

Tipología y ciclo de vida de los datos
M2.851 - Aula 3

Práctica 1: Web Scraping

World Padel Tour

Rubén Gonzalo Coll Menéndez

Índice

| | |
|---|---|
| 1. Contexto..... | 2 |
| 2. Definir un título para el dataset..... | 2 |
| 3. Descripción del dataset..... | 2 |
| 4. Representación gráfica | 3 |
| 5. Contenido..... | 3 |
| 6. Agradecimientos | 5 |
| 7. Inspiración..... | 5 |
| 8. Licencia..... | 6 |
| 9. Código | 6 |
| 10. Dataset | 7 |
| 11. Contribuciones | 7 |

1. Contexto

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

Aunque el pádel es un deporte en crecimiento y con un gran potencial, que se está expandiendo de manera muy rápida entre los diferentes países Europeos, aún no tiene la misma popularidad que otros deportes mayoritarios lo que provoca que no sea tan fácil la obtención de información de los jugadores como lo es en otros deportes como el fútbol o la Fórmula 1.

La organización principal que está detrás del pádel mundial e impulsando dicho deporte es World Padel Tour; empresa privada que ha creado un circuito de pádel profesional y que a día de hoy es el más reconocido y que más fuerza está cogiendo, pudiéndose comparar con lo que serían los campeonatos del mundo de Fórmula 1 o de MotoGP.

Con el objetivo de ayudar a la difusión de este deporte me ha parecido interesante la creación de un dataset que permita a la interesados crear sus propias visualizaciones y hacer sus propios análisis acerca, tanto de la evolución de los diferentes jugadores, como del rendimiento de los mismos.

Para ello se ha accedido a la página del World Padel Tour (WPT) y aprovechando que no había ninguna restricción en el fichero “robots.txt”, se ha realizado el escaneo y descarga de la información que se ha considerado relevante para el fin comentado anteriormente.

2. Definir un título para el dataset

Elegir un título que sea descriptivo.

Dado que hemos distinguido dos datasets, uno para el ranking de hombres y otro para el ranking de mujeres, los títulos elegidos son para cada uno de ellos es:

- wpt_ranking_masculino.csv
- wpt_ranking_femenino.csv

3. Descripción del dataset

Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El dataset creado a partir de la ejecución de este proyecto contiene información tanto demográfica como del rendimiento de los jugadores de pádel que han participado en alguna de las competiciones organizadas por World Padel Tour desde el inicio de las mismas hasta la fecha de extracción.

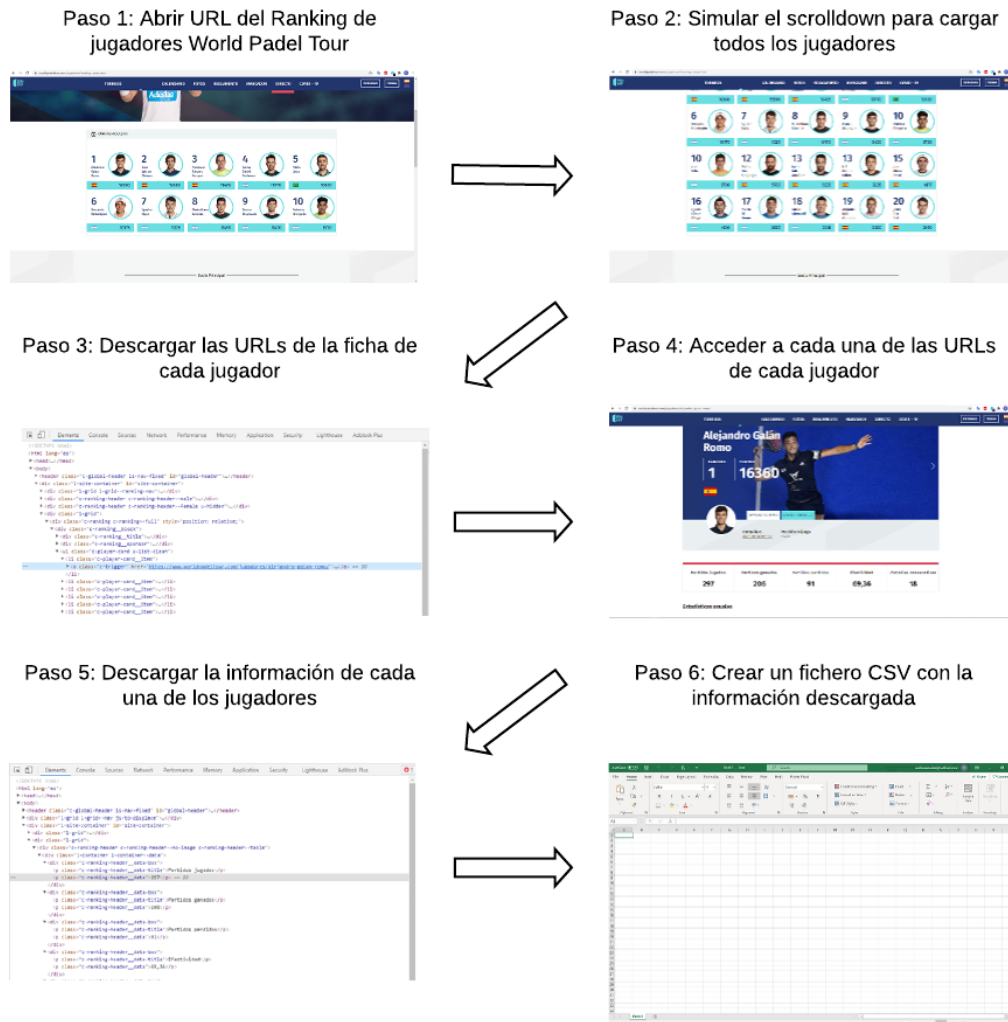
Es importante mencionar que WPT es considerada a día de hoy la competición de pádel más competitiva del mundo, en la que participan un mayor número de jugadores profesionales. Por lo tanto, con este dataset damos acceso a información al público a información de los mejores jugadores del mundo en este deporte.

Para más información sobre los campos que contiene véase el apartado 5 de este documento.

4. Representación gráfica

Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

A continuación se presenta un diagrama en el que se representa el proceso seguido para la extracción de la información de la página web seleccionada y la creación de ambos datasets.



5. Contenido

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Los dos datasets resultantes de este proyecto (“wpt_ranking_masculino o wpt_ranking_femenino”) recogen los mismos campos siendo la principal diferencia la distinción entre el ranking de hombres y de mujeres. Cada fila de los ficheros corresponde a un/a jugador/a y las columnas contienen una serie de campos que se detallan a continuación, recogiendo información desde el inicio de la creación del ranking hasta el momento de extracción.

Las dimensiones de las que consta el dataset son:

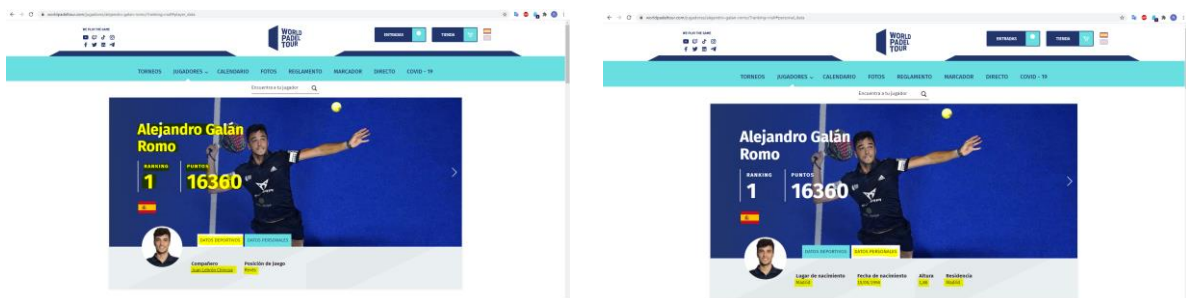
- nombre: Nombre completo del jugador.
- ranking_actual: Posición del ranking WPT a fecha de extracción de la información y según la última actualización (reference_date).

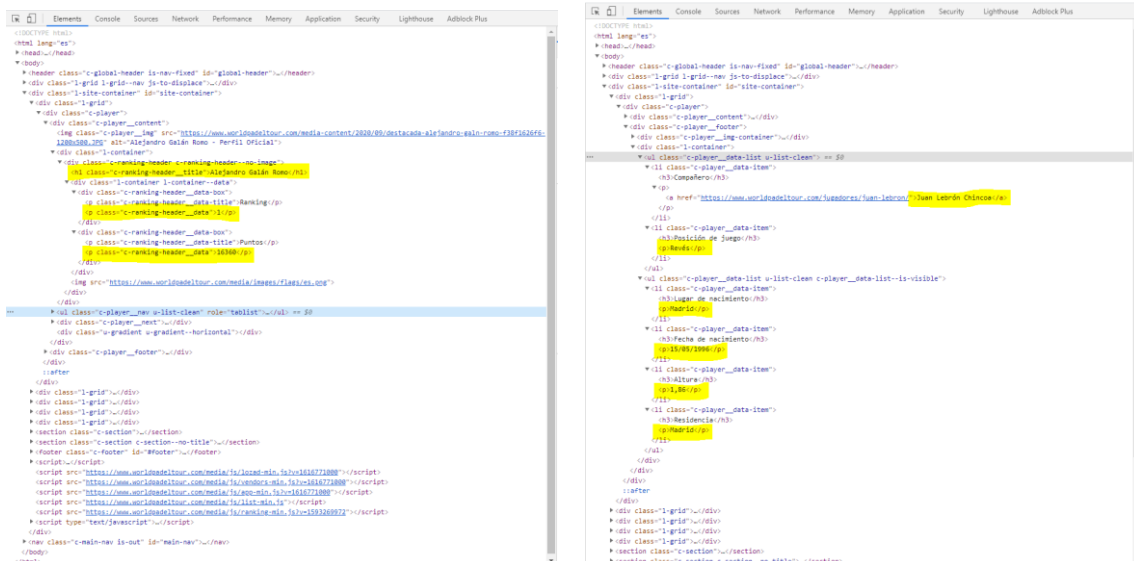
- puntos_wpt: Número total de puntos del jugador que le otorgan la posición anterior en el ranking.
- p_jugados: Número total de partidos jugados hasta la fecha en competiciones WPT.
- p_ganados: Número total de partidos ganados hasta la fecha en competiciones WPT.
- p_perdidos: Número total de partidos perdidos hasta la fecha en competiciones WPT.
- efectividad: Cálculo del número de partidos ganados entre el número total de partidos jugados.
- vict_consecutivas: Record de victorias consecutivas cosechadas por el jugador en competiciones WPT.
- companero_actual: Jugador con el que forma pareja.
- posicion_juego: Posición en la que juega el jugador. Puede ser “revés” o “derecha”.
- lugar_nacimiento: Ciudad de nacimiento del jugador.
- fecha_nacimiento: Fecha de nacimiento del jugador en formato “dd/mm/aaaa”.
- altura: Altura del jugador en metros.
- lugar_residencia: Ciudad de residencia del jugador.
- url: Enlace a la ficha del jugador de la cual se ha extraído la información.
- reference_date: Fecha a la que está actualizada la información según la página desde la que se ha extraído la información. No tiene porque coincidir con la fecha de extracción.

El método de extracción ha sido mediante la utilización de la técnica conocida como web scraping que consiste en la obtención de la información relevante a partir del código HTML de una página web. En nuestro caso, utilizando Python y la librería BeautifulSoup, hemos escaneado la página web del WPT y hemos automatizado la descarga de la información que hemos considerado necesaria para el análisis a realizar.

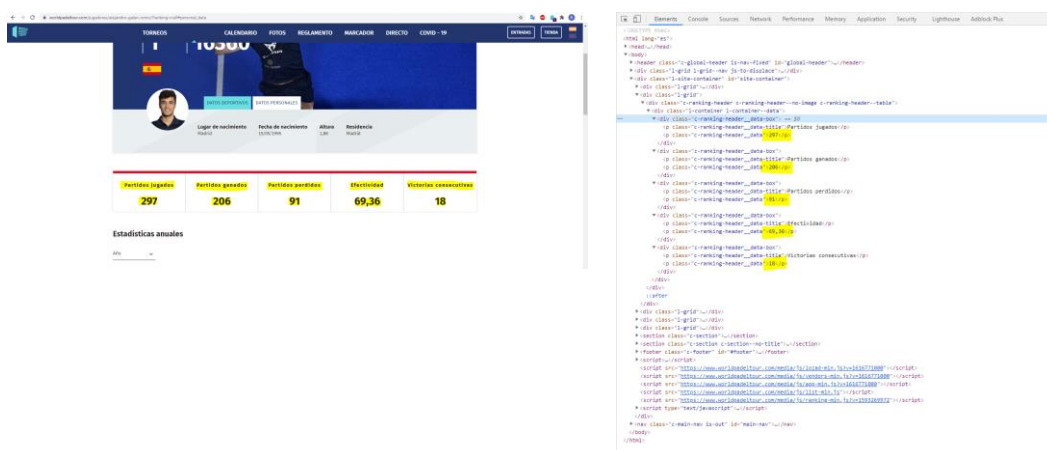
A continuación se detalla el origen dentro del código HTML de la página web, de cada uno de los campos extraídos en relación a la información de cada jugador y que componen el dataset.

- nombre, ranking_actual, puntos_wpt, compañero_actual, posición_juego, lugar_nacimiento, fecha_nacimiento, altura, lugar_residencia:





- p_jugados, p_ganados, p_perdidos, efectividad, vict_consecutivas:



6. Agradecimientos

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

Los datos han sido extraídos a partir del código HTML de la página web del World Padel Tour utilizando técnicas de “web scraping”, más concretamente con la librería BeautifulSoup escrita en Python que ha facilitado mucho la labor y nos ha permitido llevar a cabo con éxito este proyecto sin apenas conocimientos avanzados de Python.

Mención especial hay que hacer a WPT ya que sin la información que comparten en su página web, no habría sido posible obtener un dataset consolidado de los mejores jugadores de pádel de ninguna otra fuente.

7. Inspiración

Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Como se introducía en el punto 1 de esta práctica, el pádel es un deporte en crecimiento con un gran potencial. A pesar de ello, aún no está a la altura ni consta de la repercusión que otros deportes como pueden ser el fútbol, el baloncesto o la Fórmula 1 tienen. Es por ello que me parece interesante dar un primer paso en la creación de un dataset (al igual que ya existe para otros deportes) que permita a los usuarios interesados el análisis de dicha información.

Este dataset es solo un primer paso en la creación de un repositorio sobre pádel completo en el que conste toda la información relevante para el análisis de un jugador. En proyectos posteriores me gustaría desarrollar e incorporar al dataset, a partir del análisis de vídeo, estadísticas sobre los golpes ganadores y errores no forzados de los jugadores de manera que se pudiera tener una mejor idea de aquellos puntos fuertes y débiles de cada jugador basado en datos.

Algunas preguntas que el actual dataset pretende responder o aspectos que permite analizar son:

- Análisis del peso que tienen las diferentes nacionalidades en el ranking WPT.
- ¿Cuáles son los mejores jugadores de la historia del WPT?
- ¿Cuál es la distribución de la efectividad a lo largo del ranking?
- ¿Cuál es la media de puntos de los diferentes percentiles del ranking?
- ¿Cuál es la altura media de los jugadores de la posición de revés? ¿Y de derecha?

Cabe mencionar que a medida que se vayan extrayendo “snapshots” de la información en diferentes momentos del tiempo, se abrirá la puerta a realizar análisis mucho más completos, como por ejemplo el visualizar evoluciones a lo largo del tiempo de las diferentes métricas con el objetivo de entender mejor de dónde se viene y hacia dónde se va (p.ej. ¿cuál es la altura media del top 10 de jugadores de pádel ahora la altura de hace 5 años?).

8. Licencia

Seleccione una de estas licencias para su dataset y explique el motivo de su selección.

El tipo de licencia elegido para los datasets publicados es el de “Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)”.

El principal motivo para la elección del tipo de licencia anterior es que permite el uso de los datos libremente pero no su comercialización además de promover que cualquier otro estudio o trabajo basado en los datos proporcionados, se publique con el mismo tipo de licencia.

Dado que la extracción ha sido llevada a cabo para una actividad con fines educativos sobre la información obtenida de una empresa cuyo negocio esta basado en la popularización del deporte del pádel, no consideramos oportuno el poner a disposición del público un dataset que pueda ser utilizado con fines comerciales algo que creemos que queda perfectamente recogido con este tipo de licencia.

9. Código

Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

Se puede encontrar el código en el repositorio de github creado para este proyecto y al que se puede acceder a través del siguiente enlace:

https://github.com/rcollmenendezUOC/UOC_WebScraping_WPT.git




10. Dataset

Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

El data set se puede consultar en Zenodo (DOI: 10.5281/zenodo.4680125) siguiendo el siguiente enlace: <https://zenodo.org/record/4680125>

11. Contribuciones

A continuación se detallan los participantes de cada una de las secciones establecidas para esta práctica.

| Contribuciones | Firma |
|-----------------------------|--|
| Investigación previa | R.C.  |
| Redacción de las respuestas | R.C.  |
| Desarrollo código | R.C.  |