Search Engine for Premier League Teams

Afonso da Silva Pinto
up202008014@edu.fe.up.pt
FEUP, Porto
Portugal

João Pedro Reis Teixeira
up202005437@edu.fe.up.pt
FEUP, Porto
Portugal

João Rola Reis
up202007227@edu.fe.up.pt
FEUP, Porto
Portugal

Pedro Manuel da Silva Gomes
up202006322@edu.fe.up.pt
FEUP, Porto
Portugal

ABSTRACT

This report delves into the realm of Premier League football clubs, presenting the development of a dynamic information retrieval system using data analytics and web scraping methods. Inspired by the fervor and popularity of Premier League football, the project addresses the urgent demand for an improved search mechanism. The report highlights meticulous data source selection, including the Premier League's official website, a Kaggle dataset covering Premier League games from 1993 to 2023, and player statistics, adhering to proprietary licenses while ensuring data availability. The structured and reproducible data processing pipeline efficiently integrated data into an SQLite database, enhancing the information retrieval experience for Premier League fans worldwide by creating a centralized platform encompassing everything Premier League-related and providing users with real-time access to club information, player statistics, game results, and news updates.

KEYWORDS

Premier League, game, article, team, dataset, player statistics, data, processing, web, scraping, textual field, information retrieval

1 INTRODUCTION

Inspired by the Premier League football's enormous popularity and ardent fandom, the massive influx of match results and news updates provide a compelling need for an improved search mechanism, similar to the wider context of information retrieval in the modern day. This project, which is an essential component of the Information Processing and Retrieval course, aims to create a cutting-edge information retrieval system by utilising data analytics and web scraping methods.

The paper begins by explaining the use of a large dataset that includes match outcomes. Insights are also given into the selection of data sources, the assessment of their quality, the corresponding data processing pipeline, and the conceptual model. This project improves the experience of Premier League fans around the world by addressing both the status of information retrieval now and opening doors for future developments.

2 DATASETS

2.1 Data Sources

Initially, the aim was to acquire comprehensive datasets with ample quantity, quality, and substantial textual content. The exploration began by looking into Arquivo.pt[2], a valuable web archive, with a focus on gathering information about Portuguese clubs. However, after an extensive investigation, data limitations were encountered. The diversity of club websites and the absence of an online presence for some clubs made web scraping challenging.

As an alternative, attention turned to the Premier League due to its global popularity and the wealth of data it offered. A valuable resource was uncovered in the form of a Kaggle dataset [4], containing records of all Premier League games played from 1993 to 2023 in CSV format under a CC0: Public Domain license. Nevertheless, this source lacked extensive textual information.

To address this gap in the primary dataset, it was complemented with Premier League-related news obtained through web scraping from the official website's 'news' section [7]. This source provided a substantial volume of data, including over 20 thousand news articles, covering a wide array of Premier League topics.

The decision to prioritize the Premier League official website as a primary source for news data was rooted in several compelling reasons. The official website holds authority as a trustworthy and authentic source of Premier League information. Moreover, it offers comprehensive and up-to-date coverage of Premier League events, aligning seamlessly with the objective of providing users with the most current league-related news.

For player statistics, reliance was placed on the Transfermarkt website [9], which also provided data in HTML format. This additional source furnished comprehensive player statistics for each season, including details such as goals scored and yellow cards.

Transfermarkt was chosen as a player statistics data source due to its reputation for accuracy and comprehensiveness in tracking player performance. Transfermarkt is widely regarded as a reliable platform for football-related data and enjoys a strong user base within the football analytics community. This approach enhances the reliability and efficiency of data collection.

Both the Premier League and Transfermarkt websites operate under proprietary licenses. However, they provide open access to their data, ensuring transparency and adherence to data usage regulations.

2.2 Dataset Characterization

The dataset under consideration pertains to the Premier League spanning from the 2016/17 to the 2022/23 seasons. It includes thorough information encompassing team and player statistics, along with news articles published within that timeframe.

2.2.1 ***Dataset Size.***

The dataset comprises 21191 news articles and the statistics of 1839 players throughout 7 seasons, made up of, in total, 2660 games.

2.2.2 *Dataset Contents*.

The dataset consists of 3 different datasets:

- **Articles:** Contains the titles, text, and date of publishing of the news articles.
- **Games:** Contains the games' results, teams involved, and the date.
- **Players and Teams:** Contains statistics for the players and teams such as goals scored, yellow or red cards received, etc.

2.2.3 *Dataset Quality*.

The sources being very reliable contributed to high-quality data, with no missing values or outliers detected in any of the datasets.

2.2.4 *Properties Characterization*.

To gain deeper insights into the dataset's content, an analysis of some of the most crucial attributes was conducted.

The initial focus was on identifying the number of wins of each team throughout all seasons, which can be seen in Figure 3, providing valuable insights into their performance.

Additionally, an effort was made to establish a potential correlation between the number of goals scored and the number of wins. To achieve this, the average number of goals each team scored per season was computed. As illustrated in Figure 4, several unexpected values emerged. For instance, Leicester City boasted a higher total number of wins than Liverpool, yet their average goals per season were lower. From this, it can be inferred that Leicester City exhibited a robust defense but a less potent offense when compared to Liverpool over the years.

Subsequently, the aim was to identify the teams that received the most mentions in news articles, shedding light on their popularity. As anticipated, Manchester City and Liverpool, the teams dominating the Premier League in recent years, emerged as the most frequently discussed in the articles, as can be seen in Figure 5.

In alignment with this approach, an investigation was conducted to identify which weeks typically saw the highest number of published news articles. As expected, the initial and final weeks of the season, on average, witnessed a greater volume of published articles compared to the weeks in between, as illustrated in Figure 6.

Lastly, an analysis was conducted to gain a more profound comprehension of some of the highest-performing players. To accomplish this, an examination was conducted to identify the top 10 players who scored the most goals across all seasons. Referencing Figure 7, it can be observed that Premier League legends such as Harry Kane, Mohamed Salah, and Jamie Vardy claim the leading positions on this list. This analysis also underscores Vardy's pivotal role for Leicester City over the seasons, given that the team didn't boast an abundance of goals, yet he managed to secure the third-highest goal tally.

3 DATA PROCESSING PIPELINE

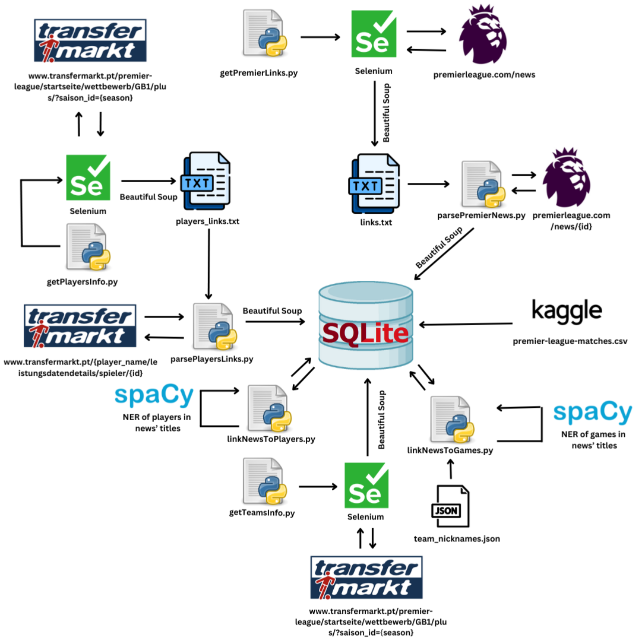
www.transfermarkt.pt/premier-league/startseite/wettbewerb/GB1/plus/?saison_id={season}
getPremierLinks.py
Selenium
premierleague.com/news
Beautiful Soup
Selenium
Beautiful Soup
players_links.txt
links.txt
parsePremierNews.py
premierleague.com/news/{id}
getPlayersInfo.py
Beautiful Soup
Beautiful Soup
kaggle
premier-league-matches.csv
www.transfermarkt.pt/{player_name/leistungsdatendetails/spieler/{id}
parsePlayersLinks.py
spaCy
NER of players in news' titles
linkNewsToPlayers.py
Beautiful Soup
linkNewsToGames.py
spaCy
NER of games in news' titles
getTeamsInfo.py
Selenium
team_nicknames.json
www.transfermarkt.pt/premier-league/startseite/wettbewerb/GB1/plus/?saison_id={season}

Figure 1: Data Processing Pipeline Diagram

The data collection and processing procedures were meticulously executed through the utilization of a structured and reproducible pipeline, as illustrated in Figure 1. The subsequent sections will provide a detailed exposition of this methodology

3.1 Data Collection

In the data collection phase, data was acquired from various sources to create a comprehensive document collection for information retrieval. To extract data from a CSV file within the Kaggle dataset, the versatile Pandas library was employed [5]. Pandas facilitated the parsing of the file into a structured data frame, and subsequently, this refined dataset was seamlessly integrated into an SQLite [6] database under the 'games' relation. Notably, team identifiers that refer to an entry in the 'teams' relation were stored instead of actual team names.

However, acquiring news data from the official Premier League website presented a more intricate challenge. Regrettably, the website lacks an official API for streamlined data retrieval. Furthermore, its news content does not load entirely upon initial rendering; instead, it employs an infinite scroll mechanism to progressively load additional news articles. Consequently, the Selenium[3] framework was employed to simulate the scrolling behavior required to access all available news articles. Upon completing this exhaustive collection, the powerful Beautiful Soup library[8] was utilized for HTML parsing and data extraction. Beautiful Soup [8] adeptly identified and stored all relevant links to news articles in a straightforward text file.

With the collection of news article links in hand, the next phase of data extraction was initiated. Individual HTTP requests were conducted for each article, leveraging Beautiful Soup[8] once more to parse the HTML content within the response payloads. For each news article, essential information, including its title, summary, full text, publication date, and URL, was meticulously stored within the 'articles' relation.

To gather information on teams and players, web scraping was utilized using the Selenium framework to extract data from the Transfermarkt website, as there was no official API available. The Beautiful Soup library[8] was also used for HTML parsing and data extraction, just like in the articles. For team data, HTML parsing was sufficient to obtain essential data such as squad size, average age, number of foreign players, average player value, and total squad value, which was then saved in the 'teams_stats' relation. For player data, the link for each player was obtained through the team's page and saved in a text file.

With the assembly of player links, HTTP requests were executed individually for each player, utilizing the Beautiful Soup [8] library for HTML content parsing. For each player, critical details such as their URL and name were systematically recorded within the 'players' relation. Additionally, to account for the distinct naming conventions within the statistical columns, players were categorized into two groups: goalkeepers and field players. Goalkeeper statistics encompass games played, goals scored, yellow cards received, double yellow cards received, red cards received, goals conceded, clean sheets, and minutes played, all of which were archived within the 'goalkeeper_stats' relation. Meanwhile, player statistics, comprising games played, goals scored, assists provided, yellow cards received, double yellow cards received, red cards received, and minutes played, were diligently stored in the 'player_stats' relation.

3.2 Data Processing

In the process of data collection, an initial dataset comprising all Premier League matches dating back to 1993 was assembled. Subsequently, a strategic decision was made to retain data exclusively from the 2016-17 season onwards. This decision was informed by the constraint that the official Premier League website provided news coverage solely for seasons commencing from 2016-17. To align the data collection efforts with the available information on the official Premier League website, exclusive focus was placed on this specific timeframe for the acquisition of match-related data.

Additionally, to ensure that the news data remained synchronized with available game and player statistics data, a data refinement step was implemented. Specifically, all news articles that were published more than 7 days after the last game for which there was data were removed. This step was essential to maintain data consistency and to align the news-related information with the existing game and player statistics, providing users with accurate and contextually relevant coverage of Premier League events.

In addition to dataset curation, advanced Natural Language Processing (NLP) techniques, specifically Named Entity Recognition (NER), were employed to identify and establish connections between news articles and their corresponding matches. The chosen NLP tool was the spaCy library [1]. For each article, the identifiers of the named teams were stored in the 'article_named_teams' relation. The heuristic to determine the relevance of news articles to specific matches relied on a temporal window: an article was considered pertinent to a match if it fell within a time frame of up to 7 days before or after the match. Furthermore, it was insisted that the named entities extracted from the article's title encompassed both teams involved in the match.

Initially, this approach yielded a relatively modest count of 128 matches. In an effort to expand the dataset, a 'team_nicknames' relation was introduced into the database. This additional relation was populated from a meticulously crafted JSON file, containing commonly used monikers and aliases for each team. This enhancement significantly broadened the matching capabilities, enabling the identification of relevant matches even when references to the teams were made using these aliases. Consequently, the number of matches detected effectively doubled.

However, upon closer examination, it became evident that not all of the identified entities found in the title were entirely accurate. A frequently encountered entity type took the form of 'team 1 v team 2', which led to those teams not being recognized. To address this issue, the SQL LIKE operator was employed to meticulously verify the presence of a team nickname within each named entity extracted from the article's title. This iterative refinement process proved to be instrumental in substantially enhancing the precision and accuracy of the matching criteria.

Through the synergistic application of these strategies, a total of 464 matches were successfully identified and linked.

Data was collected on 2237 players who registered for their teams from the 2016-17 season onwards. Out of these, only 1839 registered for the premier league. Following the same process as the teams, the identifiers of the named players in the 'article_named_players' relation were stored. As a result, a total of 1464 matches were successfully identified and linked.

4 CONCEPTUAL MODEL

Referring to Figure 2, the database is structured around eleven main tables:

- **articles** – Each article includes a title, summary, full text, publication date, URL, and a reference to the mentioned game.
- **games** – Game data includes the season it occurred in, the week of the season (out of 38), the date, home and away teams, their respective goal counts, and the full-time result.
- **teams** – Team information comprises the official name (as per Transfermarkt website), the name in the Kaggle dataset, and a short name (e.g., "Manchester City FC," "Manchester City," "MCI").
- **seasons** – Each season is represented by its name and the last year it spans (e.g., 2017-18, 2018).
- **team_nicknames** – A class to account for teams' unofficial nicknames used in news articles, mapping the official name to its nickname(s). For example, "Saints" refers to "Southampton."
- **article_named_teams** – This entity identifies the team or teams mentioned in the title of any given article.
- **players** – Each player is listed with their name and a link to their page on Transfermarkt.

- **players_stats** – Player's stats include the number of games played, goals scored, assists given, yellow cards received, double yellow cards received, red cards received, and minutes played.
- **goalkeepers_stats** – Goalkeepers stats include the number of games played, goals scored, yellow cards received, double yellow cards received, red cards received, goals conceded, clean sheets, and minutes played.
- **teams_stats** – For every team, there is a corresponding link to their page on Transfermarkt. Additionally, the squad member count, average player age, number of foreigners, average player value, and total squad value are all provided.
- **article_named_players** – This entity identifies the player or players mentioned in the title of any given article.

5 CONCLUSION

This document describes the steps that the datasets took to get to their final, usable condition.

The datasets were examined and investigated throughout this milestone in order to determine which data cleaning and preparation procedures were required for them to function towards the project aim.

Satisfied with the outcome, one can state, after careful analysis, that the datasets are prepared for the objectives of the following milestones.

6 FUTURE WORK

In the future, the objectives encompass the establishment of comprehensive search functionality within the system. The aim is to accomplish both straightforward and intricate search tasks to cater to diverse information needs.

6.1 Simple Queries

The aim is to facilitate effortless access to pertinent data. For instance, users should be able to initiate uncomplicated queries like searching for a team's name and promptly retrieving news articles related to it. Additionally, querying specific dates should yield details about games played on that day, along with related news.

6.2 Complex Queries

Beyond basic searches, the vision is to support more sophisticated queries. Users should have the ability to dynamically analyze the performance of Premier League players across multiple seasons. They should be able to enter queries such as "Performance trends of [Player Name] over [n] seasons" or "Comparison of [Player A] and [Player B] over [n] seasons".

Furthermore, there is a plan to cluster Premier League news articles into topics and offer recommendations to users based on their preferences. Users should be able to input search queries like "Recommend articles similar to [Article Title]", "News about [Player Name] injuries" and "Articles about [Player Name] transfers".

REFERENCES

[1] [SW] Explosion AI, spaCy: Industrial-Strength Natural Language Processing in Python version 4.1.4, 2023. URL: https://spacy.io/.

[2] [n. d.] Arquivo.pt. Accessed on October 10, 2023. https://arquivo.pt/.

[3] [SW] Selenium Contributors, Selenium: Browser Automation Framework version 4.1.0, 2023. URL: https://www.selenium.dev/.

[4] Evangower. 2022. Premier league matches 1992-2022. Accessed on October 10, 2023. Kaggle. https://www.kaggle.com/datasets/evangower/premier-league-matches-19922022.

[5] Wes McKinney et al. 2011. Pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14, 9, 1–9.

[6] Michael Owens. 2006. *The definitive guide to SQLite.* Springer.

[7] [n. d.] Premier league official website - news. Accessed on October 10, 2023. https://www.premierleague.com/news.

[8] [SW] Leonard Richardson, Matthew Levine, and Mark Pilgrim, Beautiful Soup: HTML Parsing Library version 4.10.0, 2023. URL: https://www.crummy.com/software/BeautifulSoup/.

[9] [n. d.] Transfermarkt. Accessed on October 10, 2023. https://www.transfermarkt.pt/.

A ANNEX

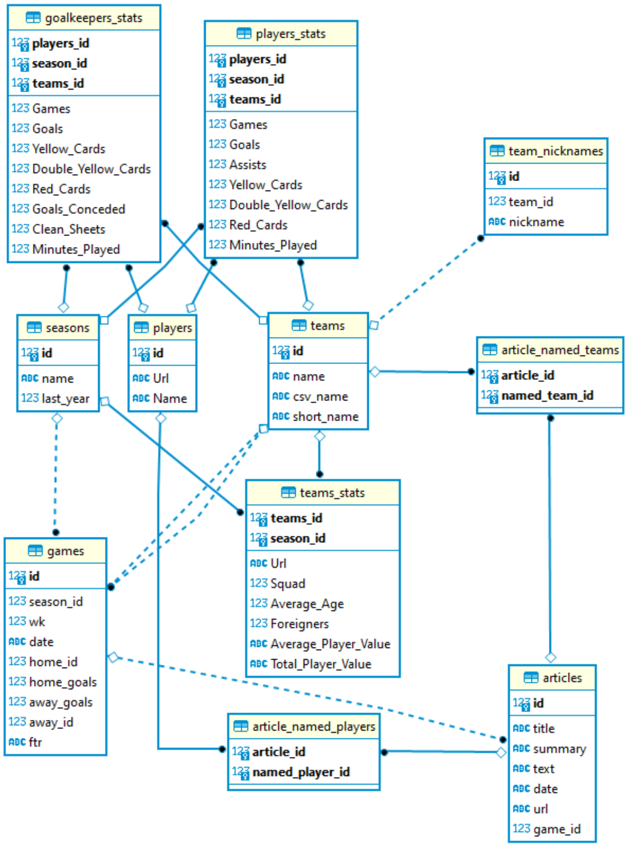

Figure 2: Domain Conceptual Diagram

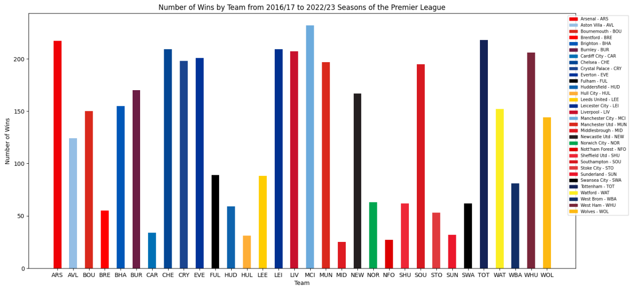

Figure 3: Number of Wins of each Team between 2016/17 and 2022/23 Seasons

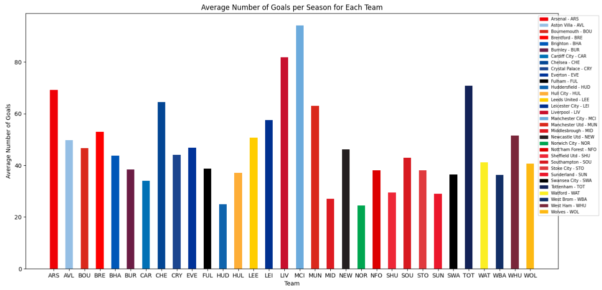

Figure 4: Average Goals scored by each Team per Season

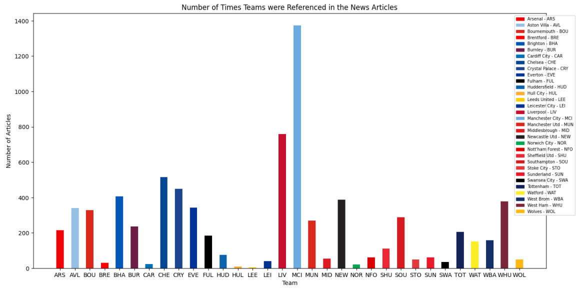

Figure 5: References in Articles for each Team

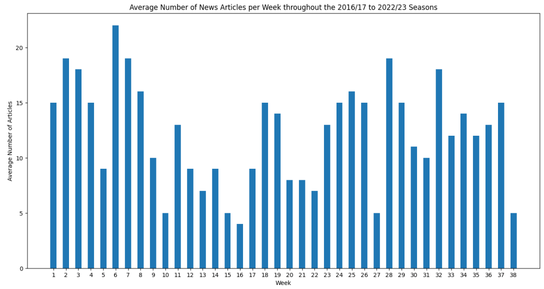

Figure 6: Average number of Articles published in each Week per Season

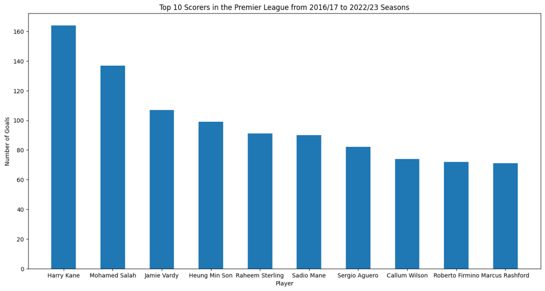

Figure 7: Top 10 Scorers across all Seasons