Universitat Oberta de Catalunya
uoc.edu

Tipología y ciclo de vida de los datos

PRAC1

Joan Miquel Forteza Fuster
41539288-T

Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona esta información.

Para la realización de nuestro Web Scraper se ha elegido la página de http://www.resultados-futbol.com donde en esta página se recogen los resultados de los partidos de fútbol de todas las ligas del mundo y es una página donde recoger un buen ejemplo de datos, en nuestro caso recogeremos los datos de la Premier League. Este sitio es de una empresa llamada BeSoccer, dedicado a todo tipo de datos de competiciones de fútbol a nivel mundial.

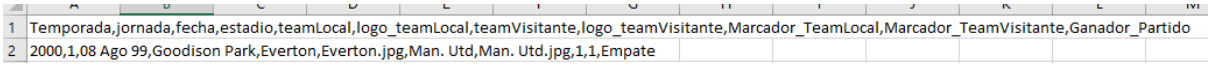
Nombre del dataset

Premier League → PremierLeague.csv

Descripción del dataset

En este DataSet se ha querido recolectar la información de las Jornadas de la Premier League, con la finalidad de poder hacer un análisis de las jornadas a lo largo de los años del siglo actual de esta Liga de fútbol.

Visualización y explicación del contenido del dataset



1	Temporada,jornada,fecha,estadio,teamLocal,logo_teamLocal,teamVisitante,logo_teamVisitante,Marcador_TeamLocal,Marcador_TeamVisitante,Ganador_Partido
2	2000,1,08 Ago 99,Goodison Park,Everton,Everton.jpg,Man. Utd,Man. Utd.jpg,1,1,Empate

- **Temporada**: Se refiere a los años que transcurre la temporada (string)
- **Jornada**: Numero de la jornada correspondiente (numeric)
- **Fecha**: Fecha en la que se disputó el partido (string)
- **Estadio**: Nombre del estadio en el que se disputó el partido (string)
- **teamLocal**: Nombre del equipo Local (string)
- **logo_TeamLocal**: nombre del logo del equipo local (string)
- **teamVisitante**: Nombre del equipo Visitante (string)
- **logo_TeamVisitante**: nombre del logo del equipo visitante (string)
- **Marcador_teamLocal**: Número de goles del local (numeric)
- **Marcador_teamVisitante**: Número de goles del visitante (numeric)
- **Ganador_Partido**: Nombre del ganador del partido, en caso de empate se pondrá Empate (string)

El periodo de tiempo de la recolección de datos es desde la temporada 1999/2000 hasta la temporada actual, se ha recogido mirando los tags html de cada jornada en el calendario. Se incluye la temporada 99/00 aunque tenga parte del siglo anterior por que la finalización recae en el siglo actual.

Lo datos se han recogido mediante un WebScrapper desarrollado en python. Dicho WebScrapper recoge los datos del siguiente ejemplo de html del directorio calendario:

```
<div class="boxtop">
    <div class="linea">
        <div class="botonlivs">
            <span class="titlen">
                <a href="/premier/grupo1/jornada10">Ver jornada</a>
            </span>
        </div>
        <span class="titlebox">Jornada 10</span>
    </div>
</div>
<div class="contentitem">
    <table id="tabla1" cellspacing="0" summary="">
        <tbody>
            <tr class="vevent">
                <td class="fecha">27 Oct 19</td>
                <td class="equipo1">
                    <a href='/Arsenal' title='Arsenal'>
                        <img width="18" src="https://thumb.resfu.com/img_data/escudos/small/263.jpg?size=37x&5" alt="Arsenal" />Arsenal
                    </a>
                </td>
                <td class="rstd">
                    <span class="summary hidden" title="Arsenal - Crystal">Arsenal - Crystal</span>
                    <span class="dtstart hidden" title="2019-10-27T17:30:00">2019-10-27T17:30:00</span>
                    <span class="location hidden">Emirates Stadium</span>
                    <span class="eventType category" title="Fútbol"></span>
                    <a class="url" href="/partido/Arsenal/Crystal-Palace-Fc">2 - 2</a>
                </td>
                <td class="equipo2">
                    <a href='/Crystal-Palace-Fc' title='Crystal'>
                        <img width="18" src="https://thumb.resfu.com/img_data/escudos/small/860.jpg?size=37x&5" alt="Crystal Palace" />Crystal
                    </a>
                </td>
                <td class="cmm">
                    <a class="c" href='/partido/Arsenal/Crystal-Palace-Fc'>192</a>
                </td>
            </tr>
            <tbody>
            </table>
            <div>
```

Como podemos ver tenemos nuestros datos del dataset en ese trozo de html y solo hay que recorrer los tags cada partido recogiendo los datos que queremos para su posterior análisis.

De estos tags hay que destacar que al obtenerse con el WebScrapper se han guardado los logos en una carpeta, llamado logos, donde se ha guardado en formato jpg, y controlando de manera que al obtener el escudo del equipo local y el visitante, si estos ya existen en la carpeta, no los vuelva a crear.

También como comentamos en la primera PAC, hemos añadido un waiter entre petición y petición para así evitar saturar el servidor con nuestra obtención de datos, se ha establecido un tiempo de 10 segundos entre petición y petición, que podría ser un poco la simulación de la navegación de un navegador web, modificando así también las cabeceras del web scraper con el user agent de nuestro navegador que podemos obtener fácilmente buscando en nuestro navegador mediante check user agent en nuestro caso al usar el navegador Chrome .

Hay que tener en cuenta también que se han formulado una serie de preguntas que se podrían responder con nuestro conjunto de datos, pero para realizar dichas contestaciones, requerimos de otro paso más, Diseño y explotación de los datos recolectados con informes o con gráficas, modelos de predicción, etc.. pero este paso se obvia ya que no es necesario para la asignatura actual de la práctica.

Agradecimientos

Agradecer al equipo BeSoccer, por proporcionar la información que requería a través de resultados-futbol.com, también se ofrece un api de pago para obtener la misma información que queremos obtener, pero por seguir lo requerido en la práctica se decide optar por el web scraper. Es cierto también que se había optado primero por la página de mismarcadores.com, pero ofrece muchas limitaciones a la hora de obtener los datos, entonces se siguió investigando y se encontró resultados-futbol.com a través de https://www.besoccer.com/, viendo así que ofrece un diseño y obtención de información más abierto que la fuente primera de recolección.

El archivo de robots.txt de resultados-futbol.com es el siguiente:

```
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:
```

Siendo esto de una manera más flexible que mismarcadores, ya que el contenido desde donde se quería obtener primeramente estaba disallowed, que sería el cuadro, información que se asemejaba al calendario de resultados de donde sacamos la información.

```
User-agent: *
Disallow: /clasificacion/
Disallow: /cuadro/
Disallow: /redirect/
Disallow: /partido/
Disallow: */x/js/browsercompatibility*.js

User-agent: SmartViper
Disallow: /

User-agent: Mediapartners-Google
Disallow: /clasificacion/
Disallow: /cuadro/
Disallow: /redirect/
Disallow: /x/
Allow: /partido/
```

Está claro que no es obligatorio tener en cuenta estas restricciones, pero para evitar ser bloqueados por el autor se ha optado por obtener los datos a través de resultados-futbol.com.

Inspiraciones

El motivo de hacer esta práctica orientada al fútbol de la Premier League, es debido a que desde pequeño llevo jugando al fútbol y viendo el fútbol en televisión y siempre he tenido gran afán por la premier league, que cuenta con la mejor afición de todas las ligas del mundo y he tenido la ocasión de poder disfrutar en directo de ciertos partidos de esta misma y comprobar que ese dicho es real. El ambiente y la euforia de la afición, no se siente en ningún otro lugar y da igual que sea el colista que la afición alentará más que si fuera la afición del campeón. Además la Premier League, es muy caótica y el nivel de competición es muy alto, es cierto que la liga española se ha ido regulando año a año y ahora los partidos están más reñidos, pero el dominio sigue siendo el mismo, FC. Barcelona y Real Madrid, en cambio en la Premier nunca se sabe, hace unos años ganaba el Leicester de una manera sorprendente, luego llegó el dominio del MC City y parecía que nadie tenía que quitarle el trono, hasta que Liverpool se entromete por enmedio. Por eso la decisión de obtener los datos de la premier league y poder contestar y predecir preguntas de quién ganará, o si ganará por que esa en campo a favor, que equipo recién ascendido ganará más partidos, etc.. En relación también al periodo seleccionado de tiempo, es cierto que la página ofrece más temporadas para obtener datos y se podría hacer un análisis de los cambios que ha habido entre siglos, pero en este caso, interesa ver como al largo de los años el cambio en las competiciones con el nivel altísimo de dinero actual ha cambiado toda

la lógica de los resultados del fútbol inglés, de menos a más dinero en el patrimonio de los clubes.

Licencia

La licencia utilizada y escogida en este caso para nuestro Dataset es **Creative Commons Reconocimiento-NoComercial (CC BY-NC)**, es la licencia que mejor se adapta en nuestro caso, ya que es un proyecto de finalidades académicas y queremos se pueda reutilizar dicha información, sin violar al autor de la información y bloqueando su posible distribución comercial ya que los datos son obtenidos por medio de otro autor y no nos interesa que esta información pueda ser aprovechada por terceros relacionados con el tratamiento similar de información que el autor para realizar comercializaciones cuando se nos ha cedido dicha información a nosotros para finalidades académicas. La información de que licencia aplicar se ha extraído de la página Licencias de uso asociadas a las iniciativas de datos abiertos en España y su posterior obtención y confirmación en creativecommons.

Recursos

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC. PID_00256968.pdf
- Teguayaco Gutiérrez González, Web scrapiing aviation accidents https://github.com/tteguayco/Web-scraping-aviation-accidents

Firma

Contribuciones	Firma
Búsqueda previa	JFF
Contestación de las respuestas	JFF
Desarrollo de código	JFF

JFF: Joan Miquel Forteza Fuster

Zenodo

El siguiente link nos lleva a la página de Zenodo donde está publicado nuestro dataset.