Tipología y ciclo de vida de los datos

PRA1 - Web Scraping

Alberto García Galindo y Federico Alejandro Floriano Pardal

7 de noviembre de 2020

Introducción

En este documento se expone el informe final del caso práctico resuelto de *web scraping* enmarcado en el contexto de la resolución de la práctica 1 de la asignatura **Tipología y ciclo de vida de los datos** del **Máster en Ciencia de Datos (*Data Science*)** de la **Universitat Oberta de Catalunya**.

En esta práctica se ha llevado a cabo la construcción de un conjunto de datos con información de los mejores jugadores de padel del mundo. En concreto, se extrajo información del sitio web del World Padel Tour y se recogió información tanto de la clasificación mundial masculina como femenina.

Para ello, se utilizó `Python` como herramienta principal, destacando el uso de la librería `Selenium`, especialmente útil a la hora de trabajar con páginas web dinámicas como a la que nos hemos enfrentado.

Contexto

A día de hoy, el pádel se sitúa como uno de los deportes más en auge de España. Cada vez son más las personas que disfrutan de él a todos los niveles y que lo practican de manera habitual. A nivel profesional, el circuito profesional del World Padel Tour se sitúa como el principal campeonato internacional, donde los mejores jugadores y jugadoras recorren el mundo compitiendo y proporcionando un espectáculo único.

No obstante, es un deporte relativamente reciente y que principalmente es practicado por hispanohablantes, por lo que su popularidad aun no llega a la de otras disciplinas como el tenis o el fútbol. En consecuencia, es uno de esos campos en los que la analítica de datos todavía no ha irrumpido con la misma magnitud que en otros ámbitos, y las contribuciones a nivel de bases de datos, investigación o tecnología son escasas.

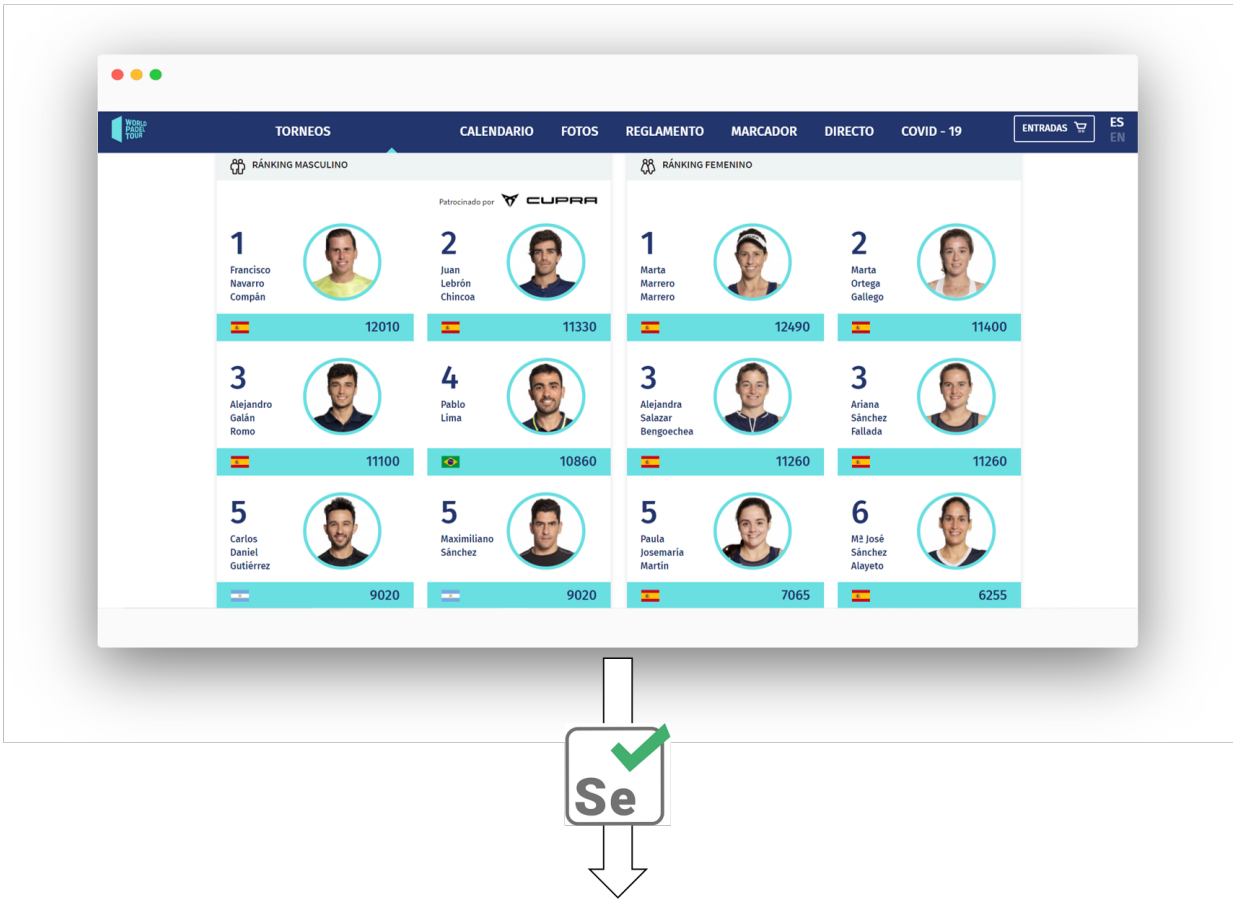
En este contexto se enmarca la elaboración de esta práctica, que pretende realizar un ejercicio de *web scraping* para extraer, de la página oficial de World Padel Tour, información tanto personal como deportiva de los mejores jugadores y jugadores del mundo, y construir un conjunto de datos con dicha información. El resultado del ejercicio supondrá el primer conjunto de datos abierto de jugadores de padel profesional y una de las primeras contribuciones a nivel internacional.

Título del conjunto de datos

Top 100 World Padel Tour Dataset

Descripción del conjunto de datos

Como resultado del ejercicio de *web scraping*, se ha obtenido un conjunto de datos que contiene la información de los 200 mejores jugadores de padel actuales del circuito World Padel Tour: los 100 mejores del ranking masculino y las 100 mejores del ranking femenino. Se extrayeron tanto características personales de los jugadores como deportivas, incluyendo datos relativos a su carrera profesional completa como los resultados de competeción más destacables obtenidos en los años mas recientes . En el apartado 5 se encuentra una descripción más detallada de los campos obtenidos del conjunto de datos.



Nombre	Posición	Puntos	Partidos jugados	Efectividad	...
Francisco Navarro Compán	1	12010	407	75.43	...
Juan Lebrón Chincoa	2	11330	302	63.25	...
Alejandro Galán Romo	3	11000	287	68.99	...
Pablo Lima	4	10680	423	85.58	...
...	...	...	...	...	...
Marta Marrero Marrero	1	12490	278	78.42	...
Marta Ortega Gallego	2	11400	260	70.00	...
Alejandra Salazar Bengoechea	3	11260	264	81.06	...
Ariana Sánchez Fallada	4	11260	231	67.53	...

200

63

Figure 1: Esquema seguido en el proyecto

Representación gráfica

En la Figura 1 del documento se incluye una representación gráfica que ilustra el proceso seguido para la realización del proyecto.

Contenido

A continuación se detallan cada una de las características que han sido recogidas para cada uno de los profesionales.

- **Nombre**: Nombre del jugador en el formato recogido por la página de World Padel Tour.
- **Posición**: Posición actualizada en el ránking World Padel Tour.
- **Puntos**: Puntuación actualizada según el sistema de puntuación de World Padel Tour.
- **Compañero**: Compañero profesional actual del jugador.
- **Pista**: Posición habitual en pista del jugador: *drive* o *revés*.
- **Lugar de nacimiento**: Lugar de nacimiento del jugador.
- **Fecha de nacimiento**: Fehca de nacimiento del jugador, en formato (DD/MM/AAAA).
- **Altura**: Altura del jugador en metros.
- **Residencia**: Lugar de residencia habitual del jugador.
- **Partidos jugados**: Número total de partidos jugados desde que se tienen registros en World Padel Tour.
- **Partidos ganados**: Número total de partidos ganados.
- **Partidos perdidos**: Número total de partidos perdidos.
- **Rendimiento**: Porcentaje total de victorias.
- **Racha**: Mayor racha de partidos ganados de manera consecutiva.
- **Circuito**. Circuito al que pertenece el jugador: *masculino* o *femenino*

Por otra parte, para cada uno de los años disponibles en cada jugador, se recogieron algunas características adicionales:

- **Partidos jugados**.
- **Partidos ganados**.
- **Partidos perdidos**.
- **Rendimiento**.
- **Torneos**: Número de torneos ganados.
- **Finales**: Número de finales alcanzadas (sin conseguir el título).

Las variables recogidas en el conjunto de datos en el momento de su extracción (7 de noviembre de 2020) se encuentran actualizados con fecha del 30 de marzo de 2020, tal y como se recoje en la página oficial. Dicho desfase se debe a la parada en la actualización en consecuencia de la pandemia causada por la enfermedad del SARS-CoV-2. Cabe mencionar que, en condicones normales, la frecuencia de actulización de la clasificación del World Padel Tour es semanal.

Agradecimientos

En primer lugar, nos gustaría agradecer a World Padel Tour la creación y mantenimiento de su página web oficial. No solo se agredece la facilidad a la hora de poder acceder a los datos de sus jugadores, sino también la disponibilidad de documentos adicionales para comprender, por ejemplo, el sistema de puntuación por torneos o las medidas adoptadas dada la situación excepcional de pandemia.

Por otra parte, nos gustaría agradecer también a la organización la permisividad a la hora de poder acceder al contenido de su sitio web, estableciendo muy pocas limitaciones a la hora de rastrearlas o, en nuestro caso, poder realizar de manera satisfactoria un ejercicio de *web scraping*. En respuesta, los autores hemos intentado en todo momento seguir unas buenas prácticas y unos principios éticos a la hora de realizar el proyecto, buscando no saturar el servidor del sitio web con un número muy elevado de peticiones consecutivas.

Finalmente, como en cualquier proyecto que utilice recursos *open source*, nos gustaría dar las gracias a todas las personas que hacen posible la existencia de herramientas como `Python` o ‘`Selenium`.

Inspiración

Existen una serie de factores que motivan a la creación del conjunto de datos reportado. Como ya hemos dicho, dicho conjunto de datos supone una de las primeras contribuciones en el ámbito del pádel en cuanto a material, lo que obviamente supone un añadido motivacional.

A nivel de análisis, surgen muchas posibilidades que pueden resultar interesantes. Por ejemplo, podemos obtener una visión comparativa entre jugadores en base a sus resultados a lo largo de su carrera, como por ejemplo, su efectividad o el número de partidos jugados. Por otra parte, al disponer de datos históricos de temporadas anteriores, podemos establecer análisis temporales que nos permitan conocer la evolución de algunos jugadores, y poder visualizar, por ejemplo, como han ido emergiendo a lo largo de los años algunos de los jugadores que hoy en día si sitúan entre los 10 mejores del mundo. Otra opción podría ser, dado un enfrentamiento en un torneo oficial, establecer una comparativa entre los 4 jugadores y conocer, por ejemplo, qué pareja acumula más partidos en la presente temporada, lo que podría traducirse en un mayor deterioro físico.

Todos estos ejemplos de análisis no parecen muy complejos. No obstante, podría resultar todo un añadido a la hora de complementar información adicional en la retransmisión de los partidos durante los torneos y enriquecer la experiencia del espectador.

Finalmente, nos gustaría comentar que este conjunto de datos podría utilizarse como base para futuras mejoras que incorporasen, por ejemplo, estadísticas específicas de golpes por partido y otros factores determinantes del juego. Toda esa información, no obsante, a día de hoy no puede obtenerse del sitio web oficial de World Padel Tour, y animamos a la organización a que algún día pueda recogerse y ser accesible para el público general.

Licencia

Tanto para el repositorio como para el conjunto de datos hemos decidio definir como licencia *MIT License*. Entendemos que, para un ámbito que tanto está por explorar y que tanto necesita de colaboradores que contribuyan al mismo, debemos de utilizar una licencia muy permisiva.

Código

El código para la realización del web scraping ha sido desarrollado en lenguaje Python, y se encuentra en el archivo XXXX.py dentro del repositorio.

Dataset

El dataset ha sido publicado en Zenodo con DOI...

Contribuciones

- Investigación previa: Alberto García Galindo y Federico Alejandro Floriano Pardal.
- Redacción de las respuestas: Alberto García Galindo y Federico Alejandro Floriano Pardal.
- Desarrollo del código: Alberto García Galindo y Federico Alejandro Floriano Pardal.