

Tipología y ciclo de vida de los datos

Prueba de Evaluación Continua
(PRACTICA1: Web Scraping)

Alumno: Cèsar Comas Solé

12 de noviembre de 2018

Universitat Oberta de Catalunya (UOC)	Alumno	César Comas Solé
Máster Universitario de Ciencia de Datos.	NIF	
Tipología y ciclo de vida de los datos		PRACTICA1

Índice de contenidos

1	Introducción	3
2	Principales datos del proyecto.....	3
3	Contexto	4
4	Contenido	4
5	Agradecimientos.....	5
6	Inspiración.....	5
7	Licencia.....	5
8	Referencias	5
9	Anexo: Trabajos preliminares	6
9.1	Preparación del entorno de trabajo	6
9.2	Concreción del objetivo.....	7
9.3	Identificación de los repositorios a rastrear	8
9.4	Análisis de restricciones al rastreo de las webs	9
9.5	Análisis de viabilidad técnica.....	12

Universitat Oberta de Catalunya (UOC)	Alumno	César Comas Solé
Máster Universitario de Ciencia de Datos.	NIF	
Tipología y ciclo de vida de los datos		PRACTICA1

1 Introducción

Muchas peñas y asociaciones deportivas españolas ofrecen entre sus servicios la participación conjunta de sus afiliados a loterías y juegos de azar. Las propias asociaciones gestionan en nombre de sus socios las apuestas en dichos sorteos, ofreciendo así la capacidad de mejorar las probabilidades de éxito gracias a la realización de apuestas combinadas.

Este proyecto propone la realización de un script de extracción automática de los resultados de la web oficial de "Loterías y Apuestas del Estado" que permita posteriormente a estas asociaciones automatizar el proceso de verificación de premiados.

Los sorteos de los que se ha preparado el script capturan sobre un dataset los resultados de los últimos sorteos de la Lotería Primitiva y del sorteo extraordinario, El Gordo de la Primitiva.

2 Principales datos del proyecto

- **Título:** "Web Scraping of Spanish 'La Primitiva' Lottery Results" (Web Scraping de los resultados del sorteo de la Lotería Primitiva)



- **Descripción breve (subtítulo):** Desarrollo de un script para la extracción automática de resultados de la lotería española La Primitiva del sitio web oficial: Loterías y Apuestas del Estado.
- **Equipo de desarrollo:** El proyecto ha sido realizada íntegramente por Cesar Comas Solé.
- **Github:** <https://github.com/ccomassole/web-scraping-spanish-lottery-results>

Universitat Oberta de Catalunya (UOC)	Alumno	César Comas Solé
Máster Universitario de Ciencia de Datos.	NIF	
Tipología y ciclo de vida de los datos		PRACTICA1

3 Contexto

El objetivo de este proyecto de web scraping es crear un servicio que permita la extracción automatizada y periódica de los resultados de la lotería española La Primitiva para posteriormente verificar automáticamente los ganadores de un hipotético servicio de reventa y publicar automáticamente los resultados en diferentes medios digitales (facebook, twitter y web).

De este modo, una posible peña deportiva que ofreciera el servicio, optimizaría sus procesos a la vez que mejoraría la propuesta de valor a sus asociados.

Los sorteos de La Primitiva se realizan los jueves (9:40 pm) y los sábados (9:40 pm). El Gordo de la Primitiva (que es un sorteo Premium) se sortea el domingo (9:30 p.m.). El sitio web oficial donde se publican los resultados de La Primitiva es Loterías y Apuestas del Estado, que pertenece a una empresa del gobierno español.

4 Contenido

Los datos finalmente extraídos se almacenan en un archivo CSV con los siguientes campos:

- **Lottery:** 'Clásico', 'ElGordo'
- **Date:** Fecha del sorteo.
- **Winner_Num_1:** Primer número ganador
- **Winner_Num_2:** Segundo número ganador
- **Winner_Num_3:** Tercer número ganador
- **Winner_Num_4:** Cuarto número otorgado
- **Winner_Num_5:** Quinto número ganador
- **Winner_Num_6:** Sexto número ganador (solo en modo clásico)
- **Complementary:** número complementario (solo modo clásico)
- **Refund:** Número que permite el reembolso del valor de la apuesta del boleto.
- **Joker:** conjunto de 6 dígitos ordenados asociados con un premio extraordinario (solo en modo clásico)

Para el propósito de este proyecto, los datos de los sorteos se han extraído entre las siguientes fechas: 16/8/18 - 10/11/18, ya que el objetivo no es la extracción de un gran volumen de información, sino la ejecución de un Script que automatice procesos semanales.

El script automatizado ha sido realizado en Python, aplicándose técnicas de Web Scraping. Las librerías asociadas utilizadas han sido BeautifulSoup y Selenium, utilizando el driver de Chrome.

Universitat Oberta de Catalunya (UOC)	Alumno	César Comas Solé
Máster Universitario de Ciencia de Datos.	NIF	
Tipología y ciclo de vida de los datos		PRACTICA1

La ejecución del script, además de las librerías de Python necesarias, requiere la instalación en la carpeta /drivers/ del driver para el control automatizado de Chrome.

5 Agradecimientos

Agradecemos a SELAE (Sociedad Española de Loterías y Apuestas del Estado), como propietaria de los datos, el uso de la información extraída de su sitio.

6 Inspiración

La inspiración que ha propiciado la idea de este proyecto nace de la startup de un amigo, el cual creó un servicio para expandir los juegos de la Lotería a asociaciones de vecinos.

7 Licencia

El presente código se comparte bajo la licencia: **CC0: Public Domain License**.

De este modo otorgamos libre uso del código sin necesidad de que se solicite permiso al respecto, y ya sea para uso comercial o no. Nos liberamos de este modo de ofrecer garantía respecto a la calidad del código y renunciamos a la responsabilidad sobre su uso.

Puesto que se trata de una primera experiencia de desarrollo, y puesto que no está en nuestros planteamientos iniciales hacer seguimiento de las mejoras o los usos que sobre la misma se realicen en el futuro, creemos que la mejor forma de liberar el código es a través de una licencia completamente abierta.

8 Referencias

- [1] Richard Lawson (2015), "Web Scraping with Python". Ed. Packt Publishing
- [2] Laia Subirats, Mireia Calvo (2018), "Web Scraping", UOC
- [3] Corey Shafer. Canal de YouTube.
<https://www.youtube.com/watch?v=ng2o98k983k>
- [4] ttguayco. <https://github.com/tteguayco/Web-scraping>

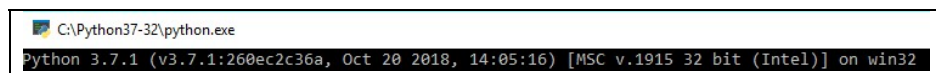
9 Anexo: Trabajos preliminares

Previa a la realización del proyecto se han desarrollado multiples tareas, tanto de preparación del entorno de desarrollo, como de análisis de la página web a scrapear. En este apartado se muestra el trabajo adicional realizado.

9.1 Preparación del entorno de trabajo

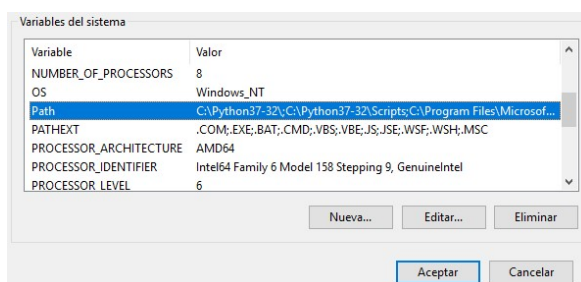
Preparamos el entorno de trabajo para asegurar la disponibilidad de los elementos tecnológicos necesarios, actualizando versiones para desarrollar la práctica:

- Lenguaje de programación: Instalamos Python v 3.7.1 (instalación directa). Lo ubicamos en la carpeta raíz: C:/Python37-32. Al ejecutar la consola de Python instalada nos devuelve la versión.



```
C:\Python37-32\python.exe
Python 3.7.1 (v3.7.1:260ec2c36a, Oct 20 2018, 14:05:16) [MSC v.1915 32 bit (Intel)] on win32
```

- Como editor utilizaremos el IDE: SublimeText3 que ya tenemos instalado y que preparamos para ejecutar Phyton a través de ANACONDA.
 - Abrimos SublimeText3
 - PREFERENCES -> PACKAGE CONTROL: Install Package
 - Escribimos "Anaconda" y esperamos hasta su instalación.
 - Aseguramos la interacción de Anaconda con Phyton:
 - Verificamos que la variable Path de nuestro entorno tiene la ruta del intérprete de la versión 3 de Python:



- En Sublime: TOOLS> BUILD SYSTEM>Automatic
 - En Sublime: TOOLS>BUILD WITH> Python
 - En Sublime: TOOLS>Save all On Build (check ON)
- Verificamos el funcionamiento creando en SublimeText un fichero Phyton y verificando que es interpretado y ejecutado adecuadamente:
 - FILE>New File

Universitat Oberta de Catalunya (UOC)	Alumno	César Comas Solé
Máster Universitario de Ciencia de Datos.	NIF	
Tipología y ciclo de vida de los datos		PRACTICA1

- Guardamos el fichero previo a lanzarlo.
- Lo lanzamos con "CTRL + B" (Shortcut de BUILD) y vemos que se ejecuta adecuadamente sobre la consola de SublimeText.

- Instalamos los paquetes relevantes de Python necesarios para la práctica desde la consola de windows, situándonos en la carpeta de nuestra instalación de Python y ejecutando "pip install [nombre_paquete]": (el listado completo de los paquetes de Python instalados puede obtenerse desde Sublime mediante la orden >help('modules')).
 - pip (gestor de paquetes de Python)
 - builtwith (verificar tecnología de una web)
 - whois (verificar propietario web)
 - requests (descarga de páginas)
 - BeautifulSoup4 (parser)
 - lxml (formato lxml)
 - html5lib (formato html5)
 - selenium (control remoto de navegadores. Requiere instalar driver para Chrome)

Fuentes consultadas:

- <https://www.python.org>
- https://www.youtube.com/watch?v=Y2q_b4ugPWk
- http://damnwidgit.github.io/anaconda/anaconda_settings/
- <https://www.youtube.com/watch?v=t1ZWknZN5kM>

9.2 Concreción del objetivo.

Se plantea como objetivo crear un servicio que permita actualizar periódicamente los resultados de la Lotería Primitiva.

Dichos resultados se publicarán a la página web y en el Facebook de nuestro cliente, una asociación futbolística que periódicamente participa en el sorteo con algunos de sus afiliados.

NOTA: El caso se plantea a modo de ejemplo (no se trata de un caso real)

9.3 Identificación de los repositorios a rastrear

La web oficial que publica periódicamente el resultado de los sorteos es: loterías y apuestas del estado: <https://www.loteriasyapuestas.es>

Los sorteos de La Primitiva se realizan los Jueves (21:40) y los Sábados (21:40). El Gordo de la Primitiva se sortea el Domingo (21:30).

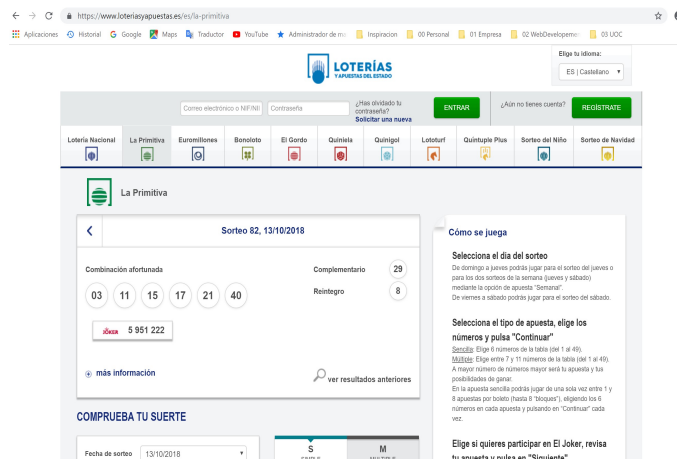
¡Sé el primero en conocer los resultados de nuestros juegos!

Aquí podrás ver EN DIRECTO la retransmisión de TODOS nuestros sorteos. Y a los pocos minutos, podrás consultar ya los resultados oficiales a través del enlace que te ofrecemos debajo. Te ofrecemos la mejor forma de saber, al minuto y con la garantía de Loterías y Apuestas del Estado, si tu apuesta ha tenido premio.

Horario de programación

- **Lunes** - a las 21:30 h sorteo de **BonoLoto**
- **Martes** - a las 21:30 h sorteo de **Euromillones** y **BonoLoto**
- **Miércoles** - a las 21:30 h sorteo de **BonoLoto**
- **Jueves** - a las 21:00 h aproximadamente sorteo de **Lotería Nacional**, a las 21:30 h sorteo de **BonoLoto** y a las 21:40 h sorteo de **La Primitiva** y **El Joker**.
- **Viernes** - a las 22:00 h aprox. programa 'La suerte en tus manos' que incluye los sorteos en diferido de **Euromillones**, **El Millón** y **BonoLoto**.
- **Sábado** - a las 13:20 h aprox. Sorteo de **Lotería Nacional**, en directo desde la extracción del Primer Premio e información de los premios anteriores. En las jornadas de los Sorteos Viajeros, se emitirá el programa aproximadamente a las 13.40 horas con un reportaje previo dedicado al municipio en el que se celebra. -a las 21:30 h sorteo de **BonoLoto** y a las 21:40 h sorteo de **La Primitiva** y **El Joker**.
- **Domingo** - a las 21:15 h sorteo de **Lototurf** -a las 21:30 h sorteo de **El Gordo de La Primitiva**
- Recuerda que sólo se dispone de señal durante la retransmisión de cada sorteo, por lo que el resto del tiempo el reproductor no mostrará imágenes.

Los resultados de los mismos se publican minutos después del sorteo en la web:



La url de acceso a la información es para el sorteo de la primitiva es:

- <https://www.loteriasyapuestas.es/es/la-primitiva>
- <https://www.loteriasyapuestas.es/es/gordo-primitiva>

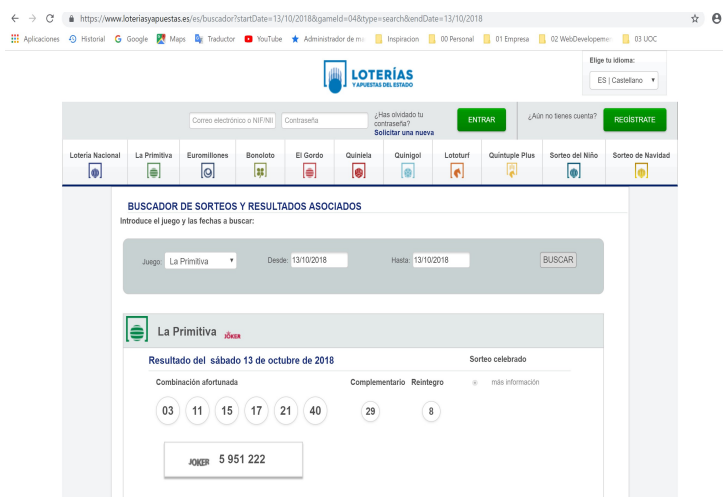
Si deseamos localizar los resultados de un sorteo en concreto, la propia web dispone de un buscador y filtros. Analizando su uso, los resultados de un sorteo dado pueden obtenerse concatenando en la url los siguientes elementos de texto:

- <https://www.loteriasypuestas.es/es/buscador?startDate=dd/mm/aaaa&gameId=ID&type=search&endDate=dd/mm/aaaa>

donde:


- para la primitiva y el gordo de la primitiva ID=04
- dd/mm/aaaa representa la fecha del sorteo.

Ejemplo: <https://www.loteriasypuestas.es/es/buscador?startDate=13/07/2018&gameId=04&type=search&endDate=13/10/2018>



9.4 Análisis de restricciones al rastreo de las webs

9.4.1 Robots.txt

#	Web	Robots.txt	Conclusiones
1	loteriasypuestasdeleestado	 robots_loteriasypuestasdeleestado.txt	Rastreo restringido a algunas carpetas

Observamos que no se permite el rastreo de urls del tipo:

- <https://www.loteriasypuestas.es/es/buscador?>

Sin embargo, analizando la navegabilidad de la web, observamos que interactuando con el navegador podríamos seleccionar la fecha de sorteo deseada en la página:

- <https://www.loteriasypuestas.es/es/la-primitiva>

The screenshot shows the 'La Primitiva' game interface on the Loterías y Apuestas del Estado website. The main content area displays the 'Sorteo 82, 13/10/2018' with the winning combination '03 11 15 17 21 40' and a complementary number '29'. The 'Reintegro' is '8'. The 'Joker' number is '5 951 222'. The 'COMPRUEBA TU SUERTE' section shows the 'Fecha de sorteo' dropdown menu highlighted with a red box. The sidebar contains instructions on how to play the game.

lo que nos refresca la pantalla con los resultados correspondientes.

The screenshot shows the 'La Primitiva' game interface on the Loterías y Apuestas del Estado website. The main content area displays the 'Sorteo 80, 06/10/2018' with the winning combination '01 11 19 26 27 34' and a complementary number '44'. The 'Reintegro' is '8'. The 'Joker' number is '1 141 111'. The 'COMPRUEBA TU SUERTE' section shows the 'Fecha de sorteo' dropdown menu set to '06/10/2018'. The sidebar contains instructions on how to play the game.

Esta última alternativa es la que utilizaremos ya que no encuentra restricciones en el fichero robots.txt.

9.4.2 Términos legales

Leídos los términos legales de uso de los contenidos web (<https://www.loteriasypuestas.es/es/paginas-informativas/politica-de-uso-de-la-web-y-sitios-moviles-webs-juego-y-corporativa.info?nomobile=1>), no se localizan restricciones respecto a la descarga automatizada de contenidos.

9.4.3 Propietario

Mediante la librería 'whois' tratamos de identificar los datos del propietario:

Universitat Oberta de Catalunya (UOC)	Alumno	César Comas Solé
Máster Universitario de Ciencia de Datos.	NIF	
Tipología y ciclo de vida de los datos		PRACTICA1

```

"LOTeriasYAPUESTAS.COM",
"loteriasyapuestas.com"
],
"registrar": "Entorno Digital, S.A.",
"whois_server": "whois.entorno.com",
"referral_url": null,
"updated_date": [
  "2018-06-28 08:44:42",
  "2018-06-28 10:44:42"
],
"creation_date": "1999-07-15 19:07:49",
"expiration_date": "2019-07-15 19:07:49",
"name_servers": [
  "ARTEMIS.TTD.NET",
  "PSINTSTL1.STL.ES",
  "artemis.ttd.net",
  "psintstl1.stl.es"
],
"status": [
  "ok https://icann.org/epp#ok",
  "ok https://www.icann.org/epp#ok"
],
"emails": "abuse@entorno.es",
"dnssec": "unsigned",
"name": null,
"org": "SOCIEDAD ESTATAL LOTERIAS Y APUESTAS DEL ESTADO",
"address": null,
"city": null,
"state": "MADRID",

```

Notamos que no se obtiene información relevante. También identificamos al propietario de la web a partir de la información legal proporcionada por la página:

Aviso legal

El titular de este sitio web y prestador de servicios de la sociedad de la información es la Sociedad Estatal Loterías y Apuestas del Estado, S.M.E., S.A. (SELAE), domiciliada en la C/ Poeta Joan Maragall 53, 28020 Madrid y su CIF es A86171964, inscrita en el Registro Mercantil de Madrid al tomo 28078, folio 202, sección 8ª, hoja M- 505970, inscripción 1ª.

Esta información legal se completa con la [política de uso](#), política de protección de datos y [política de cookies](#) de SELAE.

SELAE comercializa juegos y apuestas estatales para lo que cuenta con los correspondientes títulos habilitantes. SELAE está sometido a la supervisión del Ministerio de Hacienda y Función Pública así como del organismo regulador y supervisor en materia de juego.

Para cualquier consulta o contacto con SELAE, los usuarios podrán comunicarse a través de este correo electrónico cau@selae.es y a través de [este formulario](#) que ponemos a su disposición.

Última actualización: 29 de marzo de 2017

Copyright © Sociedad Estatal Loterías y Apuestas del Estado, S.M.E., S.A. 2017. Todos los derechos reservados.

Universitat Oberta de Catalunya (UOC)	Alumno	César Comas Solé
Máster Universitario de Ciencia de Datos.	NIF	
Tipología y ciclo de vida de los datos		PRACTICA1

9.5 Análisis de viabilidad técnica

9.5.1 Tecnología

A través la librería 'builtwith', identificamos la tecnología de la web.

```
>>> import builtwith
>>> builtwith.builtwith( ' https://www.loteriasypuestas.es ' )
{'u'javascript-frameworks': ['jQuery', 'jQuery UI'], 'u'widgets': ['u'OWL Carousel', 'u'photo-galleries': ['jQuery']}]
```

Se trata de una web dinámica desarrollada sobre jquery (javascript). En consecuencia, para interactuar con la información que proporciona deberemos interactuar desde el navegador.

9.5.2 Tamaño

