**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 כתבות החדשות עבור השחקנים מ-Ipswich Town שסטטוס ההעברה שלהם הוא: Transfer In או Loan Out

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 השחקנים בחלון ההעברות של Ipswich Town, סטטוס ההעברה של כל שחקן, לינק לכתבה המתאימה.

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. 20.06 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 i:  
   המושגים בשאילתה – Top, Pass, London, England, Player, Club, Seasons

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| W-d5 | W-d4 | W-d3 | W-d2 | W-d1 | IDF | D/dfi | Dfi | P5 | P4 | P3 | P2 | P1 | Word |
|  |  |  |  |  |  |  | 5 | 1 | 1 | 1 | 1 | 1 | Top |
|  |  |  |  |  | - | - | 0 | 0 | 0 | 0 | 0 | 0 | Pass |
|  |  |  |  |  | - | - | 0 | 0 | 0 | 0 | 0 | 0 | London |
|  |  |  |  |  |  |  | 30 | 6 | 6 | 6 | 6 | 6 | England |
|  |  |  |  |  |  |  | 24 | 3 | 7 | 4 | 7 | 3 | Player |
|  |  |  |  |  |  |  | 52 | 8 | 14 | 10 | 11 | 9 | Club |
|  |  |  |  |  |  |  | 15 | 2 | 3 | 4 | 4 | 2 | Seasons |

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 |
| 0.0823 | <https://www.itfc.co.uk/news/2024/june/30/omari-is-back> |
| 0.0823 | <https://www.itfc.co.uk/news/2024/july/01/town-sign-ben-johnson> |
| 0.0823 | <https://www.itfc.co.uk/news/2024/july/12/greaves-becomes-third-signing> |
| 0.0823 | <https://www.itfc.co.uk/news/2024/july/13/liam-delap-joins-town> |
| 0.0823 | <https://www.itfc.co.uk/news/2024/july/17/town-sign-keeper-muric> |
| 0.0823 | <https://www.itfc.co.uk/news/2024/august/01/conor-townsend-joins-ipswich-town> |
| 0.504 | <https://www.itfc.co.uk/news/2024/august/05/loan-move-for-elkan> |

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:   
A close-up of a network

Description automatically generated

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:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Recall | Precision | P7 | P6 | P5 | P4 | P3 | P2 | P1 |  |
|  |  |  | 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 במקרה זה פרמטר עבור "האם השחקן בסגל הבכיר או בקבוצת הנוער". שינוי זה יאפשר התאמה טובה יותר של תוצאות החיפוש להעדפות המשתמשים ויכול לשפר את איכות החיפושים.

1. <https://scraping-pl.vercel.app/project>
2. \*אחרי ההצגה\*
3. האתגרים בקורס ובפרויקט היו לחשוב על שאילתות שונות שאפשר להוציא מהאתר והשאילתות יחזירו קישורים מתאימים. אתגר נוסף היה לשנות שאילתות קיימות כך שהתוצאות שחזרו יהיו Hubs/Authorities.