

Práctica 1

Tipología y Ciclo de vida de los datos

Presentación:

Se ha realizado una wiki en el repositorio de Github en el siguiente enlace: <https://github.com/mishuvale91/basket-statistic-scraper/wiki>, donde se encuentra los nombres de los integrantes del grupo y una descripción de los ficheros que se han utilizado para el desarrollo de la práctica.

Descripción de la Práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

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

Se ha seleccionado la información estadística de liga española e italiana de basketball, de las siguientes páginas web:

- **Liga española:** <http://acb.com/club/estadisticas/id/>
- **Liga italiana:**
<http://web.legabasket.it/team/tbd.phtml?from=2020&to=2020&club=MIO&type=d1>.

Las mismas que permiten visualizar datos estadísticos de las temporadas, por jugador, por fases, y equipo local.

La información se ha recolectado en el contexto de obtener una base de datos de jugadores por equipo local de los cinco últimos años.

Los sitios web elegidos proporcionan información muy relevante ya que están diseñadas para que los usuarios externos puedan visualizar las estadísticas de los equipos locales de la liga española e italiana, la misma que se visualiza en tablas resumidas.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

El título que se define para el dataset es el siguiente: **basketPlayerScraper**

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 **basketPlayerScraper** contiene datos estadísticos de los jugadores de basketball de la liga española e italiana que han sido extraídos desde su página oficial, a través del uso de webscapping utilizando el lenguaje de programación Python.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente



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

El dataset **basketPlayerScraper** contiene los siguientes campos:

1. **league:** variable de tipo carácter que indica la liga de la que se extrae los datos, es decir, española o italiana.
2. **game_date:** variable de tipo "date" que indica la fecha en la que se jugó el partido
3. **local_team:** variable de tipo carácter que indica el nombre del equipo local
4. **local_team_points:** variable de tipo entero que indica los puntos del equipo local

5. **visit_team:** variable de tipo carácter que indica el nombre del equipo visitante
6. **visit_team_points:** variable de tipo entero que indica los puntos del equipo visitante
7. **player_team:** variable de tipo carácter que indica si el jugador, jugaba en el equipo local o visitante
8. **player_name:** variable de tipo carácter que indica el nombre del jugador
9. **player_total_points:** variable de tipo entero que indica el total de puntos del jugador
10. **time:** variable de tipo “time” que indica los minutos y segundos jugados, en formato mm:ss
11. **one_point_shots_get:** variable de tipo entero que indica el número de tiros conseguidos de un punto
12. **one_point_shots_made:** variable de tipo entero que indica el número de tiros realizados de un punto
13. **two_point_shots_get:** variable de tipo entero que indica el número de tiros conseguidos de dos
14. **two_point_shots_made:** variable de tipo entero que indica el número de tiros realizados de dos puntos
15. **three_point_shots_get:** variable de tipo entero que indica el número de tiros conseguidos de tres puntos
16. **three_point_shots_made:** variable de tipo entero que indica el número de tiros realizados de tres puntos
17. **rebouts:** variable de tipo entero que indica el número de rebotes capturados
18. **assists:** variable de tipo entero que indica el número de asistencias repartidas
19. **fouls:** variable de tipo entero que indica el número de faltas cometidas
20. **received_fouls:** variable de tipo entero que indica el número de faltas recibidas

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Se tiene dos propietarios del conjunto de datos que son los siguientes:

1. La Asociación de Clubs de Baloncesto – ACB

Conocida por motivos de patrocinio como la Liga Endesa, es la principal liga de baloncesto profesional de España, tomó las riendas de la liga en la temporada de 1983 – 1984 en sustitución de la Federación (FEB). Actualmente consta de

18 equipos. Sus clubes han ganado la Euroliga, Eurocup, la Liga de Campeones y la Copa Europea de la FIBA.

Su página oficial es: <http://acb.com/>, creada desde el año 2001.

Con respecto a las condiciones legales del sitio, en la siguiente url: http://acb.com/Documentos/aviso_legal_acbcom.pdf, se encuentra un documento llamado “**aviso_legal_acbcom.pdf**”, que se lo puede descargar libremente.

En este sentido, el documento mencionado tiene como condiciones de uso de los datos, en el punto 3.- Derechos de propiedad intelectual y/o industrial, donde se indica lo siguiente:

“El Usuario se obliga a usar los contenidos de forma diligente y correcta, de acuerdo con la ley, la moral y el orden público. ACEB, S.A.U. autoriza al Usuario para visualizar la información que se contiene en este sitio web; así como para efectuar reproducciones privadas (simple actividad de descarga y almacenamiento en sus sistemas informáticos), siempre y cuando los elementos sean destinados únicamente al uso personal, así como su utilización exclusivamente con fines periodísticos, siempre que, en ambos casos, se respete la integridad del mismo y se identifique que la fuente originaria ha sido el presente Sitio Web: perteneciente a ACEB, S.A.U., quedando prohibido de manera expresa cualquier tipo de utilización sesgada y contraria a su naturaleza.”

Cabe mencionar que, por el párrafo expuesto, se sobre entiende que no existe ninguna prohibición en extraer los datos del sitio web ACB, siempre y cuando se cite el trabajo de la fuente originaria, como se lo ha realizado mediante la publicación en Zenodo. Además, nuestras intenciones de análisis de los datos son para la obtención de informes completos de los jugadores, y que se podrían incluir dentro del ámbito periodístico, por lo que se recalca que en nuestro caso no existe ningún fin de lucro.

Adicional, se puede indicar que la liga ACB tiene un “**robots.txt**”, prácticamente vacío, que solo contiene las siguientes líneas:

User-agent: Twitterbot
Disallow:

Las mismas, que no restringe el acceso a ningún directorio, ni especifican otro tipo de restricción que afecte al proceso de extracción de los datos.

2. La Lega Basket Serie A – LBA

Es una liga de baloncesto de clubes profesionales masculinos que se organiza en Italia desde 1920 y está dirigida por Lega Basket, que está regulada por la Federazioni Italiana Pallacanestro (FIP). La LBA juega bajo las reglas de la FIBA y actualmente consta de 17 equipos. Hoy en día es considerado como una parte superior de las ligas de baloncesto nacional Europea. Sus clubes han ganado la mayor cantidad de campeonatos de la Euroliga, la mayor cantidad Copas FIBA Saporta, y, la mayor cantidad de Copas FIBA Korac.

Su página oficial es: <http://www.legabasket.it/> , creada desde el año 2001.

En referencia a las condiciones legales no se ha encontrado ninguna página que contenga información al respecto, lo único que se especifica al pie de la página de su sitio web oficial es el siguiente texto: *“Tutti i diritti riservati. Legabasket.it”*. Por lo que, damos sobre entendido que al ser una liga que requiere de federación, los datos de los partidos son publicados en la prensa, tanto escrita como digital.

En este sentido, nosotros hemos citado el origen de los datos tanto en el repositorio de código Github como en la publicación en Zenodo.

Además, la liga Lega no tiene robots.txt en base del dominio.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El conjunto de datos es interesante como primer punto debido a que a los integrantes del grupo nos gusta el basketball, como segundo punto, la información se encuentra explícita en datos estadísticos, fáciles de manipular.

Con este conjunto de datos se pretende responder las siguientes preguntas:

1. ¿Qué jugador dentro de los últimos cinco años ha realizado más tiros de un punto?
2. ¿Qué jugador dentro de los últimos cinco años ha realizado más tiros de dos puntos?
3. ¿Qué jugador dentro de los últimos cinco años ha realizado más tiros de tres puntos?
4. ¿Qué jugador dentro de los últimos cinco años ha recibido más faltas?
5. ¿Qué jugador dentro de los últimos cinco años ha cometido faltas?
6. ¿Cuál ha sido el jugador estrella en el año 2019?

8. Licencia. Seleccione una de estas licencias para su dataset y explique el

motivo de su selección:

La licencia que se ha elegido para el dataset es:

Released Under CC0: Public Domain License

El motivo de la selección de esta licencia se debe a que cuenta con menos restricciones, es decir, que se otorgan derechos de dominio público.

Se utilizan para hacer que los trabajos con derechos de autor sean utilizables por cualquier persona sin condiciones, evitando las complejidades con otras licencias. No se requiere ningún permiso o licencia para un trabajo en el dominio público, el mismo que puede ser copiado a voluntad.

En este sentido, a los integrantes del grupo no nos interesa que se nos cite cuando usamos los datos, ni que se restrinja su uso comercial, tal como, se hace utilizando la licencias de: Released Under CC BY-NC-SA 4.0 License y Released Under CC BY-SA 4.0 License

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

Para realizar la práctica se eligió como lenguaje de programación Python, ya que se tenía experiencia previa con este lenguaje.

Se empezó el proyecto analizando varias páginas web de ligas profesionales de basketball, la idea es hacer scraping a la mayor cantidad de páginas posibles, para obtener datos de todos los jugadores incluso si estos se movieran de una liga a otra.

Se abordó las ligas nacional ACB y Lega, ya que son las más importantes a nivel europeo, y al ser aficionados al baloncesto teníamos claro desde un principio que datos como mínimos queríamos obtener por cada jugador.

Primero se comenzó por mirar la página de la liga ACB, priorizando los datos estadísticos de los partidos, vimos que era factible la extracción de los datos, una vez visto de donde podríamos extraer los datos, pensamos como obtenerlos por toda una temporada, por lo que encontramos que ya existía una búsqueda que los desplegaba, obtuvimos el identificador del partido para realizar la búsqueda. Con esto solo nos quedaba saber que equipos jugaban en la liga por cada temporada, ya que la página ya contaba con la búsqueda que se necesitaba.

Una vez, que se fue desarrollando el código, nos dimos cuenta que se estaba extrayendo los datos duplicados, de igual forma al extraer todos los partidos de un equipo para una temporada se estaba obteniendo parte de los partidos de los otros equipos contra los que jugaba. Para evitar esta duplicidad decidimos que solo se iba a obtener los datos de los partidos jugados como local por cada equipo, para esta parte se tuvo que desarrollar código exclusivamente que realice este filtro, ya que la página no realiza esta búsqueda de partidos.

Para extraer los datos de la liga Lega se siguió un camino parecido en la investigación de la página, pero, para esto ya se tenía claro que páginas nos hacían falta, ya que esta liga no tiene una página que buscara equipos que juegan cada temporada, pero si era posible extraer todos los equipos que juegan o jugaron en esta liga. Así que para cada temporada hicimos la búsqueda de partidos para todos los equipos, y, aquellos que no juegan para esta temporada no devolverían datos de los partidos.

Sin embargo, la liga Lega si permite filtrar la búsqueda para que solo se desplieguen los equipos que juegan como local.

Adicional, se analizaron otras ligas como:

- **Euroliga:** <https://www.euroleague.net/>
- **BBL, liga alemana:** <https://www.easycredit-bbl.de/de/>
- **Liga griega:** <http://www.esake.gr/>

Parece factible también extraer los datos de estas ligas, ya que todas contienen en sus páginas la posibilidad de consultar las estadísticas de los partidos, pero por el tiempo de la práctica no se las tomó como prioridad.

A continuación, se adjunta el enlace de GitHub, donde se encuentra el proyecto “**basket-statistic-scraper**”: <https://github.com/mishuvale91/basket-statistic-scraper>

También se cuenta con una carpeta “**src**” en la siguiente ruta <https://github.com/mishuvale91/basket-statistic-scraper/tree/main/src>, donde se han depositado los archivos python, que contienen el código que se ha utilizado para la extracción de la información de las páginas anteriormente mencionadas. Los cuales se detallan a continuación:

1. **ACBStatisticsScraper.py**

Este fichero se encarga de realizar la extracción de los datos de la liga de baloncesto ACB.

El método principal es ***getSeasonPlaters(season)***, esta llama al

método **getSeasonTeams**, para obtener todos los equipos de una temporada.

Después recorre equipo por equipo a través del método **getTeamGames**, que el mismo devuelve una lista de los partidos jugados como local por el equipo.

Para obtener cada partido se llama al método **getGamePlayers**, que se encarga de extraer finalmente las estadísticas de los jugadores que jugaron el partido respectivo.

2. LegaStatisticsScraper.py

Este fichero se encarga de realizar la extracción de los datos de la liga de baloncesto Lega.

El proceso de extracción de la información se lo realiza de igual forma como se detalló en el fichero “**ACBStatisticsScraper.py**”.

3. LeagueScraperFactory.py

Es un fichero que siguiendo el patrón “**Factory**”, devuelve una instancia del scraper de la liga que se pida. En la versión actual sólo devuelve scrapers para la liga ACB y Lega.

4. main.py

Este fichero es el programa principal, que debe ser ejecutado a través de la línea de comandos.

Se le utilizan tres parámetros:

1. **startSeason**: la temporada inicial desde la que se va a obtener los datos, para la práctica se toma los datos desde el año 2016.
2. **endSeason**: la temporada final hasta la que se va a obtener los datos, para la práctica se toma los datos hasta el año 2020.
3. **league**: se escoge la liga de donde se quiere obtener los datos. Para la práctica se ha decidido solo tomar la liga ABC y Lega (“abc-lega”).

Adicional, cabe mencionar que se pueden solicitar los datos de varias ligas separándolas por “-”.

Ejemplo de ejecución:

```
main.py --startSeason 2016 --endSeason 2020 --league lega
```

La ejecución de la instrucción es muy sencilla, se puede visualizar el llamado de los tres parámetros.

A continuación, se recorre las ligas seleccionadas en el parámetro **league**. Para cada

liga se obtiene una instancia del scraper y se llama al método **getSeasonPalyers**, para cada una de las temporadas se llama a los parámetros **startSeason** y **endSeason**.

Finalmente, el resultado de cada temporada es almacenado en un fichero con formato .csv.

Con la siguiente instrucción:

player_[league]_[startSeason]_[endSeason].csv

5. Utils.py

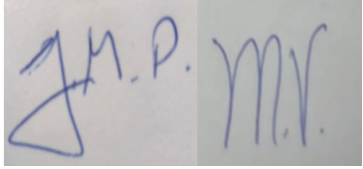
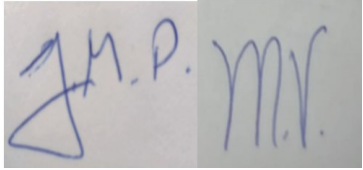
Es un fichero que contiene métodos de utilidades:

- **getArgs:** devuelve los argumentos pasados por línea de comandos en un objeto args.
- **getLogger:** devuelve una instancia del objeto logger.
- **getFilepath:** devuelve el nombre del fichero donde se va almacenar los datos.
- **writeToCSV:** escribe los datos extraídos al fichero csv.
- **getRequest:** realiza la petición get. Se encarga además de controlar el tiempo que tardan las peticiones y añadir un delay en caso de que excedan en cierto margen.

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

Para la publicación del dataset “**basketPlayerScraper**” en Zenodo se ha obtenido el siguiente DOI: **10.5281/zenodo.4253042**, que utiliza el siguiente link: <https://doi.org/10.5281/zenodo.4253042>

El dataset se encuentra publicado en formato .csv, que cuenta con información estadística de las ligas ABC y Lega con sus respectivas temporadas, jugadores, fases y sus equipos locales entre los años 2016 y 2020.

Contribuciones	Firma
Investigación previa	
Redacción de las respuestas	
Desarrollo código	