Estadístiques de la temporada 2018/2019 de la primera divisió de futbol espanyola

Pràctica 1 - Tipologia i cicle de vida de les dades

Francesc Valls Tor i Pol Codinachs Latorre

Context

La temàtica que s'ha escollit en aquest treball és sobre l'esport, concretament el futbol. Tractarem les dades de la lliga espanyola de futbol per tal d'observar en una temporada tots els enfrontaments que s'han disputat, i poder analitzar amb més detall el rendiment que ha tingut cada equip. Un anàlisi exhaustiu de cada jornada que s'ha disputat durant la temporada 2018/2019.

Títol

El títol escollit és el següent: **Estadístiques de la temporada 2018/2019 de la primera divisió de futbol espanyola.**

Descripció del dataset

El dataset es composa de diferents estadístiques extretes de tots els partits de futbol de la primera divisió espanyola jugats la temporada 2018/2019. La majoria de les variables obtingudes són numèriques i fan referència paràmetres del joc, com per exemple, percentatge de possesió, gols encaixats o faltes provocades, entre d'altres.

D'un partit jugat s'obtenen les mateixes variables tant per l'equip local com per l'equip visitant.

Representació gràfica

El projecte de web scrapping es centra en la web www.resultados-futbol.com d'on s'utilitzen dues URL per a l'extracció de totes les dades necessaries.

D'una banda, en l'url www.resultados-futbol.com/primera/{temporada}/grupo1/jornada{n} s'extreu un llistat de tots els enfrontaments disputats en la temporada de competició especificada.

Aquesta llistat de partits és el que s'utilitza en la segona url www.resultados-futbol.com/partido/{equip local}/{equip visitant}/{temporada} per extreure les estadístiques de cada enfrontament.

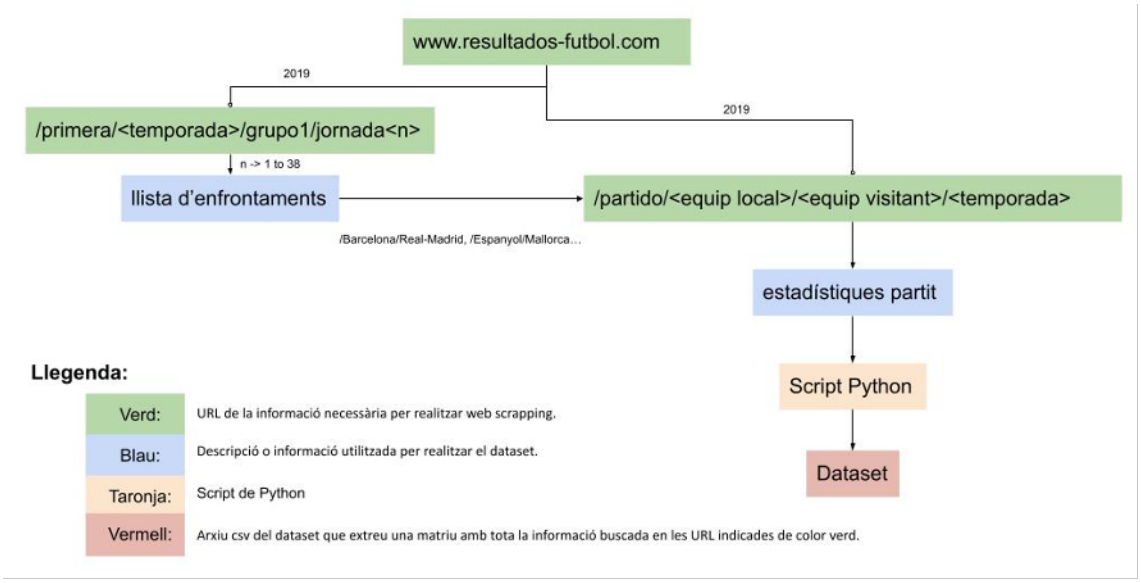

Figure 1: Diagrama de flux del projecte

Contingut

El dataset es composa de 37 variables que es detallen a continuació:

- Temporada:
Variable data que indica el any que s'ha analitzat les jornades de futbol.
- Jornada:
Variable numérica que indica el número de jornada que s'està disputant en aquella temporada.
- Partit:
Variable alfanumérica que indica el partit que disputen entre dos equips.
- L:
Variable alfanumérica que indica el equip que juga com a local (a casa).
- V:
Variable alfanumérica que indica el equip que juga com a visitant (a fora de casa).
- L-P:
Variable numérica que indica en porcentatge la possessió de pilota que ha tingut el equip local durant aquell partit.
- L-G:
Variable numérica que indica els gols que ha marcat el equip local durant aquell partit.
- L-XP:
Variable numérica que indica la quantitat de xuts que ha fet el equip local a la porteria del equip visitant.
- L-XF:
Variable numérica que indica la quantitat de xuts que ha fet el equip local direcció a la porteria del equip visitant pero que han anat fora de la poteria.
- L-XT:
Variable numérica que indica el total de xuts que ha fet el equip local.
- L-PP:
Variable numérica que indica la quantitat de parades que ha fet el porter del equip local.
- L-C:
Variable numérica que indica la quantitat de córners que ha llençat el equip local.
- L-FJ:
Variable numérica que indica la quantitat de fora de jocs que ha provocat el equip local.
- L-TG:
Variable numérica que indica la quantitat de targetes grogues que ha percebut el equip local.
- L-TV:
Variable numérica que indica la quantitat de targetes vermelles que ha percebut el equip local.
- L-AS:
Variable numérica que indica la quantitat d'assistències que ha fet el equip local.
- L-XPA:
Variable numérica que indica la quantitat de xuts del equip local que ha anat a algun dels 3 pals de la porteria del equip visitant.
- L-L:
Variable numérica que indica la quantitat de lesions que han patit el equip local.
- L-S:
Variable numérica que indica la quantitat de substitucions que ha fet el equip local.
- L-F:
Variable numérica que indica la quantitat de faltes que ha fet el equip local.
- L-PC:
Variable numérica que indica la quantitat de penaltis que han provocat el equip local al equip visitant.
- V-P:
Variable numérica que indica en porcentatge la possessió de pilota que ha tingut el equip visitant durant aquell partit.
- V-G:
Variable numérica que indica els gols que ha marcat el equip visitant durant aquell partit.
- V-XP:

Variable numérica que indica la quantitat de xuts que ha fet el equip visitant a la porteria del equip local.
- V-XF:
Variable numérica que indica la quantitat de xuts que ha fet el equip visitant direcció a la porteria del equip local pero que han anat fora de la poteria.
- V-XT:
Variable numérica que indica el total de xuts que ha fet el equip visitant.
- V-PP:
Variable numérica que indica la quantitat de parades que ha fet el porter del equip visitant.
- V-C:
Variable numérica que indica la quantitat de córners que ha llençat el equip visitant.
- V-FJ:
Variable numérica que indica la quantitat de fora de jocs que ha provocat el equip visitant.
- V-TG:
Variable numérica que indica la quantitat de targetes grogues que ha percebut el equip visitant.
- V-TV:
Variable numérica que indica la quantitat de targetes vermelles que ha percebut el equip visitant.
- V-AS:
Variable numérica que indica la quantitat d'assistències que ha fet el equip visitant.
- V-XPA:
Variable numérica que indica la quantitat de xuts del equip visitant que ha anat a algun dels 3 pals de la porteria del equip local.
- V-L:
Variable numérica que indica la quantitat de lesions que han patit el equip visitant.
- V-S:
Variable numérica que indica la quantitat de substitucions que ha fet el equip visitant.
- V-F:
Variable numérica que indica la quantitat de faltes que ha fet el equip visitant.
- V-PC:
Variable numérica que indica la quantitat de penaltis que han provocat el equip visitant al equip local.

Propietari

Les dades són extretes de la pàgina web www.resultados-futbol.com que és una pàgina web dedicada al futbol internacional, on nosaltres ens hem centrat en recollir les dades de la lliga de futbol professional d'Espanya per poder realitzar la pràctica de web scrapping, concretament de la temporada 2018/2019.

Inspiració

Hem escollit aquest tema per fer la pràctica de web scrapping perquè ens resulta interessant els anàlisis i/o estadístiques que es poden extreure observant tots els enfrontaments d'una temporada de futbol i així poder trobar certes tendències o patrons pel que fa les variables de joc en un partit.

Llicència

Hem decidit escollir la següent llicència **CC0: Public Domain License** ja que no volem posar cap mena de restricció al nostre dataset publicat i que el proper usuari que en fagi ús sigui liure de poder copiar, modificar o distribuir-lo sense limitacions ni haver de demanar permís als autors.

Codi

El codi del nostre projecte de webscrapping ha estat generat en llenguatge Python i consisteix, en trets generals, mitjançant la llibreria BeautifulSoup, apuntar diferents pàgines de www.resultados-futbol.com per tal d'extreure les dades desitjades.

En primer lloc, es pot executar l'script desenvolupat de la següent manera per tal d'obtenir les estadístiques de qualsevol temporada de futbol de la primera divisió espanyola posterior a l'any 2018/2019.

Per obtenir les estadístiques de la temporada 2018/2019, cal escriure "2019". Si per exemple, es volen les dades de la temporada 2021/2022, l'any que s'escriuria en la següent comanda seria "2022".

```
python3 main.py 2019
```

Pel que fa l'script, en primer lloc es presenten les llibreries a importar i s'escriuren els headers utilitzats:

```
from sys import argv
import requests
import pandas as pd
import argparse
from bs4 import BeautifulSoup

headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36(KHTML,
like Gecko) Chrome/56.0.2924.76 Safari/537.36', "Upgrade-Insecure-Requests": "1",
"DNT": "1","Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8",
"Accept-Language": "en-US,en;q=0.5","Accept-Encoding": "gzip, deflate"}
```

Seguidament, es realitza el primer request mitjançant BeautifulSoup per tal d'extreure un llistat de tots els enfrontaments que hi ha hagut en la temporada seleccionada. Es realitzen 38 iteracions (ja que les temporades espanyoles contenen 38 jornades) a la següent url:

www.resultados-futbol.com/primera/{any}/grupo1/jornada{n}

On {any} és la temporada especificada en la comanda i {n} és cadascuna de les jornades 1..38. En aquestes iteracions, s'accedeix a les files d'una taula amb classe "vevent" que és d'on s'extreurà els noms dels equips locals i equips visitants que s'enfronten en aquella jornada.

```
def partits (year):
    matchlist = []
    for i in range(1,39):
        url = ("https://www.resultados-futbol.com/primera" + str(year) + "/grupo1/jornada"
        + str(i))
        page = requests.get(url, headers=headers)
        soup = BeautifulSoup(page.content, features="html.parser")
        for item in soup.find_all('tr', class_="vevent"):
            for match in item.find_all('a', class_="url", href=True):
                if str(match['href'][9:][-5:]) == ("/" + str(year)):
                    matchlist.append(str(match['href'][9:][:-5]))
                else:
                    matchlist.append(str(match['href'][9:]))
    return [matchlist, year]
```

Després s'utilitza aquesta llista de sortida, que conté un llistat amb tots els partits i l'any de joc, en una altra funció que realitzarà iteracions sobre cada enfrontament de la llista i utilitzar-lo en la següent url:

www.resultados-futbol.com/partido/{equip local}/{equip visitant}/{any}

S'accedeix a una taula continguda a cada pàgina i s'accedeixen a les diferents columnes (td) de les files (tr) amb classe "barstyle bar4". Amb l'informació extreta es genera una llista per a les dades de l'equip local i una llista per a les dades de l'equip visitant que s'acaben concatenant.

Finalment, mitjançant la llibreria pandas, es genera el dataset de sortida amb les respectives columnes.

Una de les dificultats trobades a l'hora de desenvolupar l'script, ha estat que no tots els partits contenien les mateixes dades. Així doncs, s'ha decidit inicialitzar les llistes de local i visitant a zero i generar una llista amb els noms definitius de les columnes. D'aquesta manera, quan s'extreuen dades d'un partit en concret, si la variable que s'està observant no es troba en la mateixa posició actual de la llista "parameters", es deixa amb un zero i es continua a la següent iteració.

```
def dataset (partits):
    llista_partits = partits[0]
    any = partits[1]
    local_visitant = []
    df = []
    n = 1
    n_partit = 0
    print("Jornada " + str(n))
    for i in llista_partits:
        local = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
        visitant = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
        if (any == 2023):
            url = ("https://www.resultados-futbol.com/partido/" + str(i))
        else:
            url = ("https://www.resultados-futbol.com/partido/" + str(i) + "/" + str(any))
        page = requests.get(url, headers=headers)
        soup = BeautifulSoup(page.content, features="html.parser")
        parameters = ['Posesión del balón', 'Goles', 'Tiros a puerta', 'Tiros fuera',
        'Total tiros', 'Paradas del portero', 'Saques de esquina', 'Fueras de juego',
        'Tarjetas Amarillas', 'Tarjetas Rojas', 'Asistencias', 'Tiros al palo', 'Lesiones',
        'Sustituciones', 'Faltas', 'Penalti cometido']
        for index, table in enumerate(soup.find_all('tr', class_="barstyle bar4")):
            local[parameters.index(
              table.find_all('td')[1].text)] = table.find_all('td')[0].text
            visitant[parameters.index(
              table.find_all('td')[1].text)] = table.find_all('td')[2].text
        n_partit = n_partit + 1
        local_visitant = local + visitant
        print("=" * n_partit)
        local_visitant.insert(0, str(int(any)-1) + "/" + str(any))
        local_visitant.insert(1, n)
        local_visitant.insert(2, i)
        local_visitant.insert(3, str(i).split('/')[0])
        local_visitant.insert(4, str(i).split('/')[1])
        df.append(local_visitant)
        if n_partit == 10:
            n = n + 1
            n_partit = 0
            print("Jornada " + str(n))
```

```
    df = pd.DataFrame(df)
    column_names = ['Temporada', 'Jornada', 'Partit', 'L', 'V', 'L-P', 'L-G', 'L-XP',
    'L-XF', 'L-XT', 'L-PP', 'L-C', 'L-FJ', 'L-TG', 'L-TV', 'L-AS', 'L-XPA', 'L-L',
    'L-S', 'L-F', 'L-PC', 'V-P', 'V-G', 'V-XP', 'V-XF', 'V-XT', 'V-PP', 'V-C', 'V-FJ',
    'V-TG', 'V-TV', 'V-AS', 'V-XPA', 'V-L', 'V-S', 'V-F', 'V-PC']
    df.columns = column_names
    df.to_csv('dataset.csv')
    print(df)
```

Dataset

El dataset del projecte ha estat publicat a la plataforma Zendono amb el següent **DOI:** 10.5281/zenodo.7339755

Vídeo

El vídeo explicatiu de la pràctica es pot trobar en el següent enllaç: https://drive.google.com/file/d/1XdCsSsH-8C9WxmsLy4p-X7ewjdqJTzvY/view?usp=sharing

Contribucions	Signatura
Investigació prèvia	FVT, PCL
Redacció de les respostes	FVT, PCL
Desenvolupament del codi	FVT, PCL
Participació al vídeo	FVT, PCL