אחזור מידע – פרויקט סוף – חורף התשפ"ה

מגישים: ליאור צוקר 316375187 || נבו גוטליב 206234353 || רז קסל 206017725  
 || יוני אזרף 209459239

תרגיל בית 1

תרגיל בית 2

תרגיל בית 3

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

**Answer:**

The website offers users comprehensive information about NBA games, including schedules, results, highlights, and game analyses. It also provides detailed statistics, news, and updates about teams and players, as well as unique content such as interviews and behind-the-scenes stories. Users can browse information on subscription plans for viewing personalized content, along with details on device and platform support.

<https://www.nba.com/games>

1. 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.

**Answer:**

**Player performance trends over the last 10 games  
Query: “What are the average points, assists, and rebounds of a certain player over his last 10 games?”**

**Issue on NBA.com: Player statistics are generally displayed per game. Users must manually calculate averages by looking at multiple game summaries.**

**Solution:**

**Extract individual game stats for the player.**

**Compute averages across the last 10 games.**

**Display a clear table or chart summarizing the data (e.g., points, assists, rebounds).**

**Head-to-head comparison between teams  
Query: “How did two teams perform against each other in their last 5 encounters?”**

**Issue on NBA.com: Historical comparisons between teams are not centralized on the site. Users must look up each game separately.**

**Solution:**

**Retrieve historical game data (results, dates, statistics) for the selected teams.**

**Provide a comparison that includes:**

**Wins/losses for each matchup.**

**Average points per team.**

**Key players in those games.**

**Free-throw shooting percentages  
Query: “Top 5 NBA teams by free-throw percentage over each team’s last three games.”**

**Issue on NBA.com: Accurate team statistics appear on different links, making it difficult to view them all on a single page.**

**Solution:**

**Gather data from separate games for each team.**

**Calculate average free-throw percentages over the last three games for each team.**

**Display the top 5 teams with the highest percentages.**

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

**Answer:**

**You can find our crawler code in our GitHub repository:**[**https://github.com/liorZucker11/Result-Wizards/tree/main/scrap-project**](https://github.com/liorZucker11/Result-Wizards/tree/main/scrap-project)

1. רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.

**Answer:**

**Selenium: An automation tool used for programming crawlers to automatically access and control web browsers to collect data from websites.**

**PrettyTable: A Python library that allows creating and displaying tables in a readable text format (console, files, or any other textual format), making it easier to style and organize data.**

**BeautifulSoup: A Python library for parsing and processing HTML and XML code, making it simpler and more efficient to extract information from web pages.**

**Flask: A lightweight Python framework for web application development, allowing the quick and simple creation of interactive interfaces.**

**NetworkX: A Python library for creating, analyzing, and manipulating graphs, with an emphasis on analyzing connections and statistics between nodes, especially suited for information retrieval.**

**Matplotlib: A Python library for creating visual graphs of data, offering flexibility and customization for various requirements.**

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

**Answer:**

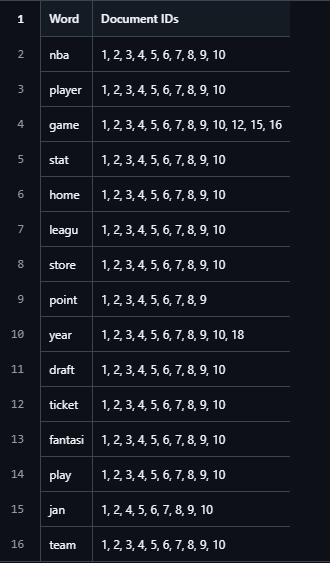
**Query A: “What are the average points, assists, and rebounds of a certain player over his last 10 games?” Runtime: 20.01 seconds.**

**Query B: “How did two teams perform against each other in their last 5 encounters?” Runtime: 51.37 seconds.**

**Query C: “Ranking all NBA teams by free-throw percentage over each team’s last three games.” Runtime: 212.40 seconds.**

**The runtime of the queries depends on the number of links the crawler traverses and the volume of data being retrieved. During execution, we also wait for the page to fully render before starting to extract data, so there is a deliberate delay factored into the runtime. We could optimize these wait times more effectively, instead of relying on a fixed preset delay.**

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

**Answer:  
The table below shows the words and their corresponding Document IDs:  
We corrected the table after the presentation and added stemming to the words that were found to exclude duplicates.  
By doing that we are being able to receive more accurate results.**

1. בחרו את אחת השאילתות, וחשבו tfIdf של המושגים בשאילתא.

**Answer:**

**We calculate the term frequency–inverse document frequency (tfIdf) for the chosen query, considering the frequency of each term in the query and across all documents.**

***Chosen Query 1:* “What are the average points, assists, and rebounds of a certain player over his last 10 games?”**

**Below are four tables illustrating different stages of data collection and the TF–IDF computation process:**

### A screenshot of a computer Description automatically generated**Table 1: Scraped Player Data**

### **Table 2: Raw Term Frequencies, DF, and IDF**

A screenshot of a computer

Description automatically generated

**Table 3: TF Values per Page**

**A screenshot of a computer

Description automatically generated**

**Table 4: TF-IDF Table for All Terms (3 Decimal Points)**

A screenshot of a computer

Description automatically generated

**Finally, by multiplying each term’s TF by its IDF, we obtain the TF–IDF score for each term in each page.**

1. האם בדפים שהוחזרו קיימים hubs? Authorities? נמקו.

**Answer:  
Yes, the pages that were returned from the query are all Authorities.  
Those pages contain significant data over each player and stats over the last games. The query doesn’t returns any Hubs because we are in a directly way go to a specific URL without using Hubs for navigation**

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

**Answer:**

**Below is a list of 10 links used to calculate PageRank:**

**https://www.nba.com/players → 0.329**

**https://www.nba.com/player/1630166/deni-avdija → 0.067**

**https://www.nba.com/stats/player/1630166 → 0.043**

**https://www.nba.com/player/1630166/deni-avdija/bio → 0.121**

**https://www.nba.com/player/1630166/deni-avdija/videos → 0.121**

**https://www.nba.com/team/1610612757/blazers → 0.043**

**https://www.nba.com/team/1610612757/schedule → 0.043**

**https://www.nba.com/stats/team/1610612757/traditional → 0.080**

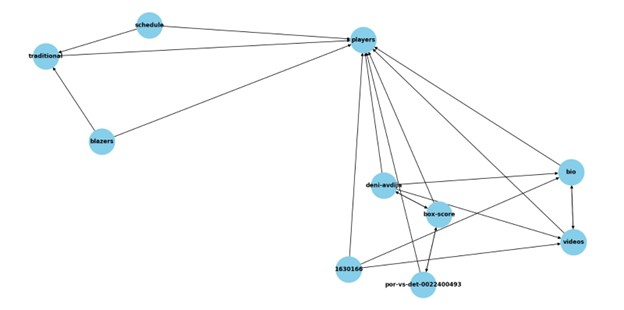
**https://www.nba.com/game/por-vs-det-0022400493 → 0.067**

**https://www.nba.com/game/por-vs-det-0022400493/box-score → 0.086**

**These calculations were performed using the NetworkX library. We created a directed graph in which each node represents a link, and used a script to locate href references among these nodes on each page. When a link to another page was found, we added a corresponding edge to the graph.**

**We performed PageRank with a custom alpha value (a = 0.85) using NetworkX.**

In addition, to improve page rankings, it is possible to pre-filter and catalog the returned pages into groups according to topics.   
This way the ratings will be more accurate after additional filtering.



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

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

**Answer:**

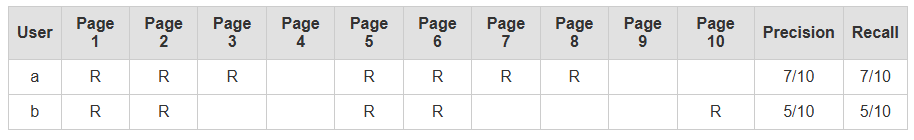
**According to the PageRank results, one page stands out with a relatively high score –**[**https://www.nba.com/players**](https://www.nba.com/players)**(0.329) – while the other pages are ranked with lower scores. To evaluate the relevance of the results, we asked two users to mark the results as relevant or not:**

**User A: Marked pages 1,2,3,5,6,7,8 as relevant.** <https://www.nba.com/players>  
 <https://www.nba.com/player/1630166/deni-avdija>  
 <https://www.nba.com/stats/player/1630166>  
 <https://www.nba.com/player/1630166/deni-avdija/videos>  
 <https://www.nba.com/team/1610612757/blazers>  
 <https://www.nba.com/team/1610612757/schedule>  
 <https://www.nba.com/stats/team/1610612757/traditional>

**User B: Marked pages 1,2,5,6,10 as relevant.** <https://www.nba.com/players>  
 <https://www.nba.com/player/1630166/deni-avdija>  
 <https://www.nba.com/player/1630166/deni-avdija/videos>  
 <https://www.nba.com/team/1610612757/blazers>  
 https://www.nba.com/game/por-vs-det-0022400493/box-score

**Based on this feedback, we can propose improving the queries by incorporating additional ranking parameters—for example, personal relevance (whether the content focuses on active players, personal stats, or game details). We could prioritize pages that offer general player information (e.g., “players”) and tailor them to user preferences by filtering out single-player-only pages.  
Such enhancements can improve the quality and relevance of the search results.**

The number of pages that the crawler returned is **10**, and we consider all the pages relevant.  
The crawler is directed to go to a specific URLs to search after the statistics so all the pages are relevant.  
So the precision and the recall in this case will be equal.



1. 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.  
     
   **https://result-wizards.onrender.com/**
2. 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 |
| שיפור זמן ריצה של השאילתא השלישית | הזחלן בשאילתא השלישית עובר בין המון דפים ולכן זמן הריצה ארוך יותר, אכן יהיה יעיל יותר לשפר את זמן הריצה ע"י שימוש בthreads וריצה במקביל. |
| למצוא פתרון לקטע של הSLEEP | אנו משתמשים בsleep לטובת הבטחת רינדור הדפים בתהליך הזחילה. זמן זה הוא קבוע בקוד ואכן ניתן לשפר אותו ולהתאים את זמן ההמתנה לצורה דינמית. |

|  |  |
| --- | --- |
| הוספת stemming לinverted index | אכן השינוי שהוצע יעיל ומדייק את התוצאות, בתגובה לכך אכן הוספנו stemming בקוד והתוצאות אכן היו יותר מדויקות! |

1. מהו היו האתגרים בקורס / פרויקט כצוות?

**Deployment Challenges:  
Deploying our project to a live site was one of the most annoying parts. We tried using Git Pages first, but it didn’t work well with our project, and we couldn’t figure out how to make it work. We also tried some other ways to deploy but had problems with those too. Finally, we used Render, and it worked. It was a big lesson about picking the right tools for deployment.**

**Scheduling Meetings:  
Since we are 4 busy team members, it was really hard to find times when everyone could meet. We all had different schedules with classes, work, and personal stuff, so planning meetings was a big problem. Sometimes, we had to work on tasks separately and communicate through messages, which made it harder to stay on the same page.**

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!