PRA 1: ¿Cómo podemos capturar los datos de la web?

Douglas Hernández Valerry

Ernesto Míguez Fernández

Tipología y ciclo de vida de los datos

En este segundo reto se elabora un caso práctico orientado a aprender a identificar datos relevantes para un proyecto analítico y usar las herramientas de extracción de datos.

1. Contexto

Para este proyecto se decidió realizar una investigación relacionada con una de las disciplinas deportivas con más seguidores a nivel mundial como el fútbol.

Estaremos realizando la extracción de los resultados de los partidos jugados en la liga premier de Inglaterra de la temporada actual con el fin principal de realizar un análisis descriptivo que nos brinde mejor información sobre la situación deportiva de esta liga como por ejemplo ver el desempeño de los equipos, entender si hay algún equipo que sea más efectivo de visitante que de local, etc.

Se decidió extraer los datos de esta liga ya que en comparación con las otras ligas principales como la española o alemana, ésta parece ser la más competitiva de todas.

Para implementar las técnicas de web scraping hacemos uso del lenguaje de programación R y del paquete rvest. La página que consultaremos será www.fbref.com la cual se dedica a la publicación de datos de las principales ligas de fútbol del mundo.

Los datos extraídos muestran los resultados de las distintas jornadas desde la temporada 2012/2013 a la fecha y su link de acceso es el siguiente:

Football Reference – Premier League 22/23

2. Título

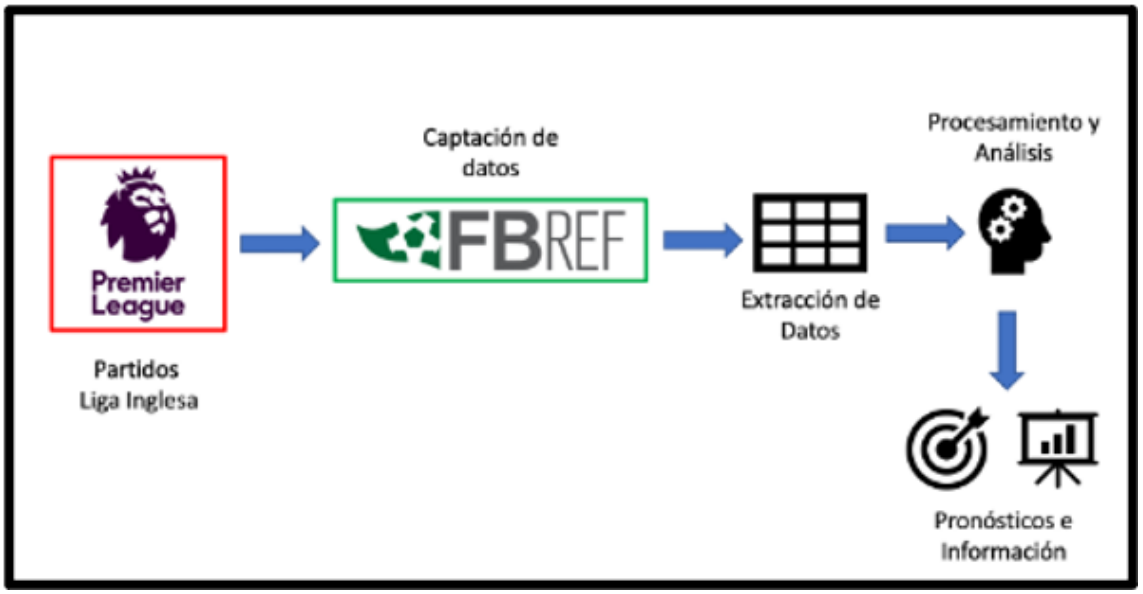
Histórico de resultados de partidos de fútbol de la liga inglesa para la temporada 22/23.

3. Descripción del dataset

El dataset contiene un total de 14 columnas y 380 registros correspondientes a los resultados de partidos de la Liga Premier de Inglaterra de la temporada actual.

Cada registro hace referencia a un partido del que no solo se tiene información del resultado del mismo, sino también de la fecha, el estadio o el árbitro del encuentro entre otros.

4. Representación gráfica



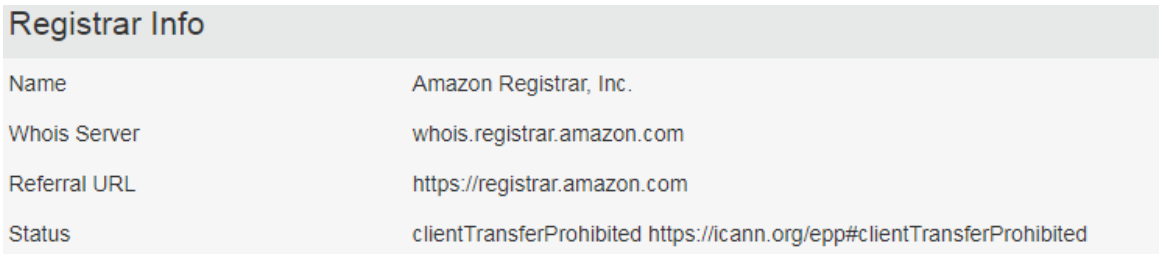
5. Contenido

En la siguiente tabla se muestra una breve descripción de los campos obtenidos mediante web scraping.

Campo	Descripción	Tipo de dato
wk	Número de semana de la temporada en la que se llevó a cabo el partido	Int
day	Día de la semana en el que se jugó el partido	String
date	Fecha en la que se registró el enfrentamiento	Date
time	Hora en la que inició el juego	String
home	Equipo Local	String
xgh	Goles esperados del equipo local	Float
score	Resultado del partido	Int
xga	Goles esperados del equipo visitante	Float
away	Equipo Visitante	String
attendance	Cantidad de público asistente	Int
venue	Estadio donde se jugó el partido	String
referee	Árbitro principal del partido	String
match_report	Indicador de si existe un resumen sobre el juego (no observable)	String
Notes	Anotaciones sobre el partido	String

6. Propietario

Los datos del propietario de la página web se muestran en la siguiente imagen, después de haber sido consultado en la página *https://who.is*.



Adicionalmente, con la intención de actuar de acuerdo con los principios éticos y legales se ha consultado el fichero robots.txt de la página con la intención de seguir las instrucciones a la hora de usar la página.
Como se muestra en la imagen a continuación, nuestro user-agent no estaba en la lista de restringidos por lo que no ha sido necesario cambiarlo, y el sitio web al que hemos accedido (*https://fbref.com/en/comps/*) tampoco estaba restringido según el fichero robots.txt.

De esta manera hemos procedido a realizar nuestra práctica utilizando el sitio web con normalidad.

```
User-agent:*
# Disallow: /cbb/
# Disallow: /cfb/
# Disallow: /olympics/

# Disallow: /awards/
# Disallow: /blog/
# Disallow: /boxscores/
# Disallow: /coaches/
# Disallow: /draft/
# Disallow: /executives/
# Disallow: /friv/
# Disallow: /hof/
# Disallow: /leaders/
# Disallow: /play-index/
# Disallow: /players/
# Disallow: /route.cgi
# Disallow: /schools/
# Disallow: /search/
# Disallow: /stadiums/
# Disallow: /static/
# Disallow: /teams/
# Disallow: /years/

Disallow: /feedback/
Disallow: /linker/
Disallow: /my/

Disallow: /news/
Disallow: /en/news/
Disallow: /pt/news/
Disallow: /de/news/
Disallow: /fr/news/
Disallow: /es/news/
Disallow: /it/news/
Disallow: /news/

Disallow: /req/
Disallow: /short/
Disallow: /nocdn/


User-agent: AhrefsBot
Disallow: /

User-agent: AhrefsBot/5.0
Disallow: /


## sitemaps generated by copyit/sitemaps/build_sitemaps.pl
##
Sitemap: https://fbref.com/sitemaps/sitemap.xml
```

7. Inspiración

Para esta práctica hemos decidido obtener información acerca de la liga inglesa de fútbol ya que ambos somos aficionados a este deporte y además creemos que puede ser interesante analizar esta liga, puesto que en los últimos años se está consolidando como la liga más competitiva del panorama internacional.

El objetivo es analizar con detalle esta competición, ver qué equipos son los más competitivos y quizá pueda servir como base para futuros análisis de otras competiciones con las que hacer una comparación y encontrar las razones que hacen a la Premier League la competición de fútbol más competitiva con el fin de identificar los aspectos que se pueden mejorar en otras competiciones para conseguir que sean más atractivas.

8. Licencia

La licencia bajo la que se publicará el dataset será Creative Commons Zero (CC0), con la que renunciamos a los derechos de autor para que otros analistas puedan utilizar el dataset sin ningún tipo de restricción. La razón de esta elección es ir en sintonía con el modelo open source, que en nuestra opinión ofrece un entorno favorable para la investigación y el desarrollo de nueva tecnología.

9. Código

```r

# SCRIPT PARA EXTRAER RESULTADOS DE LAS TEMPORADAS ANTERIORES

scrape_function <- function(temporada = 1992){

  #Scrape
  link <- paste0("https://fbref.com/en/comps/9/",temporada,"-",temporada+1,
                 "/schedule/",temporada,"-",temporada+1,"-Premier-League-Scores-and-Fixtures")

  bow(link)
  epl_page <- read_html(GET(link, timeout(10)))


  resultados <- link %>%
    read_html() %>%
    html_table() %>%
    .[[1]] %>%
    clean_names()

  resultados_clean <- resultados %>%
    filter(wk != "NA")

  write_csv(resultados_clean, "~/Documents/Douglas/Projects/futbol/R_data/scrap_directory/data_ligas/epl_data/epl_resultados.csv")
}

scrape_function(temporada = 2022)
```

10. Dataset

El DOI generado al publicar el dataset en Zenodo es:
`10.5281/zenodo.7337929`

Se accede mediante el siguiente link:

https://zenodo.org/record/7337929#.Y3j9kL3MLIU

11. Contribuciones

Investigación previa	Douglas Hernández | Ernesto Míguez
Redacción de las respuestas	Douglas Hernández | Ernesto Míguez
Desarrollo del código	Douglas Hernández | Ernesto Míguez
Participación en el vídeo	Douglas Hernández | Ernesto Míguez

Bibliografía

- **Subirats, L. & Calvo, M. (2019).** Web Scraping. Editorial UOC
- *https://guides.github.com/activities/hello-world*
- *https://fbref.com*
- *https://fbref.com/robots.txt*
- *https://who.is*