

## PRÁCTICA 1. Web Scraping Euromillones

Javier Plo Moreno

Elena Segundo Martín

### CONTEXTO

El contexto corresponde a los resultados de sorteos de Euromillones desde 2004. La web de la cual obtendremos la información es <https://www.euromillones.com.es>. En concreto, dicha información es accesible desde la página <https://www.euromillones.com.es/historico/euromillones-anos-anteriores.html>.



COMO SE JUEGA ⓘ

SORTEOS ANTERIORES ⓘ

JUGAR A EUROMILLONES >

< El juego de Euromillones / Histórico sorteos

NÚMEROS PREMIADOS EN EUROMILLONES POR AÑO

2020	2019
2018	2017
2016	2015
2014	2013
2012	2011
2010	2009
2008	2007
2006	2005
2004	

Consulta las combinaciones premiadas en Euromillones desde el primer sorteo, celebrado el día 13 de febrero de 2004 hasta el último año completo.

RESULTADOS DE LOTERÍAS:



JUEGA EUROMILLONES

77.000.000 €

Aquí siempre hay suerte

JUGAR

La Primitiva

3.900.000 €

>

Bonoloto

500.000 €

>

Gordo Primitiva

6.700.000 €

>

Los resultados se encuentran agrupados por años. En la página inicial de resultados encontramos un enlace para cada año que nos lleva a la página HTML que contiene los resultados.



COMO SE JUEGA ⓘ

SORTEOS AN

◀ Euromillones 2015

Euromillones 2017 ▶

#### SORTEOS DE EUROMILLONES 2016

SEM.	SORTEO	DIA	COMBINACION GANADORA							EL MILLÓN
			NÚMEROS					ESTRELLAS		
1	002	5-ene	36	10	6	39	31	6	10	
	003	8-ene	35	33	26	40	5	3	8	
2	004	12-ene	2	10	30	44	1	1	8	
	005	15-ene	43	38	19	10	46	1	11	
3	006	19-ene	2	30	38	43	46	7	2	
	007	22-ene	27	10	30	47	12	9	8	
4	008	26-ene	15	40	24	48	38	2	9	
	009	29-ene	29	32	23	1	5	1	7	
5	010	2-feb	36	21	10	6	9	6	2	
	011	5-feb	46	32	27	3	41	4	8	
6	012	9-feb	9	6	13	28	37	4	5	
	013	12-feb	31	49	28	20	3	5	2	
7	014	16-feb	3	22	50	10	37	6	10	
	015	19-feb	32	13	14	39	30	3	9	
8	016	23-feb	42	32	23	25	37	11	1	
	017	26-feb	13	50	5	33	15	9	11	

Para cada año, los resultados se muestran en una tabla ordenados por fecha. Cada combinación ganadora consta de 5 números y otros dos números denominados estrellas.

## TÍTULO DEL DATASET

**Resultados\_Euromillones**

## DESCRIPCIÓN DEL DATASET

El conjunto de datos contiene el resultado de los sorteos del juego Euromillones que se han realizado desde el año 2004.

## REPRESENTACIÓN GRÁFICA

SORTEO
Anyo
Sorteo
Semana
Fecha
Numero1
Numero2
Numero3
Numero4
Numero5
Estrella1
Estrella2
ElMillon

## CONTENIDO

Cada registro del conjunto de datos corresponde a un sorteo de Euromillones. Podemos encontrar todos los sorteos desde el 13 de febrero de 2004 hasta el último sorteo realizado en el momento de la extracción de los datos.

El conjunto de datos incluye los siguientes campos:

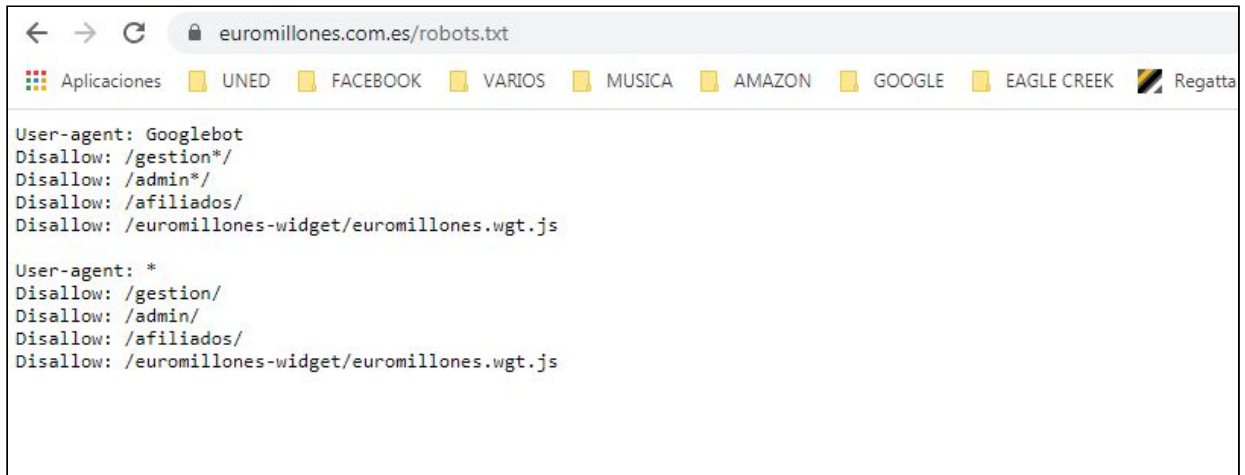
- Anyo: Indica el año del sorteo.
- Sorteo: Identificador del sorteo. Es único para cada año del sorteo.
- Semana: Número de la semana del año.
- Fecha: Día y mes del sorteo. Formato “DD-mes”.
- Numero1: Primer número premiado del sorteo.
- Numero2: Segundo número premiado del sorteo.
- Numero3: Tercer número premiado del sorteo.
- Numero4: Cuarto número premiado del sorteo.
- Numero5: Quinto número premiado del sorteo.
- Estrella1: Primer número premiado correspondiente a la denominada “Estrella” del sorteo.
- Estrella2: Segundo número premiado correspondiente a la denominada “Estrella” del sorteo.

- ElMillon: Número premiado correspondiente al denominado “El millón” del sorteo.

## Archivos Robots.txt y Sitemap

El sitio web no dispone de mapa.

Como vemos en la siguiente captura, el archivo Robots.txt no excluye de la búsqueda a las páginas sobre las que aplicaremos web scraping.



```
← → ↻ euromillones.com.es/robots.txt
Aplicaciones UNED FACEBOOK VARIOS MUSICA AMAZON GOOGLE EAGLE CREEK Regatta

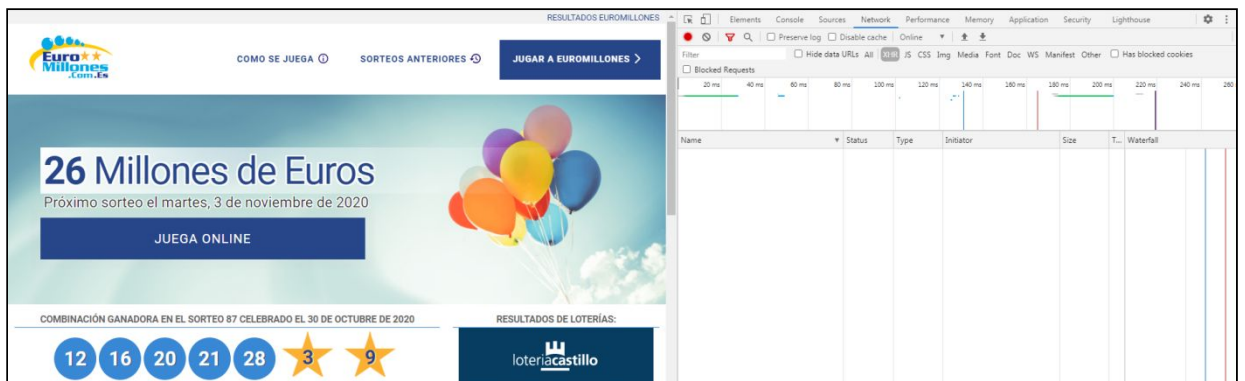
User-agent: Googlebot
Disallow: /gestion*/
Disallow: /admin*/
Disallow: /afiliados/
Disallow: /euromillones-widget/euromillones.wgt.js

User-agent: *
Disallow: /gestion/
Disallow: /admin/
Disallow: /afiliados/
Disallow: /euromillones-widget/euromillones.wgt.js
```

## API

<http://www.gregreda.com/2015/02/15/web-scraping-finding-the-api/>

En las developer tools de Chrome comprobamos mediante el menú Network/XHR, que no podemos realizar ningún tipo de llamada XMLHttpRequest (Get) para obtener algún tipo de información:



RESULTADOS EUROMILLONES

COMO SE JUEGA SORTEOS ANTERIORES JUGAR A EUROMILLONES

**26 Millones de Euros**  
Próximo sorteo el martes, 3 de noviembre de 2020  
JUEGA ONLINE

COMBINACIÓN GANADORA EN EL SORTEO 87 CELEBRADO EL 30 DE OCTUBRE DE 2020

12 16 20 21 28 3 9

RESULTADOS DE LOTERÍAS: loteríaCastillo

Network tab showing requests:

Name	Status	Type	Initiator	Size	T...	Waterfall
Blocked Requests						

### **Saturación de llamadas**

En nuestro caso, el número de llamadas será pequeño, ya que estamos manejando información de sobre 17 años, a 52 semanas por año, con dos sorteos semanales. Precisamos de una llamada inicial para obtener la página de resultados anteriores y una por cada año. Por este motivo no se considera ninguna estrategia para evitar la saturación de llamadas del servidor.

### **Modificar el User Agent**

No hemos encontrado llamadas en las que se restrinja el acceso. Es decir, el sitio no revisa dicha cabecera.

### **Chequeo del navegador**

No hemos tenido problemas a la hora de acceder a la información, por lo que no son necesarias técnicas como la de utilizar la navegación privada o el modo incógnito, para asegurar que el conjunto de cookies está vacío, o la utilización del comando *curl* para debugar casos complejos.

### **Asumir que el web scraper dejará de funcionar**

Hemos detectado algunas diferencias de estructura y de contenido dependiendo de los años, que se han solucionado en el código:

- *Algunas semanas tienen más de un sorteo*
- *Algunos años no tienen campo "Semana" (aquellos que solo tienen un sorteo por semana)*
- *Algunos años no tienen campo "El Millón"*
- *Algunos sorteo tienen más de un número de "El Millón"*

### **Calidad y robustez de los datos**

- *Compleitud*
  - o *Se almacenan todos los datos disponibles.*
- *Unicidad*
  - o *Los datos son únicos.*
- *Puntualidad*
  - o *Los datos representan la realidad fielmente. Aunque en la web se añaden información sobre próximos sorteos no realizados en la que la información de los números premiados aparece en blanco, informando sobre la semana, el número del sorteo y la fecha.*
- *Validez*
  - o *Los datos son válidos*
- *Exactitud:*
  - o *Los datos son exactos*
- *Consistencia*

- o Los datos son consistentes

### Aspectos legales

Se revisa la página <https://www.euromillones.com.es/legal.html> para asegurarnos de que no hay ningún tipo de alusión a condiciones sobre el web scraper.

### Beautifulsoup vs Selenium

Se presentan dos versiones del código, una que no utiliza Selenium y otra que hace un uso combinado de esta librería junto con BeautifulSoup.

Selenium no fue diseñado para hacer web scraping. En realidad, fue desarrollado para pruebas web. Se utiliza para pruebas automatizadas de aplicaciones web. Automatiza los navegadores web y puede usarlo para realizar acciones en entornos de navegador en su nombre. Sin embargo, desde entonces se ha incorporado al web scraping.

Beautifulsoup es una buena herramienta para principiantes del web scraping, dispone de buena documentación, etc... En cambio tiene la desventaja de que depende en gran medida de otras librerías para funcionar, y que no tiene la capacidad de enviar solicitudes web; por lo que se debe utilizar el módulo *requests*, o el módulo estándar de Python para enviar dichas solicitudes.

Hemos encontrado muchas opiniones sobre cuándo usar uno u otro, y a modo de resumen:

*Se recomienda usar Selenium para cosas como interactuar con páginas web, ya sea en un navegador completo o en un navegador en modo sin headers, como Chrome sin headers. BeautifulSoup es mejor para observar y escribir declaraciones que dependen de si se encuentra un elemento o qué se encuentra, y luego usar Selenium para ejecutar tareas interactivas con la página.*

En nuestro caso, no existen tareas interactivas con la página, por lo que en la versión en la que hemos utilizado Selenium, lo hemos hecho en los siguientes puntos:

- Carga de la página inicial
- Obtención de los enlaces

Referencias:

- <https://stackoverflow.com/questions/17436014/selenium-versus-beautifulsoup-for-web-scraping>
- <https://www.bestproxyreviews.com/scrapy-vs-selenium-vs-beautifulsoup-for-web-scraping/>

## AGRADECIMIENTOS

MarinWbanet, S.L.L , es el titular de la web [www.euromillones.com.es](http://www.euromillones.com.es) y de los datos extraídos.

## INSPIRACIÓN

Puede ser interesante analizar los resultados para buscar los números más premiados, patrones, intentar predecir el número premiado o, como explica [este artículo](#), demostrar que no es posible predecir el número premiado dado que se trata de un proceso totalmente aleatorio y no existen factores externos para los que se pueda encontrar alguna correlación con los resultados.

## LICENCIA

### **Released Under CC BY-NC-SA 4.0 License**

Permite la reutilización y difusión siempre y cuando se haga referencia al autor de la misma y la obra no sufra cambio o alteración alguna.

## CÓDIGO

El código con el que se ha generado el dataset, desarrollado en Python, se encuentra el repositorio de GitHub: [https://github.com/esegundoUOC/Practica1\\_Visualizacion\\_datos](https://github.com/esegundoUOC/Practica1_Visualizacion_datos).

Como se ha explicado en el apartado [CONTENIDO](#), en el repositorio encontramos dos versiones:

- “Euromillones\_WebScraping.py” versión del código sin utilizar Selenium.
- “Euromillones\_WebScraping\_Selenium.py” versión del código utilizando Selenium.

## DATASET

El dataset ha sido publicado en formato CSV en Zenodo.

**DOI:** 10.5281/zenodo.4226114

**URL:** <https://zenodo.org/badge/DOI/10.5281/zenodo.4226114.svg>

Contribuciones	Firma
Investigación previa	JPM, ESM
Redacción de las respuestas	JPM, ESM
Desarrollo código	JPM, ESM