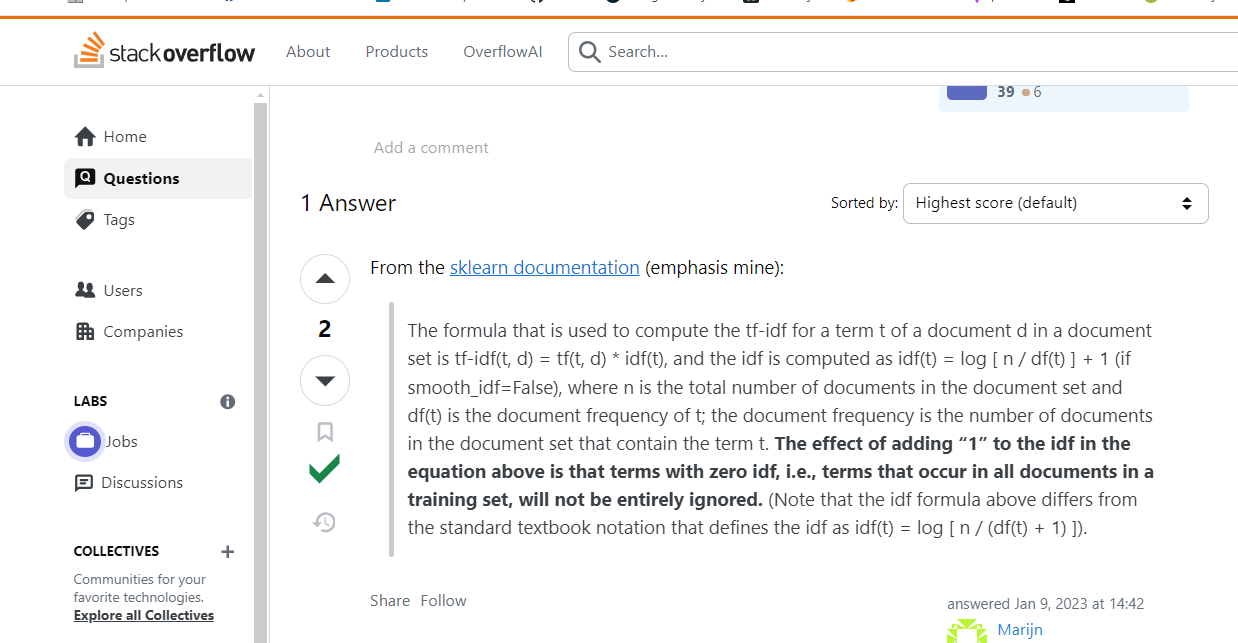
עבור השאילתא הראשונה, ביצענו tfidf עבור כל דף. את חישוב התוצאות ביצענו באמצעות סקריפט במהלך ריצת הזחלן. קוד לסקריפט ניתן לראות פה:[**https://github.com/SharonMor/Query-Kings---A/blob/main/HW2/HW2.js**](https://github.com/SharonMor/Query-Kings---A/blob/main/HW2/HW2.js)

**A screenshot of a computer

Description automatically generated**

הערה חשובה: בquery הנ"ל היו 2 מילים שחזרו בכל הדפים עקב מילה שכיחה (top) ומילה שהופיעה בתפריט, אשר הופיע בכל אחד מהדפים (tennis). עקב כך קיבלנו כי כל התוצאות שחזרו היו 0 בעקבות חישוב הidf של העמוד :

Math.log((documents.length + 1) / (documentsWithWord + 1)) = Math.log(1) = 0

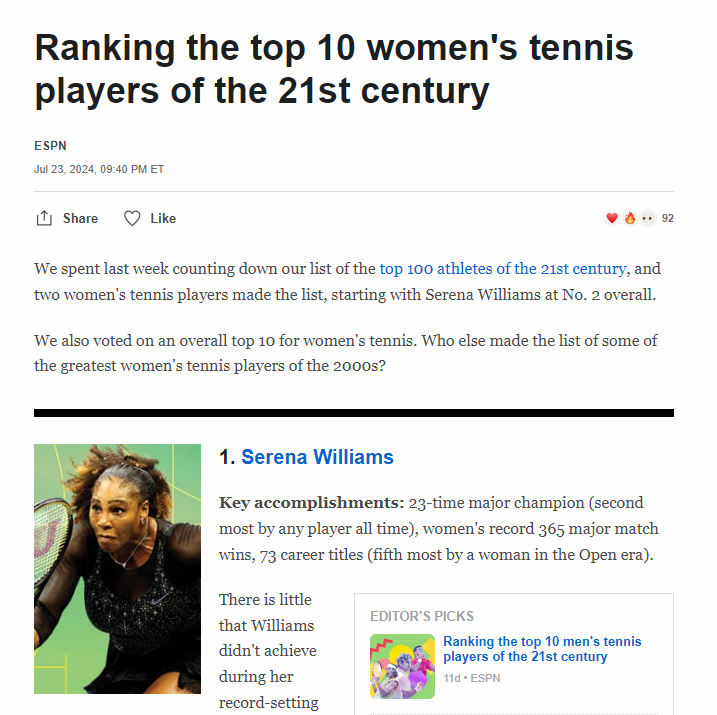
כלומר, TF \* 0 = 0.  
לא רצינו שכל התוצאות יתאפסו ולא נוכל להפיק מידע מדבר זה, לכן הוספנו 1 לכל תוצאת IDF ע"מ לתת ביטוי לחישוב הTF של מילת הQUERY ובכך להפיק כיוון מסוים לכל מילה בכל מקרה שיהיה. את הרעיון להוספת האחד מצאנו בstackoverflow:  
לאחר תוצאות ראינו כי 2 התוצאות בעלות הניקוד הגבוה ביותר הן גם הרלוונטיות ביותר.

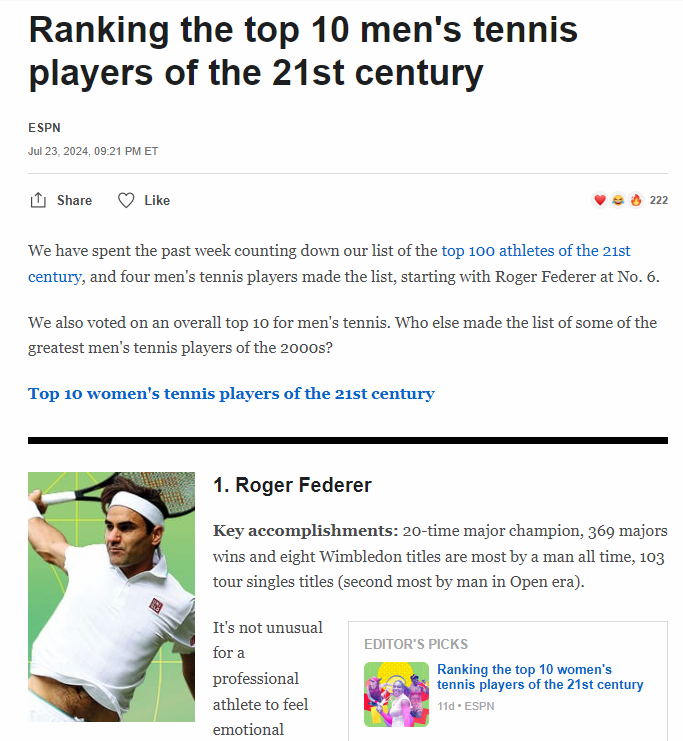
**8. האם בדפים שהוחזרו קיימים hubs? Authorities? נמקו.**

בדפים שהוחזרו קיימים hubs ו-authorities.  
ניתן לדוגמא לראות שבדףRanking the top 10 women's tennis players of the 21st century שהוחזר, יש הפנייה לדף אחר שגם הוחזר   
Ranking the top 10 men's tennis players of the 21st century וההפך.

כלומר, ישנה קישוריות דו-כיוונית בין הדפים האלו.

אפשר לראות שיש לא מעט לינקים בעמוד שמובילים לעמודים אחרים (hubs) וההפך (authorities)- לינקים שנכנסים לעמוד הרצוי. אפשר לראות את זה בדוגמא הבאה:





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

ביצענו חישוב Page rank עבור 10 הדפים שהוחזרו מהשאילתא: top 10 tennis men ranking. ביצענו את החישוב בצורה ידנית- עברנו על כל דף (מעמודה שמאלית) וראינו לאיזה דף אחר הוא מפנה. מוצגת כאן מטריצה של 10 העמודים (מצורף לינק לכל אות שמייצגת עמוד מסויים) שהוחזרו כאשר ריבוע אדום מייצג שלא קיימת הפנייה לדף אחר וריבוע ירוק מייצג שכן קיימת הפנייה לדף אחר, ריבוע שחור לא רלוונטי (הפנייה עצמית).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | [A](https://www.espn.com/mma/story?id=24067525&_slug_=mma-pound-pound-rankings) | [B](https://www.espn.com/tennis/story?id=40581478&_slug_=ranking-top-10-men-tennis-players-21st-century) | [C](https://www.espn.com/boxing/story?id=40588404&_slug_=boxing-ranking-top-10-men-boxers-2000) | [D](https://www.espn.com/tennis/story?id=40581041&_slug_=ranking-top-10-women-tennis-players-21st-century) | [E](https://www.espn.com/fantasy/football/story?id=40196245&_slug_=2024-fantasy-football-rankings-ppr-field-yates-qb-rb-wr-te) | [F](https://www.espn.com/mma/story?id=40588557&_slug_=ranking-top-10-men-mma-fighters-2000) | [G](https://www.espn.com/boxing/story?id=40590132&_slug_=boxing-ranking-top-10-women-boxers-2000) | [H](https://www.espn.com/mens-college-basketball/story?id=40632246&_slug_=100-days-2024-25-mens-college-basketball-season-top-stories-players-games) | [I](https://www.espn.com/soccer/story?id=40490059&_slug_=ranking-top-25-men-soccer-players-21st-century) | [J](https://www.espn.com/espn/story?id=40446224&_slug_=top-100-athletes-21st-century) |
| [A](https://www.espn.com/mma/story?id=24067525&_slug_=mma-pound-pound-rankings) - id=24067525 | - |  |  |  |  |  |  |  |  |  |
| [B](https://www.espn.com/tennis/story?id=40581478&_slug_=ranking-top-10-men-tennis-players-21st-century) - id=40581478 |  | - |  |  |  |  |  |  |  |  |
| [C](https://www.espn.com/boxing/story?id=40588404&_slug_=boxing-ranking-top-10-men-boxers-2000) - id=40588404 |  |  | - |  |  |  |  |  |  |  |
| [D](https://www.espn.com/tennis/story?id=40581041&_slug_=ranking-top-10-women-tennis-players-21st-century) - id=40581041 |  |  |  | - |  |  |  |  |  |  |
| [E](https://www.espn.com/fantasy/football/story?id=40196245&_slug_=2024-fantasy-football-rankings-ppr-field-yates-qb-rb-wr-te) - id=40196245 |  |  |  |  | - |  |  |  |  |  |
| [F](https://www.espn.com/mma/story?id=40588557&_slug_=ranking-top-10-men-mma-fighters-2000) - id=40588557 |  |  |  |  |  | - |  |  |  |  |
| [G](https://www.espn.com/boxing/story?id=40590132&_slug_=boxing-ranking-top-10-women-boxers-2000) - id=40590132 |  |  |  |  |  |  | - |  |  |  |
| [H](https://www.espn.com/mens-college-basketball/story?id=40632246&_slug_=100-days-2024-25-mens-college-basketball-season-top-stories-players-games) - id=40632246 |  |  |  |  |  |  |  | - |  |  |
| [I](https://www.espn.com/soccer/story?id=40490059&_slug_=ranking-top-25-men-soccer-players-21st-century) - id=40490059 |  |  |  |  |  |  |  |  | - |  |
| [J](https://www.espn.com/espn/story?id=40446224&_slug_=top-100-athletes-21st-century) - id=40446224 |  |  |  |  |  |  |  |  |  | - |

מצורפת טבלה עם ערכים שבהם חושב דירוג ה-pageRank בכל איטרציה (סה"כ 4 איטרציות)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration 0 | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 |
| PR(A) = 1 | PR(A) = 0.150 + 0.85(1.000/2) = 0.575 | PR(A) = 0.150 + 0.85(0.150/2) = 0.214 | PR(A) = 0.150 + 0.85(0.150/2) = 0.214 | PR(A) = 0.150 + 0.85(0.150/2) = 0.214 |
| PR(B) = 1 | PR(B) = 0.150 + 0.85(1.000/2) = 0.575 | PR(B) = 0.150 + 0.85(0.394/2) = 0.318 | PR(B) = 0.150 + 0.85(0.285/2) = 0.271 | PR(B) = 0.150 + 0.85(0.265/2) = 0.263 |
| PR(C) = 1 | PR(C) = 0.150 + 0.85(1.000/2) = 0.575 | PR(C) = 0.150 + 0.85(0.150/2) = 0.214 | PR(C) = 0.150 + 0.85(0.150/2) = 0.214 | PR(C) = 0.150 + 0.85(0.150/2) = 0.214 |
| PR(D) = 1 | PR(D) = 0.150 + 0.85(0.575/2) = 0.394 | PR(D) = 0.150 + 0.85(0.318/2) = 0.285 | PR(D) = 0.150 + 0.85(0.271/2) = 0.265 | PR(D) = 0.150 + 0.85(0.263/2) = 0.262 |
| PR(E) = 1 | PR(E) = 0.150 = 0.150 | PR(E) = 0.150 = 0.150 | PR(E) = 0.150 = 0.150 | PR(E) = 0.150 = 0.150 |
| PR(F) = 1 | PR(F) = 0.150 = 0.150 | PR(F) = 0.150 = 0.150 | PR(F) = 0.150 = 0.150 | PR(F) = 0.150 = 0.150 |
| PR(G) = 1 | PR(G) = 0.150 = 0.150 | PR(G) = 0.150 = 0.150 | PR(G) = 0.150 = 0.150 | PR(G) = 0.150 = 0.150 |
| PR(H) = 1 | PR(H) = 0.150 = 0.150 | PR(H) = 0.150 = 0.150 | PR(H) = 0.150 = 0.150 | PR(H) = 0.150 = 0.150 |
| PR(I) = 1 | PR(I) = 0.150 = 0.150 | PR(I) = 0.150 = 0.150 | PR(I) = 0.150 = 0.150 | PR(I) = 0.150 = 0.150 |
| PR(J) = 1 | PR(J) = 0.150 + 0.85(0.575/2 + 0.575/1 + 0.394/2 + 0.150/2 + 0.150/2 + 0.150/1) = 1.306 | PR(J) = 0.150 + 0.85(0.318/2 + 0.214/1 + 0.285/2 + 0.150/2 + 0.150/2 + 0.150/1) = 0.843 | PR(J) = 0.150 + 0.85(0.271/2 + 0.214/1 + 0.265/2 + 0.150/2 + 0.150/2 + 0.150/1) = 0.815 | PR(J) = 0.150 + 0.85(0.263/2 + 0.214/1 + 0.262/2 + 0.150/2 + 0.150/2 + 0.150/1) = 0.810 |

הנוסחא שהשתמשנו בה כדי לחשב את הpage rank היא:

תמונה שמכילה טקסט, גופן, צילום מסך, לבן

התיאור נוצר באופן אוטומטי כאשר d פאקטור הוא 0.85 במקרה שלנו.

ניתן לראות של-J יש את ה PageRank הגבוה ביותר כאשר J הוא העמוד Ranking the top 100 professional athletes since 2000   
כלומר הרבה לינקים ניגשים אליו (ניתן לראות גם במטריצה שצורפה תחילה, שבעמודה J יש הרבה ריבועים ירוקים, מה שמראה שלהרבה דפים יש קישורים לדף זה).

תמונה שמכילה קו, תרשים

התיאור נוצר באופן אוטומטי

מצורף גרף המתאר את הקישוריות בין 10 הדפים שהוחזרו מהשאילתא.

תוצאות סופיות לפי סדר הדירוג:

דירוג של J: 0.81

דירוג של B: 0.263

דירוג של D: 0.262

דירוג של A: 0.214  
דירוג של C: 0.214

דירוג של E: 0.15

דירוג של F: 0.15  
דירוג של G: 0.15  
דירוג של H: 0.15  
דירוג של I: 0.15

**10. הראו לשני משתמשים שונים את הדירוג מהסעיף הקודם, בקשו מהם לסמן relevance feedback .**

משתמש 1: בחר שהדפים הרלוונטים הם- B,D,J

משתמש 2: בחר שהדפים הרלוונטים הם- B

B- Ranking the top 10 men's tennis players of the 21st century

D- Ranking the top 10 women's tennis players of the 21st century

J- Ranking the top 100 professional athletes since 2000

עבור השאילתא המקורית top 10 tennis men ranking נקבל:

תמונה שמכילה טקסט, צילום מסך, פני אדם, אדם

התיאור נוצר באופן אוטומטי

**חשבו precision ו - recall.האם ניתן להציע שאילתא מותאמת על מנת לשפר תוצאות?**

סך הכל הוחזרו כ50 עמודים מהשאילתה הנ"ל.

המסמכים שנמצאו בעינינו כרלוונטיים:

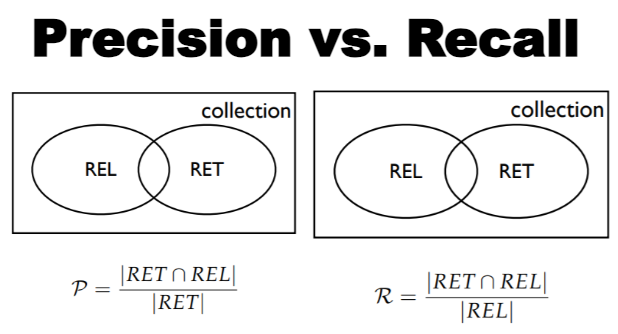
[Ranking the top 10 men's tennis players of the 21st century](https://www.espn.com/tennis/story?id=40581478&_slug_=ranking-top-10-men-tennis-players-21st-century%20)

[Ranking the top 100 professional athletes since 2000](https://www.espn.com/espn/story/_/id/40446224/top-100-athletes-21st-century)

[ESPN's top athletes of the 21st Century: Every sport ranking](https://www.espn.com/nfl/story/_/id/40618858/ranking-top-athletes-21st-century)

כלומר 3 מסמכים רלוונטים מתוך 50 מסמכים סך הכל.

נסתכל על ה10 דפים הראשונים שחזרו. מתוך ה10 דפים הראשונים שחזרו, 2 מהם רלוונטים.

לכן נקבל לפי החישוב:  


|RET inter ReL| = 2

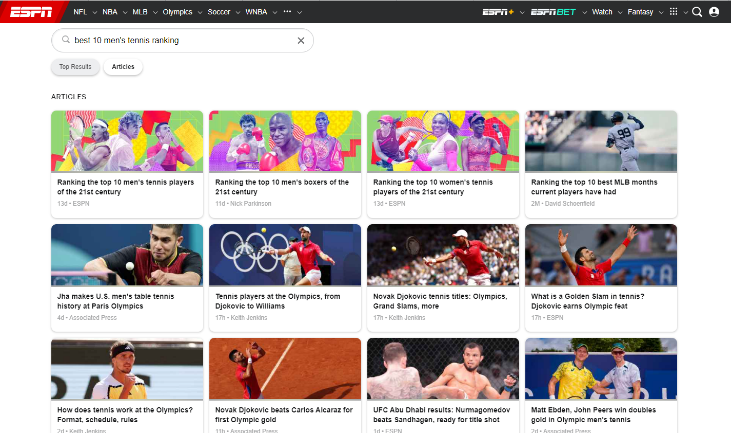
|RET| = 10

|REL| = 3

Precision = 2/10

Recall = 2/3

שאילתא חדשה: best 10 men's tennis ranking



ניתן לראות שע"י שינוי השאילתא, אנחנו מקבלים תוצאות שהן יותר רלוונטיות למה שהמשתמש מחפש.

**11. 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.**

קישור לאתר:[**https://sharonmor.github.io/IR-Project/**](https://sharonmor.github.io/IR-Project/)

קישור לריפו: [**https://github.com/SharonMor/Query-Kings---A**](https://github.com/SharonMor/Query-Kings---A)

**12. בתאריך 12.8 תציגו את הדף מסעיף 11, הכולל את סעיפים 1-10. תקבלו משוב מחבריכם. ענו בטבלה:**

|  |  |
| --- | --- |
| **איזה שינוי הוצע?** | **האם לדעתכם יש מקום לשיפור זה? אם כן - הסבירו כיצד ניתן לשפר. אם לא, נמקו** |
|  |  |
|  |  |

**13. מהו היו האתגרים בקורס / פרויקט כצוות?**

**יש להגיש קישור לריפו שלכם ב -GIT, הכולל את כל קבצי האתר, ותיקייה בשם project שבה קובץ וורד ובו מענה על השאלות.**

**ניתן לפנות אליי בכל שאלה.**

**בהצלחה!**