CAPTURA MASIVA DE INFORMACIÓN ESTRUCTURADA

Antes de realizar la captura masiva de información se tuvo que elaborar un catálogo de datos con fuentes fiables donde poder conseguirla. Tras las recomendaciones de los tutores y una búsqueda exhaustiva se obtuvieron las siguientes fuentes:

|  |
| --- |
| * <https://fbref.com/es/> |
| * <https://www.transfermarkt.es/> |
| * <https://es.whoscored.com/> |
| * <https://www.footballdatabase.eu/> |
| * <https://one-versus-one.com/es/> |
| * <https://www.livescore.com/en/> |
|  |

A continuación, se analizaron las distintas fuentes contemplando lo que podían ofrecer cada una, y en una primera extracción se decidió utilizar la fuente <https://fbref.com/es/>

Esta fuente es la que más estadísticas a nivel técnico-táctico ofrece contando con la posibilidad de generar reportes de scouting y de crear widgets personalizados con el tipo de estadísticas que deseen obtener.

En un enfoque más técnico, para realizar el scraping de datos de FBREF, se crearon widgets\* personalizados(\*urls dentro de la propia página que dan acceso a determinados datos en función de los filtros que se apliquen) para obtener todos las datos generadas por un jugador perteneciente a las 5 ligas Europeas de Fútbol más importantes: La Liga(España) , Serie A(Italia), Ligue 1(Francia), Bundesliga(Alemania).

Ejemplos de widget que se ha creado en FBREF:

* 'https://widgets.sports-reference.com/wg.fcgi?css=1&site=fb&url=%2Fen%2Fcomps%2F{league\_name\_short}%2F{season}%2Fshooting%2Fplayers%2F{season}-{league\_name\_long}&div=div\_stats\_shooting'

Con este widget se accede al código html en forma de tabla de FBREF donde se proporcionan las estadísticas de disparos en función de la temporada y liga que se desee. En mi caso, realizo una extracción por temporada para no saturar el server y tener más controlada la ejecución en Collab y cuanto a las ligas se seleccionaron las 5 más importantes de Europa ya mencionadas anteriormente.

Teniendo esto la extracción es muy sencilla, basta con utilizar la función read\_html de pandas pasándole como input el link del widget, y generar un Dataframe con la salida.

En este ejemplo de widget se obtenían estadísticas de disparo, pero también se han obtenido otro tipo de estadísticas como por ejemplo de posesión, regate, defensa, tiempo de juego o incluso estadísticas de porteros los cuales pueden generar otros datos que los jugadores de campo no generan como por ejemplo: % de paradas, precisión en saque de puerta o goles encajados.

Finalmente se realiza una concatenación de todos los dataframes obtenidos por cada widget y se obtiene como salida un dataframe final con una gran cantidad de variables, que se exporta en csv y está preparado para el posterior preproceso y transformación.

Explicar ahora aquí a parte de como he hecho el cuaderno de twitter análisis