TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRÁCTICA 1

Realizado por:
Renán Freire Zurita Manuel García Guillén

CONTENIDO

INTRODUCCIÓN

Elaboramos este documento para completar la tarea Práctica 1: Web scraping correspondiente a la asignatura tipología y Ciclo de Vida de los Datos.

La práctica se centrará en un ejercicio de aplicación de esta técnica y la resolución de algunas cuestiones relacionadas con este tipo de prácticas.

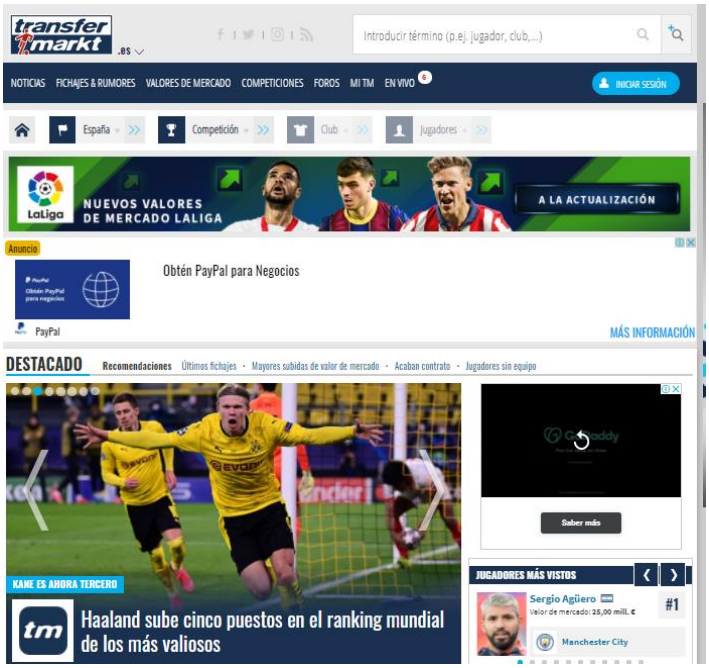
En concreto, nos centraremos en obtener a los futbolistas de los mejores 20 clubs de fútbol del mundo de la página https://www.transfermarkt.es/.

CONTEXTO

Dentro de los muchos usos de la Ciencia de Datos, hay una que nos parece interesante y es la aplicación de las técnicas de esta disciplina en el deporte, ya hemos visto películas como Moneyball, película dirigida en el 2011 y que profundiza en la aplicación real del big data y de la ciencia de datos para mejorar el rendimiento de un equipo de béisbol.

La industria del futbol por su naturaleza es muy impredecible, pero de alguna manera la aparición de big data en este deporte puede ser de gran ayuda para la extracción de información importante para la toma decisiones, donde los beneficiados serán principalmente los directores técnicos y sus asistentes, así como también del periodismo deportivo.

En este contexto, buscamos alguna versión parecida en el mundo del futbol y encontramos la base de datos en línea que provee la página https://www.transfermarkt.es/:



Esta página se está convirtiendo en un referente mundial en este deporte y en ella se publican datos de jugadores profesionales de todo el mundo, incluyendo su situación actual, datos físicos, datos de rendimiento y valoración económica en el mercado.

Como podemos leer en este artículo, incluso se toma como referencia para la compraventa de

futbolistas y como prueba exculpatoria en juicios.

https://www.mediotiempo.com/futbol/historia-transfermarkt-irreal-clubes-tomen-serio

Es por esto, que se nos hizo un tema apasionante intentar obtener un número reducido de los futbolistas que aquí aparecen e investigar un poco más el funcionamiento de esta página tan interesante.

DEFINIR UN TÍTULO PARA EL DATASET

El título de nuestro Dataset seria:

Estadísticas y Valor de Mercado Mejores Futbolistas 2021

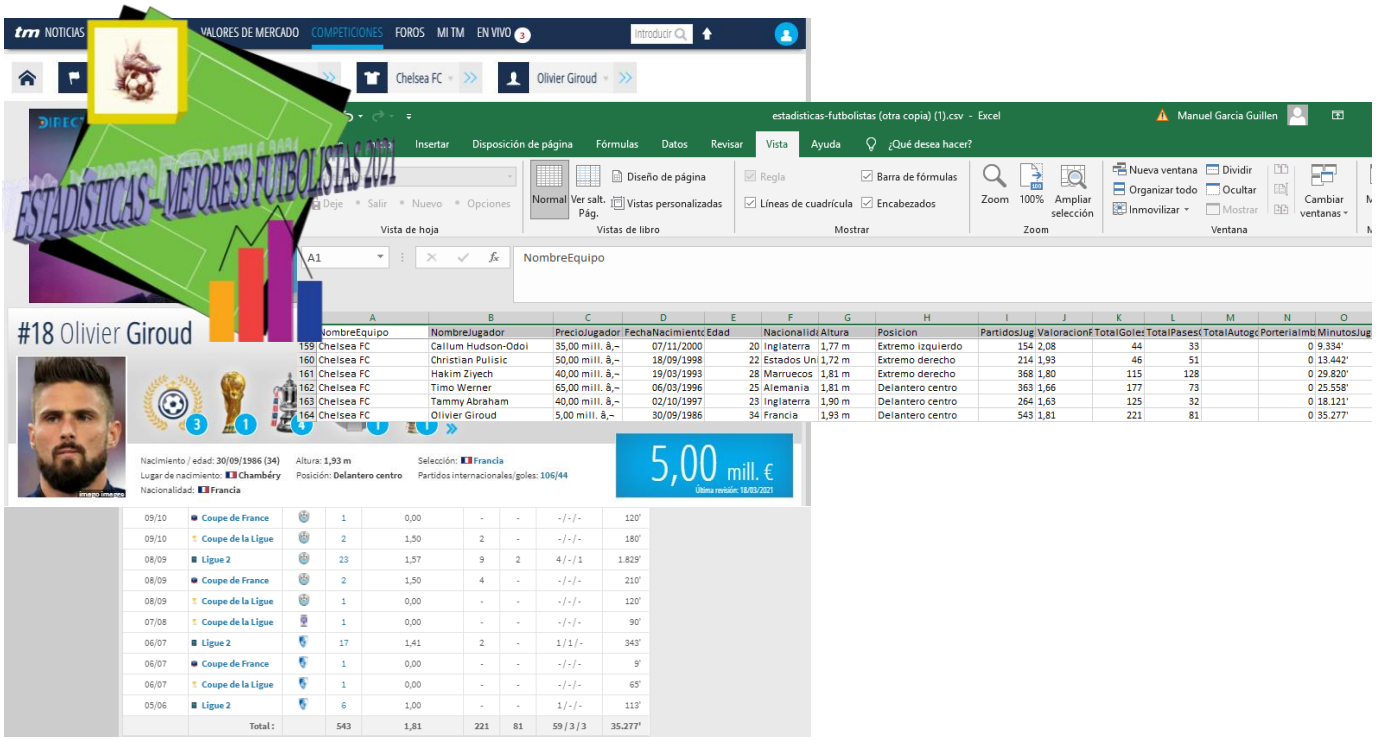
En él intentamos dar información del contenido de forma clara y concisa y fijarlo en su espacio temporal.

DESCRIPCIÓN DEL DATASET

Para obtener esta información, hemos partido del enlace dentro de https://www.transfermarkt.es/ que muestra los mejores 20 clubes del mundo. A partir de ahí, hemos rastreado a todos los futbolistas que militan en estos clubs, cada futbolista ocupara un registro dentro de nuestro dataset, junto con su nombre, club y atributos más importantes entre los que se pueden destacar su valor en el mercado y las diferentes variables de rendimiento deportivo (Totales goles, pases, minutos jugados, etc.) que resumen todas las temporadas que el jugador ha participado a lo largo de su vida futbolística.

REPRESENTACIÓN GRÁFICA

En estas imágenes podemos ver como la información de la página se llevó a un CSV



CONTENIDO

La página que estamos explorando, es una base de datos con más de 833.000 jugadores, nosotros nos hemos limitado a conseguir la información de los jugadores pertenecientes a los 20 clubes más valiosos del mundo.

Partiendo de la lista de estos clubes, hemos navegado por sus plantillas, entrando en la página de cada futbolista y obteniendo sus datos de procedencia, altura, estadísticas sobre su juego a lo largo de su carrera y el valor actual de mercado, todo ello en el momento más actual posible, esto es marzo del 2021.

Cada registro de nuestro dataset tendrá la siguiente información:

NombreEquipo: Nombre del Club al que pertenece actualmente el deportista, campo String

NombreJugador: Nombre del futbolista, campo String

Preciojugador: Valor de mercado del deportista, campo String

FechaNacimiento: Fecha de nacimiento del futbolista, campo en formato Date

Edad: Edad del futbolista, campo en formato Integer

Nacionalidad: País de nacimiento, campo de tipo String

Altura: Altura del futbolista, campo de tipo Double, con dos decimales, se corresponde con la altura en metros.

Posición: Posición que ocupa el futbolista en el campo, campo de tipo String.

PartidosJugados: Número de partidos que el futbolista fue parte de la alineación, campo de tipo Integer

ValoracionPromedio: Valoración promedio que ha recibido el jugador en todos los partidos en los cuales él participó, campo de tipo Double con dos decimales.

TotalGoles: Goles que anotó el futbolista a lo largo de su carrera, tipo de campo Integer.

TotalPasesGol: Asistencias que dio el futbolista a lo largo de su carrera, tipo de campo Integer.

MinutosJugados: Número de minutos jugados en su carrera, campo de tipo Integer.

Los siguientes campos, son exclusivos para porteros y por tanto para el resto de los jugadores tendrán un 0.

TotalGolesRecibidos: Número de goles que encajó, campo de tipo Integer.

PorteriaImbatida: Número de partidos en los que el portero no recibió goles, campo de tipo Integer.

AGRADECIMIENTOS

Los datos que se han recogido de la página https://www.transfermarkt.es/ son propiedad de Matthias Seidel y la compañía Axel Salmer publishing house. En especial hacemos un agradecimiento al sitio web por permitir el libre scrapring de su página especificado en su archivo de reglas robots.txt.

Para su extracción se han utilizado las librerías de Python BeautifulSoup, request y selenium junto con las técnicas de web scraping aprendidas en esta Maestría.

INSPIRACIÓN

Como comentamos en el apartado Contexto, nos interesa mucho la posibilidad de aplicar Ciencia de Datos y técnicas de Big Data al mundo del fútbol.

Llegados a este punto nos hacemos las siguientes preguntas:

- ¿Es posible predecir mediante Machine Learning el valor de mercado futuro de un profesional del fútbol con estos datos?
- ¿Sería posible predecir mediante estas técnicas, futuras estrellas a bajo precio?
- ¿Cabría la posibilidad de componer equipos a medida a bajo costo?
- ¿Podría ser posible buscar posiciones alternativas no habituales a los jugadores?
- Determinar si existe alguna relación o dependencia significativa de las variables de rendimiento de un jugador con respecto a su valor de mercado.
- ¿Es posible sugerir por medio de análisis predictivo un 11 titular modesto para poder ganar una liga?
- ¿Es posible mediante técnicas de minería de datos determinar cuál es el mejor equipo no por su valor de mercado si no por las diferentes características de rendimiento de sus jugadores?

Como trabajo futuro se podría navegar más a fondo en la página web para extracción de datos más complejos como opiniones, comentarios sobre cierto tema o jugador y poder realizar un análisis de sentimientos y entender la percepción de los usuarios.

LICENCIA

Para nuestro dataset daremos la licencia Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Nuestro motivo de publicación es meramente divulgativo y dentro del marco de una práctica de maestría, esta licencia permite al consumidor del dataset, compartir, copiar y redistribuir el material en cualquier medio o formato, también le permite remezclar, transformar y construir a partir del material bajo las mismas condiciones de acceso, uso y exclusivamente para fines académicos y científicos.

Además, el consumidor nos debe dar crédito adecuadamente, indicando si se han realizado cambios y de forma que no parezca que tiene nuestro apoyo para ello.

Sin embargo, el consumidor no podrá hacer uso comercial de estos datos y en caso de compartir cualquier resultado obtenido de nuestros datos debe hacerlo bajo la misma licencia.

En conclusión, buscamos que los datos puedan ser utilizados libremente, si es posible, cierto reconocimiento, pero en ningún caso lucrar con este trabajo estudiantil.

CÓDIGO

Para la práctica de WebScraping, se creó un proyecto colaborativo en un repositorio público de Github, La URL del mismo es la siguiente:

https://github.com/rfreirez/web-scraping-transfermarkt

DATASET

Aún no nos damos de alta en ZENODO porque nuestros datos necesitan pasar por un proceso de limpieza para que sean aceptados por el curador de la página.

BIBLIOGRAFÍA

- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Selenium with Python https://selenium-python.readthedocs.io/

TRABAJO COLABORATIVO

Contribuciones	Firma
Investigación Previa	Renán Freire Zurita Manuel García Guillén
Redacción de las respuestas	Renán Freire Zurita Manuel García Guillén
Desarrollo del código	Renán Freire Zurita Manuel García Guillén