PRA 1: Web Scraping

En aquest document s'ha copiat tot el contingut del README del projecte. La informació més actualitzada es pot trobar al propi repositori.

Football Scraper

En aquest projecte s'ha creat un web scraper amb finalitats acadèmiques. Concretament pretén resoldre la PAC1-Web Scraping de l'assignatura M2.951 - Tipologia i cicle de vida de les dades del Màster universitari de Ciència de dades impartit per la Universitat Oberta de Catalunya (UOC).

En aquest cas s'ha escollit agafar les dades de la lliga espanyola de futbol. L'extracció de dades s'ha realitzat sobre el web http://www.resultados-futbol.com el qual té un històric de totes les jornades de la lliga espanyola des del seu inici l'any 1929 fins a l'actualitat.

La motivació en aquest projecte ve donada arran de situacions quotidianes on diem que guanyarà un equip o un altre basant-nos solament amb el nom de l'equip. Per exemple, si es fes una enquesta on és preguntes "Qui guanyarà el partit Barcelona-Eibar?" és molt probable que hagis respost Barcelona, solament pel renom que té i la carrera que porta com a club. Basant-me en aquest fet, m'agradaria generar un dataset on es recollís els partits jugats en la lliga espanyola juntament amb el seu resultat. En aquesta assignatura no es contempla la creació de model predictius, però la idea seria crear un model que fes prediccions d'acord amb els noms dels equips que s'enfronten i digués quin dels dos és més probable que guanyi.

M'agradaria agrair a l'equip de resultados-futbol per oferir totes les dades necessàries per la realització d'aquesta pràctica. S'ha de dir que ofereixen una API de pagament per fer consultes directes de les dades, però com bé s'ha dit, en aquest cas es farà ús d'un web scraper per la recol·lecció de les dades.

Membres del grup

L'activitat ha estat realitzada de manera individual per Albert Eduard Merino Pulido.

Enllaç al projecte

https://github.com/data-science-env/Football-Scraper

Fitxers i carpetes del projecte

Estructura

Football-Scraper
├── logos
│ ├── equip1.jpg
│ ├── equip2.jpg
│ └── equipN.jpg
├── src
│ ├── chromedriver_78.exe
│ ├── models.py
│ └── scraper.py
├──
├── football_spanish_league.csv
├── PRA1-Web Scraping.pdf
├── README.md
└── requirements.txt

Descripció

- `logos` conté els logos dels diferents equips en format jpg
- `src/chromedriver_78.exe` s'utilitza per a executar l'exemple de Selinux.
- `src/models.py` conté les diferents classes utilitzades per representar el objectes de Equip, Jornada i Lliga.
- `src/scraper.py` conté el proces principal per generar el dataset.
- `football_spanish_league.csv` dataset generat com a solució de la practica.
- `PRA1-Web Scraping.pdf` conté la documentació de la practica. El contingut es el mateix que aquest README.
- `README.md` conté l'informació del projecte.
- `requirements.txt` conté les dependencies per poder executar el projecte.

Executar el web scraper

```
# Crear i activar un entorn virtual
virtualenv env
source env/bin/activate

# Instal·lar les dependencies de projecte
pip install -r requirements.txt

# Executar scraper
python scraper.py [-h] [--selenium] [--no_data]
```

Arguments:

- -h, --help mostra un missatge d'ajuda

- --selenium Indica si volem executar el codi de Selenium
- --no_data Indica si no volem agafar dades

Robots.txt

Durant la realització d'aquesta pràctica, el fitxer robots.txt era el següent.

```
User-agent :  *
Disallow :  /muro
Disallow :  /perfil
Disallow :  /amigos
Disallow :  /mensajes
Disallow :  /notificaciones
Disallow :  /misgrupos
Disallow :  /misfotos
Disallow :  /misvideos
Disallow :  /misblogs
Disallow :  /misnoticias
Disallow :  /misjuegos
Disallow :  /control
Disallow :  /editor
Disallow :  /legal
Disallow :  /normas_uso
Disallow :  /video/
Disallow :  /videos/
Disallow :  /fotos/usuario/
Disallow :  /noticias/usuario/
Disallow :  /videos/usuario/
Disallow :  /ajax/load_extension.php
Disallow :  /ajax/preload_extension.php

Allow :  /

User-agent :  Mediapartners-Google
Disallow :

User-agent :  grapeshot
Disallow :
```

Important : En utilitzacions futures, cal revisar l'arxiu robots.txt de nou per comprovar que se segueix tenint permís per extraure les dades del web.

Estructura HTML d'un partit

```
<tr class="vevent">
   <td class="fecha">10 Feb 29</td>
   <td class="equipo1">
       <a href="/Arenas-Club/1929" title="Arenas"><img width="18"
src="https://thumb.resfu.com/img_data/escudos/small/4657.jpg?size=37x&amp;5"
alt="Arenas de Getxo">Arenas</a>
   </td>
   <td class="rstd">
       <span class="summary hidden" title="Arenas - Atlético">Arenas - Atlético</span>
       <span class="dtstart hidden" title="1929-02-10T00 : 00 : 00">1929-02-10T00 : 00
: 00</span>
       <span class="location hidden">Municipal de Gobela</span>
       <span class="eventType category" title="Fútbol"></span>

       <a class="url" href="/partido/Arenas-Club/Atletico-Madrid/1929">2 -
 3</a> </td>
   <td class="equipo2">
       <a href="/Atletico-Madrid/1929" title="Atlético"><img width="18"
src="https://thumb.resfu.com/img_data/escudos/small/369.jpg?size=37x&amp;5"
alt="Atlético">Atlético</a>
   </td>
   <td class="cmm"><a class="c" href="/partido/Arenas-Club/Atletico-
Madrid/1929">5</a></td>
</tr>
```

Com podem observar l'estructura és ben clara. El **primer** element marcat amb *class="fecha"* correspon a data en la qual es va jugar el partit de futbol.
El **segon** element marcat amb *class="equipo1"* conté la informació de l'equip 1. En aquest cas s'ha agafat el nom de l'equip juntament amb el seu escut (logo). El **tercer** element marcat amb *class="rstd"* conté informació del partit. D'aquest element s'ha agafat l'estadi on es va disputar el partit i el resultat. El **quart** element marcat amb *class="equipo2"* segueix el mateix esquema que l'equip 1. Per últim apareix el nombre de comentaris que s'ha fet sobre el partit els quals no s'han agafat, ja que no apareixen enlloc.

Contingut dataset

El dataset generat conté les dades dels partits de futbol de la lliga espanyola des de la temporada 1929 fins a la temporada 2020. Cal destacar que no hi han dades dels anys 1937, 1938, i 1939, ja que a causa de la guerra civil que va patir Espanya no va haver-hi lliga. El nom escollit per al dataset ha estat `football_spanish_league.csv`
A continuació es mostren les capçaleres del dataset juntament amb el tipus de valor que contenen.

- year : numeric
- jornada : numeric

- date : string
- stadium : string
- teamA : string
- logo_teamA : string
- teamB : string
- logo_teamB : string
- scoreTeamA : numeric
- scoreTeamB : numeric
- winner : string
- winnerAsNumeric : numeric

Com es pot veure, s'han guardat els logos de l'equip A i l'equip B. En el dataset surten representats amb el nom del fitxer que conte la imatge. A la carpeta logos, estan tots els logos dels diferents equips. Els logos segueixen tots el format nom_equip.jpg, per exemple, el fitxer amb el logo l'equip Eibar s'anomena Eibar.jpg.

Aspectes a destacar

- S'han modificat les capçaleres del scraper per tal de fer-nos passar per un navegador. Utilitzant el següent link podem saber el *user agent* del nostre navegador.
- S'han fet ús de temporitzadors per tal de no saturar el servidor amb les peticions.
- S'han emmagatzemat els escuts dels diferents equips en format jpg i s'ha afegit el nom del fitxer al csv. Per tal de no descarregar cada vegada l'escut d'un equip s'ha posat una mesura per detectar si ja s'havia descarregat anteriorment.
- Fent ús de Selenium s'ha gestionat el login de la pagina. L'usuari que s'utilitza és *scraper* i la contrasenya *scraper*. Evidentment és un usuari vàlid i a la vegada fals. S'ha utilitzat un correu temporal per al registre.
- El programa gestiona quan hi han problemes inesperats com podria ser la caiguda de connexió amb el servidor.

A mode d'exemple, s'ha utilitzat Selenium per fer login en la mateixa pàgina que s'ha realitzat la pràctica. Solament s'ha utilitzat per recuperar informació bàsica de l'usuari. En cap moment s'ha utilitzat per generar cap dataset nou. Per poder fer ús de Selenium s'ha hagut de descarregar un driver de Chorme des del següent link. En el meu cas he descarregat el driver per a la versió 78 de Chrome el qual deixo en aquesta mateixa carpeta.

Llicencia

El dataset generat en el fitxer `football_spanish_league.csv` està subjecte a la següent llicencia Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

S'ha escollit aquesta llicència d'acord amb les seves característiques. Com podem veure aquesta llicència permet que es comparteixi adaptacions de les dades sempre que aquestes es comparteixin de la mateixa manera. S'ha decidit que al ser dades extretes per a la resolució d'una pràctica estudiantil, aquestes dades no tinguis permís per ser utilitzades amb fins comercials.

Contribucions en el projecte

Contribucions	Signa
Recerca prèvia	AMP
Redacció de les respostes	AMP
Desenvolupament codi	AMP