DESARROLLO EN PYTHON DE UN *WEB SCRAPER* DE DATOS DE LOS PARTIDOS DE LA LFP

Roger Cervantes Sentenà & Rodrigo Rico Gómez

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS (PRA 1)

WEB SCRAPING

Máster Universitario en Ciencia de Datos

Universitat Oberta de Catalunya

9 de noviembre de 2020

Introducción

En este documento se presenta el trabajo llevado a cabo por los autores del mismo en la primera práctica de la asignatura *Tipología y ciclo de vida de los datos*. En ella, se pretende realizar un recolector de datos que extraiga la información interesante de una *web* para luego poder disponer de ella en formato CSV. Popularmente, este proceso es conocido con el nombre de *web scraping*.

En cada uno de los siguientes apartados se contesta a cada una de las preguntas planteadas en el enunciado de la práctica. Se comienza estableciendo el contexto en el cual se enmarca la realización del trabajo y porqué se ha elegido ese sitio *web* como fuente de los datos. Luego se realizará una descripción del *dataset* que se desea obtener como producto final, previa definición del título. Posteriormente, se mostrará el contenido, es decir, los campos o atributos que componen el *dataset*. Por último, dedicaremos un apartado a mostrar nuestro agradecimiento al propietario del sitio *web* que hemos usado para extraer la información necesaria, otro apartado a explicar en qué medida nos hemos sentido inspirados por este tema y por qué creemos que puede resultar de utilidad nuestro trabajo. Se concluirá especificando la licencia bajo la cuál publicamos nuestro *dataset*.

Al final de este documento se adjuntan dos anexos en los que se presenta el código de nuestro *web scraper* y una vista previa de los tres *datasets* generados.

1. Contexto

El trabajo desarrollado en esta práctica se enmarca dentro de la disciplina del *web scraping*. El *web scraping* es una técnica empleada para obtener información útil de un sitio *web* disponible en Internet, para ser utilizada posteriormente en un proyecto de datos [1]. En este caso, nos centramos en la etapa de extracción de la información, por tanto, la materia prima del trabajo es el sitio *web* seleccionado, y el producto final son los datos estructurados en tres *dataset* en formato *CSV*.

Concretamente, la temática del proyecto serán los datos de los partidos de la LFP (Liga de Fútbol Profesional de España). Por tanto, es necesario encontrar una fuente que nos provea de dichos datos en Internet. La primera pregunta que cualquier recolector de datos debe hacerse es si existe una página oficial que permita la descarga de los datos necesarios mediante una *API* [1]. Al no existir ninguna herramienta de tipo *API* que nos permita obtener la información, el *web scraping* se vuelve una alternativa interesante [1]. En este contexto de incapacidad de obtener la información de los partidos de la LFP a través de medios preparados para ello (*APIs*), presentamos en este trabajo el diseño de un *web scraper* que cumpla dicha función.

El sitio *web* elegido para obtener los datos es una página dedicada a recoger los datos principales de cada partido (local, visitante, resultado, fecha y hora, etc.), y que contiene enlaces a páginas que contienen información más detallada de cada partido en particular (como pueden ser las estadísticas comparativas, las alineaciones, los cambios, las tarjetas, etc.). Esta estructura se puede visualizar de forma esquemática en la Fig. 1.

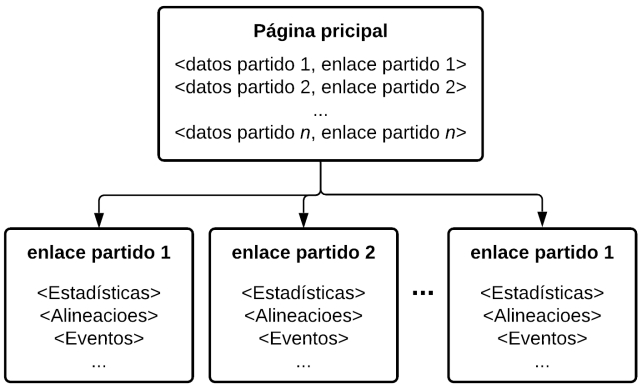

Figura 1: Estructura básica del sitio web.

Dada esta estructura, es posible acceder a toda la información que contiene el sitio *web* mediante técnicas de *web scraping*. El desarrollo del código se ha realizado en lenguaje *Python*, apoyándonos en la librería especializada BeautifulSoap para acceder a los distintos campos de la estructura *html* de las páginas *web* visitadas. Se puede acceder a la página principal a través de este enlace. Un ejemplo de página *web* de uno de los partidos se puede consultar aquí: (partido Granada - Athletic de Bilbao).

Tras inspeccionar e investigar varios sitios *web* alternativos que nos puedan aportar la misma información, se ha llegado a la conclusión de que éste es el mejor de los que se han barajado, debido a su estructura sencilla (Fig. 1) y al fácil acceso de la información interesante desde el formato *html*.

2. Definición del título

Por razones expuestas en el apartado 3, se ha decidido almacenar los datos de salida en tres *datasets* distintos. El título asignado a cada uno de ellos va en relación con los datos que alberga (Cuadro 1).

Datos	Título del dataset
Datos y estadísticas de cada partido	Partidos
Alineaciones	Alineaciones
Eventos del partido	Eventos

Cuadro 1: Título de cada *dataset*

Los nombres asignados a cada *dataset* hacen referencia a la entidad que representan. A continuación, se explica la existencia de los tres *datasets* y su descripción.

3. Descripción del *dataset*

Aunque el objetivo de la práctica es la obtención de un único *dataset*, dada la cantidad y la utilidad de la información encontrada, se ha decidido ampliar el número de *datasets* a tres. En el apartado anterior se ha comentado que los datos que extraeremos del sitio *web* serán las estadísticas de cada partido, las alineaciones y los eventos. Ante esta cantidad de información existen dos opciones:

- Almacenar toda ella en un mismo *dataset*.
- Separarla en función de sus peculiaridades.

Cada una de ellas tiene ventajas e inconvenientes. En *data warehousing*, a la primera opción se le conoce como estructura desnormalizada [2], que resulta más eficiente frente a consultas complejas pero almacena gran cantidad de información redundante, lo cual puede derivar en inconsistencias y falta de estabilidad [2]. A la segunda opción se le denomina estructura normalizada [2], y, aunque no es tan eficiente frente a consultas complejas como la anterior, su estructura resulta mucho más intuitiva y estable.
Como ejemplo, imaginemos que almacenamos toda la información en el mismo *dataset*, tendríamos en la misma tabla los eventos y las estadísticas. Si durante un partido ocurren alrededor de 20 eventos, tendríamos que almacenar todos los datos de las estadísticas 20 veces. La redundancia de información sería tal, que merece la pena dedicar un *dataset* a guardar cada partido como una fila con sus estadísticas correspondientes. Y en otro *dataset* distinto, tener un registro de cada evento ocurrido, donde el partido se encuentre referenciado por un *ID*.

Dicho esto, es preciso reaizar una descripción breve de los datos almacenados en cada uno de los *datasets*. La información detallada de cada uno de los campos se encuentra en el apartado 5.

Partidos:
El objetivo de este *dataset* es que guarde en cada fila la información de un partido. Dicha información serán los datos principales que describen el encuentro: equipo local, equipo visitante, fecha y hora, estadio, árbitro, etc. y las estadísticas comparativas de cada equipo: remates, faltas, tarjetas, saques de esquina, etc.

Alineaciones:
En este *dataset* cada fila es la participación de un jugador en un partido. Los atributos caracterizan cómo ha sido el rendimiento de dicho jugador en cada encuentro. El más conocido es la calificación que ha obtenido el jugador en el partido.

Eventos:
Por último, en este *dataset* cada fila representará, como su nombre indica, un evento ocurrido en un partido determinado. Los atributos que alberga sirven para caracterizar cada evento: partido en el que ocurrió, minuto, jugador o jugadores involucrados, etc.

4. Representación gráfica

En esta sección se muestra una imagen que describe visualmente el contenido de los tres *datasets*. Se puede consultar el esquema visual en la Fig. 2.

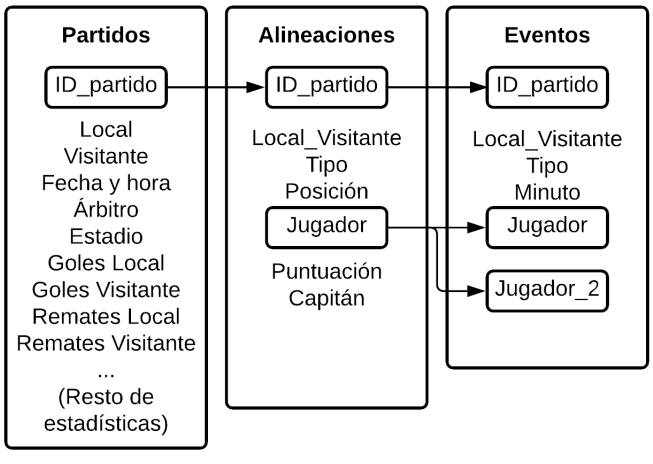

Figura 2: Esquema de los tres *datasets*.

Se ha considerado de interés remarcar aquellos campos que tienen en común varios *datasets*. Conocer esta información resulta de utilidad a la hora de hacer consultas a los datos almacenados. Especialmente cuando se trata de consultas complejas, ya que al dividir la información en tres *datasets* estamos perjudicando el rendimiento en situaciones como esas.

Referencias

[1] Subirats, L. & Calvo, M., *Web scraping*, Recursos de aprendizaje de la asignatura: *Tipología y ciclo de vida de los datos*, (2020), Editorial UOC, Barcelona, España.

[2] Abelló-Gamazo, A. & Curto-Díaz, J. & Rius-Gavidia, À. & Serra-Vizern, M. & Samos-Jiménez, J. & Vidal-Gil, J. & Díaz-Arias, D., *Introducción a las bases de datos analíticas*, Recursos de aprendizaje de la asignatura: *Diseño y uso de bases de datos analíticas*, (2020), Editorial UOC, Barcelona, España.