

Master Universitario en Ciencia de datos

Asignatura: Tipología y ciclo de vida de los datos aula 1

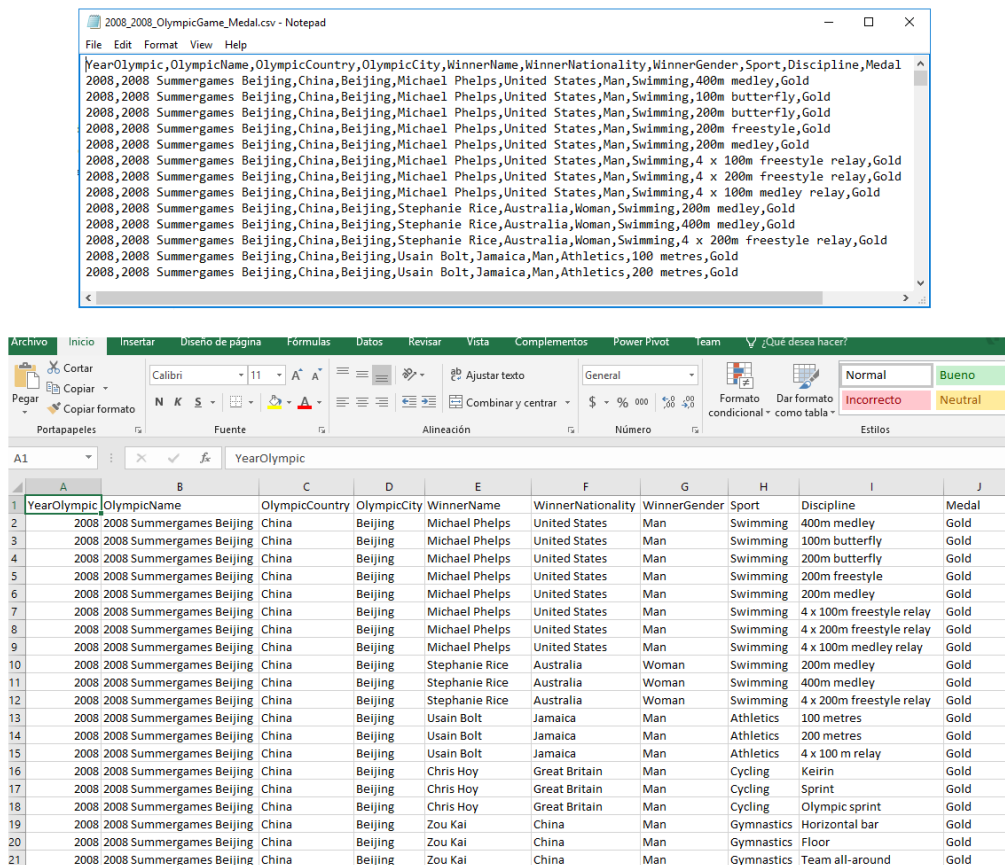
Practica 01: Web Scraping

Estudiante: José Ahias Lopez Portillo

Título del dataset: Ganadores de Juegos Olímpicos desde 1896 a 2008

Subtítulo del dataset: El programa desarrollado accede al sitio web: <http://www.theolympicdatabase.nl>, el cual contiene datos estadísticos de los juegos olímpicos realizados desde 1896 a 2008, por medio de un parámetro de entrada indica los años a extraer, se consulta los eventos realizados, ganadores de medallas y el detalle de las medalla que cada atleta olímpico gano en dicho juego, una vez consolidad toda la información se genera un DataSet en formato CSV.

Imagen.



The image shows two screenshots. The top one is a Notepad window displaying a CSV file named '2008_2008_OlympicGame_Medal.csv'. The bottom one is an Excel spreadsheet showing the data from the CSV file in a structured table format.

| YearOlympic | OlympicName | OlympicCountry | OlympicCity | WinnerName | WinnerNationality | WinnerGender | Sport | Discipline | Medal |
|-------------|--------------------------|----------------|-------------|----------------|-------------------|--------------|------------|--------------------------|-------|
| 2008 | 2008 Summergames Beijing | China | Beijing | Michael Phelps | United States | Man | Swimming | 400m medley | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Michael Phelps | United States | Man | Swimming | 100m butterfly | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Michael Phelps | United States | Man | Swimming | 200m butterfly | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Michael Phelps | United States | Man | Swimming | 200m freestyle | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Michael Phelps | United States | Man | Swimming | 200m medley | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Michael Phelps | United States | Man | Swimming | 4 x 100m freestyle relay | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Michael Phelps | United States | Man | Swimming | 4 x 200m freestyle relay | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Michael Phelps | United States | Man | Swimming | 4 x 100m medley relay | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Stephanie Rice | Australia | Woman | Swimming | 200m medley | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Stephanie Rice | Australia | Woman | Swimming | 400m medley | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Stephanie Rice | Australia | Woman | Swimming | 4 x 200m freestyle relay | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Usain Bolt | Jamaica | Man | Athletics | 100 metres | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Usain Bolt | Jamaica | Man | Athletics | 200 metres | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Usain Bolt | Jamaica | Man | Athletics | 4 x 100 m relay | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Chris Hoy | Great Britain | Man | Cycling | Keirin | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Chris Hoy | Great Britain | Man | Cycling | Sprint | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Chris Hoy | Great Britain | Man | Cycling | Olympic sprint | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Zou Kai | China | Man | Gymnastics | Horizontal bar | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Zou Kai | China | Man | Gymnastics | Floor | Gold |
| 2008 | 2008 Summergames Beijing | China | Beijing | Zou Kai | China | Man | Gymnastics | Team all-around | Gold |

¿Cuál es la materia del conjunto de datos?

Información detalla de las medallas ganadas en los juegos olímpicos según los años que el usuario quiere consultar, la información disponible de los juegos olímpicos desde 1896 a 2008.

¿Qué campos incluye?

| Nombre Columna | Descripción |
|--------------------------|--|
| YearOlympic | Almacena el año en que se realizó el juego olímpico. |
| OlympicName | Almacena una descripción del juego olímpico. |
| OlympicCountry | Almacena el país en donde se realizó el juego olímpico |
| OlympicCity | Almacena la ciudad donde se realizó el juego olímpico |
| WinnerName | Almacena el nombre del atleta ganador |
| WinnerNationality | Almacena la nacionalidad del atleta ganador |
| WinnerGender | Almacena el género del atleta ganador |
| Sport | Almacena el deporte en que el atleta gano la medalla olímpica |
| Discipline | Almacena la disciplina deportiva en que gano la medalla olímpica |
| Medal | Almacena el tipo de medalla ganada |

¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?

Inicialmente los juegos olímpicos se desarrollaban cada 4 años, teniendo una edición de verano y invierno, posteriormente los juegos olímpicos se realizan cada 2 años.

Theolympicdatabase, ha realizado una recolección de datos de los resultados de los juegos olímpicos desde 1896 a 2008.

¿Quién es propietario del conjunto de datos?

Los juegos olímpicos son un evento internacional organizado por Federaciones Internacionales de cada deporte, Comités Olímpicos Nacionales y Comités Organizadores de cada edición, la información generada de cada juego olímpico es de carácter público.

En el desarrollo de la practica se accede a la información recolectada por **Theolympicdatabase**, la cual es una organización independiente no relacionada al comité olímpico internacional o cualquier institución organizadora.

¿Por qué es interesante este conjunto de datos?

Los datos estadísticos que incluyen a más de un país siempre son datos de gran interés de análisis, en el caso de los juegos olímpicos es una base de datos muy rica, exacta y con diversidad de información con historia de 112 años.

¿Qué preguntas le gustaría responder la comunidad?

- ¿Quién es el país más ganador de medallas?
- ¿Quién es el mejor atleta de todos los tiempos?
- ¿Qué deporte tiene más disciplinas?
- ¿Quiénes ganan más medallas hombre o mujeres?
- Distribución de medallas por países y juego olímpico.
- Variación del total ganado de medallas por juego olímpicos.

Licencia GNU General Public License v3.0 (GNU General Public License)

Es la licencia mas utilizada en el mundo de software libre y código abierto, garantizando que futuros estudiantes o cualquier persona, tenga la libertad de compartir, copiar y modificar el código fuente desarrollado, al ser un programa de carácter académica con información publica, no quiero limitar el uso de lo desarrollado, siempre y cuando estén en los términos legales permitidos en la licencia.

URL de Código: <https://github.com/Ahias/WebScrapJuegosOlimpicos/tree/master/Fuentes>

- **WebScraJuegosOlimpicos.py** (Código fuente)
- **WebScraJuegosOlimpicos.cpython-36.pyc** (Código compilado)

Url DataSet: <https://github.com/Ahias/WebScrapJuegosOlimpicos/tree/master/DataSet>

- **1992_2008_OlympicGame_Medal.csv** (Dataset generado ingresando Año de inicio 1992 y Año de fin 2008)