**404 Brain Not Found** - פרויקט מסכם

מגישים:

איתמר קראוס – 318304763

חי פדידה – 318302783

תומר לזרוביץ – 316377100

גיא פריאנטה – 206132250  
  
גיט:

<https://github.com/ChayFadida/scrapingPL>

אתר:

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

תשובות

1. אתר הפרמייר ליג משמש בעיקר כפלטפורמה המקוונת הרשמית של ליגת הכדורגל הבכירה באנגליה, ומציע סיקור מקיף של פעילות הליגה. העסק העיקרי של האתר סובב סביב החדשות האחרונות, עדכונים חיים ממשחקים, וניתוח מעמיק של קבוצות ושחקנים, כמו כן סטטיסטיקות מפורטות. בנוסף, האתר מציע תכונות אינטראקטיביות כמו כדורגל פנטזי, חנות למוצרים רשמיים, וגישה לכרטיסים למשחקים, מה שיוצר חוויה מרתקת לחובבי הכדורגל.

<https://www.premierleague.com/>

1. 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 ההכרזות על שחקן/מאמן החודש בין אפריל 2023 לאפריל 2024 שאינן מלונדון (השחקן / המאמן לא משחקים/מאמנים בקבוצה מלונדון).

מידע נדרש:

* + רשימת הקבוצות הלונדוניות
  + רשימת שחקנים והמידע הבא עליהם: ארץ מוצא, קבוצה, כמות פסים בכל הזמנים.
  + רשימת שחקנים, תפקיד בקבוצה, ארץ מוצא, כמות "רשת נקייה"
  + רשימת הקבוצות הלונדוניות.
  + רשימת הכרזות של מאמן החודש.
  + רשימת הכרזות של שחקן החודש.

1. קישור לזחלן:  
   <https://github.com/ChayFadida/scrapingPL/blob/master/scraper/plScraper.py>  
   קישור לשאילתות(שמפעילות את הזחלן):

<https://github.com/ChayFadida/scrapingPL/blob/master/scraper/queries.py>

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. זמני הריצה של השאילתות השונות (לפי הסדר בסעיף 2):
   1. sec95.85
   2. 94.5 sec
   3. 190.63 sec

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

1. 1. External -> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   2. Link -> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   3. League-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   4. Pl-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   5. Competitions-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   6. Premier-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   7. Cup-> [1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15]
   8. United-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   9. Official-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]
   10. Stats-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   11. Fpl-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   12. Manchester-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   13. Fa-> [1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15]
   14. Fantasy-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
   15. Club-> [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]

1. נבחר את שאילתה iii:  
   המושגים בשאילתה - Player, Award, Month, Manager, April, London

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IDF | D/Dfi | Dfi | P10 | P9 | P8 | P7 | P6 | P5 | P4 | P3 | P2 | P1 | Word |
|  |  | 91 | 11 | 6 | 6 | 11 | 7 | 7 | 13 | 13 | 9 | 8 | Player |
|  |  | 74 | 5 | 2 | 4 | 6 | 11 | 12 | 8 | 8 | 12 | 6 | Award |
|  |  | 124 | 9 | 4 | 15 | 14 | 14 | 15 | 12 | 12 | 17 | 12 | Month |
|  |  | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | London |
|  |  | 12 | 4 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 5 | April |
|  |  | 50 | 1 | 0 | 8 | 3 | 6 | 10 | 2 | 2 | 8 | 10 | Manager |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| W-10 | W-9 | W-8 | W-7 | W-6 | W-5 | W-4 | W-3 | W-2 | W-1 | Word |
| 27.06 | 14.76 | 14.76 | 27.06 | 17.22 | 17.22 | 31.98 | 31.98 | 22.14 | 19.68 | Player |
| 12.75 | 5.1 | 10.2 | 15.3 | 28.05 | 30.6 | 20.4 | 20.4 | 30.6 | 15.3 | Award |
| 20.88 | 9.28 | 34.8 | 32.48 | 32.48 | 34.8 | 27.84 | 27.84 | 39.44 | 27.84 | Month |
| - | - | - | - | - | - | - | - | - | - | London |
| 13.36 | 0 | 0 | 0 | 6.68 | 0 | 0 | 0 | 3.34 | 16.7 | April |
| 2.72 | 0 | 21.76 | 8.16 | 16.32 | 27.2 | 5.44 | 5.44 | 21.76 | 27.2 | Manager |

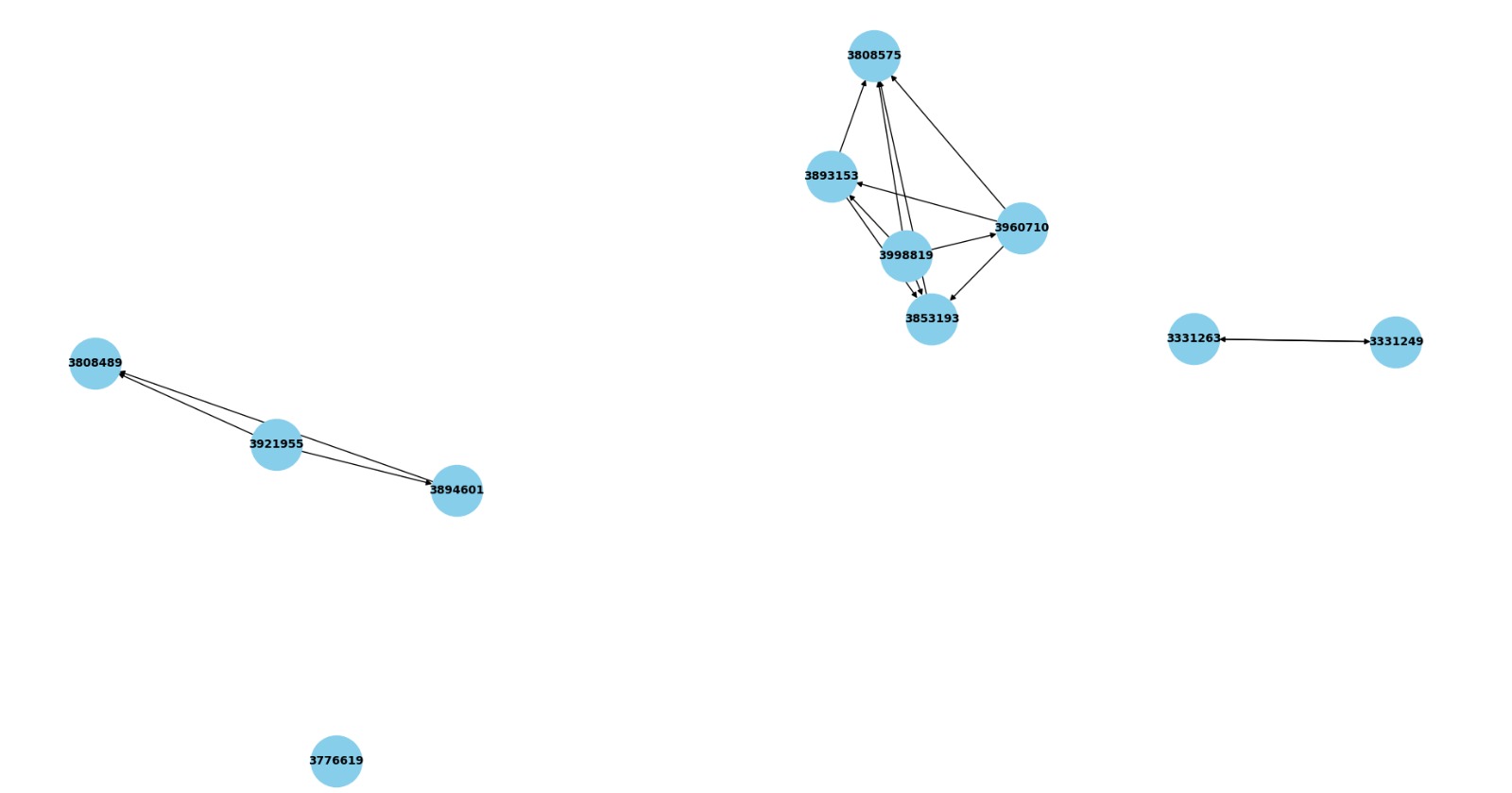
1. שתי השאילתות הראשונות מחזירות לינקים לעמודי השחקנים, כל עמוד הוא עצמאי ואינו מכיל קישור לעמוד שחקן אחר, לכן בשתי שאילתות אלו לא יהיה לנו Hubs & Authorities.  
   בשאילתה האחרונה אנו מחזירים עמוד חדשות על ההעברה של שחקן, בעמוד זה יש קישורים לחדשות אחרות לכן ייתכן ויהיו קיימים Hubs & Authorities.
2. נבחר את הדפים שהזחלן החזיר עבור השאילתה האחרונה(iii):

|  |  |
| --- | --- |
| Page Rank | Page Link |
| 0.034 | <https://www.premierleague.com/news/3998819> |
| 0.041 | <https://www.premierleague.com/news/3960710> |
| 0.034 | <https://www.premierleague.com/news/3921955> |
| 0.048 | <https://www.premierleague.com/news/3894601> |
| 0.053 | <https://www.premierleague.com/news/3893153> |
| 0.075 | <https://www.premierleague.com/news/3853193> |
| 0.089 | <https://www.premierleague.com/news/3808489> |
| 0.139 | <https://www.premierleague.com/news/3808575> |
| 0.034 | <https://www.premierleague.com/news/3776619> |
| 0.226 | <https://www.premierleague.com/news/3331263> |

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.

לינק לסקריפט: <https://github.com/ChayFadida/scrapingPL/blob/master/scraper/PageRankQ3.py>

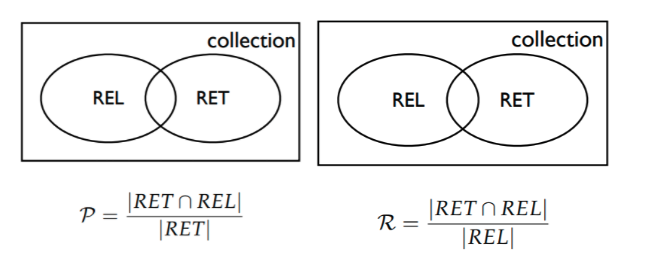
הצגת הגרף שהתקבל באמצעות matplotlib:



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:

* Total relevant – 12. יש סה״כ 12 תוצאות רלוונטיות, הכוללות שחקנים / מאמנים שבזמן שקיבלו את התואר שחקן/מאמן החודש שיחקו/אימנו בקבוצות מחוץ ללונדון.
* אופן בחירת הדפים הרלוונטים:
  1. משתמש a הינו איש מבוגר, מעריץ שרוף של כדורגל ושל הליגה האנגלית בפרט. משתמש זה ידע להבחין ולזהות כי חלק מתוצאות השאילתה הינם שחקנים שבעונה הרלוונטית לא שיחקו/אימנו בקבוצה בלונדון אך מופיעים כרגע כשייכים לקבוצה מלונדון כי עברו אליה הקיץ הזה.
  2. משתמש b הינו ילד, גם מעריץ שרוף של הליגה האנגלית אך עדיין אינו שולט בגיאוגרפיה של אנגליה ולכן לא זיהה כי שחקנים/מאמנים ממנצ׳סטר יונייטד לא שייכים לקבוצות מלנדון.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Recall | Precision | P10 | P9 | P8 | P7 | P6 | P5 | P4 | P3 | P2 | P1 |  |
|  |  |  | R | R | R | R |  | R | R | R | R | User A |
|  |  | R | R | R |  |  | R | R |  | R | R | User B |



בהתבסס על המשוב וחישוב הPrecision/Recall, ניתן לשפר את השאילתות על ידי התאמת קריטריונים נוספים לדירוג כגון הוספת פרמטרים של רלוונטיות אישית למשתמשים, כדוגמא החזר את כל ההודעות על שחקן/מאמן החודש בין אפריל 2023 לאפריל 2024 עבור אנשים שלא היו קשורים למועדונים כלשהם בלונדון בזמן ההכרזה. ודא שהשיוך למועדון מאומת על סמך מיקום המועדון במהלך חודש הפרס ולא לכלול כל אדם שעבר למועדונים בלונדון או ממנו לאחר תאריך ההכרזה.  
בנוסף, ניתן להוסיף רשימה של המועדונים בלונדון על מנת למנוע לחלוטין בלבול.

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

|  |  |
| --- | --- |
| **איזה שינוי הוצע?** | **האם לדעתכם יש מקום לשיפור זה? אם כן הסבירו כיצד ניתן לשפר. אם לא נמקו!** |
| ״ הם השתמשו בסלניום ולנו הוא עשה הרבה בעיות אז אולי הייתי משפר את זה״ | במהלך הפרויקט שלנו השימוש בסלניום היה ללא בעיות. |
| ״ לתת קצת יותר חופש בשאילתות כדי שלא יחזרו מאוד ספציפיות ולתת למשתמש לפלטר.״ | אנחנו מסכימים עם ההערה, אך למעשה מימוש השאילתות מאחורי הקלעים מתבסס על גמישות בקוד כך שאם נרצה להריץ שאילתה זהה עם פילטר שונה נפנה לאותה מחלקה רק עם סוג הפילטר החדש שאותו אנחנו רוצים לחפש. |
| ״מהירות החזרת התשובה״ | לדעתנו אין מקום לשיפור זה שכן באופן יחסי לכמות המידע שבאתר השאילתות שלנו מחזירות תוצאות די מהר. |
| ״ שיפור זמני חיפוש של selnium, ניתן להשתמש בטכנולוגיות אחרות כדי לשפר זמני ריצה.״ | מכיוון שהאתר מכיל פרסומות, התראות ובנוסף כל הטבלאות בשאילתות i ו- ii נמצאות באותו url, על מנת להגיע למידע הרלוונטי עלינו לבצע ״הקלקות״ בתוך האתר. בכדי לדמות משתמש שמקליק באתר, עלינו להשתמש בחבילת סלניום. בנוסף, לפני שהאתר לא יזרוק אותנו ויחשוב שאנחנו בוטים עלינו לחכות זמן מסוים בין קליק לקליק ולכן זמני הריצה ארוכים מהרגיל ולא ניתן לשפר זאת. |
| ״לא מצאתי משהו לשיפור, אולי לתקן את חישוב הריקול״ | בהמשך לשיפור השאילתה השלישית, גם חישוב הריקול תוקן☺ |

1. במהלך הקורס ובמהלך העבודה על הפרויקט, נתקלנו בכמה אתגרים משמעותיים שדרשו מאיתנו להתמודד עם בעיות שונות ולמצוא פתרונות יצירתיים:

1- הגדרת שאילתות וניתוחן:

ָ פיתוח שאילתות מדויקות: אחד האתגרים המרכזיים היה להגדיר שאילתות שיביאו נתונים מדויקים ורלוונטיים מהאתר. השאילתות שניסינו לפתח נדרשו להתמודד עם מגוון רחב של פרמטרים וקישורים, מה שיצר קושי בהשגת תוצאות עקביות. חיפשנו דרכים לחדד את השאילתות כך שיביאו תוצאות מדויקות יותר, דבר שדרש ניסוי וטעייה.

ָ אפיון מדויק של הנתונים: ניתוח הנתונים שהתקבלו מהשאילתות דרש להבין את התוכן וההקשרים של המידע באתר. התמודדנו עם בעיות בהבנה מעמיקה של הקשרים בין נתונים שונים ושינויים בדינמיקה של האתר, דבר שהשפיע על התוצאה הסופית של הניתוחים שביצענו.

2- עבודה עם כלים טכנולוגיים:

ָ שימוש ב-Selenium: העבודה עם Selenium הציבה אתגרים מבחינת ביצוע אינטראקציות עם דפי אינטרנט דינמיים. במיוחד, נתקלנו בקשיים בעבודה עם פרסומות, התראות ואלמנטים שמופיעים או נעלמים בזמן אמת. זה השפיע על מהירות הביצוע ועל יכולת השאיבת הנתונים, ודורש שיפוט מדויק לגבי מתי ואיך להפעיל את כלי הזחילה.

ָ חישובים מתקדמים עם Networkx: ביצוע חישובים כמו PageRank בעזרת ספריית Networkx דרש הבנה מעמיקה של מבני הגרפים וכיצד הם משפיעים על התוצאות. האתגר היה להבין את האופן שבו שינויים בקשרים ובנתונים משפיעים על חישוב PageRank ולהבטיח שהתוצאות יהיו מדויקות ונכונות.

ָ ויזואליזציה עם Matplotlib: הצגת התוצאות בצורה ויזואלית היוותה אתגר נוסף. נדרשנו למצוא דרכים להציג את הנתונים בצורה ברורה וקריאה, ולבצע התאמות בתצוגה לפי הצרכים של המשתמשים.

3- סיווג ודיוק התוצאות:

ָ הערכת רלוונטיות התוצאות: היה צורך להעריך את רלוונטיות התוצאות שהתקבלו מהשאילתות. ביצענו הערכות של Precision ו-Recall כדי לבדוק את האיכות של התוצאות ואת מידת ההתאמה שלהן לציפיות שלנו. בנוסף, סיווג התוצאות נעשה גם על ידי משתמשים שונים, מה שיצר אתגר נוסף בהבנת ההשפעה של שיקולים אישיים על ההערכות.

ָ התמודדות עם שיקולים אישיים: כל סיווג של התוצאות הושפע מהשיקולים האישיים של המשתמשים ומהנחות מוקדמות שלהם. נדרשנו למצוא דרכים להתמודד עם השפעות אלו ולוודא שהתוצאות שאנחנו מקבלים הן אמינות ונותנות תמונה מדויקת של הנתונים.

סיכום: האתגרים שהצגנו דרשו מאיתנו להיות יצירתיים ולבצע התאמות שוטפות במהלך הפרויקט. התמודדות עם מורכבות הנתונים והתנהגות האתר חידדה את כישורינו בתחום פיתוח הזחלים, ניתוח הנתונים והבנת הטכנולוגיות הקשורות לאחזור מידע. העבודה על הפרויקט אפשרה לנו לפתח פתרונות חדשניים וללמוד הרבה על כלים וטכניקות מתקדמות בתחום.