

Práctica 1: Web Scraping

Puntos para desarrollar

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

Inicialmente se plantea realizar *web scraping* sobre una plataforma conocida de cursos de formación, <https://www.skillshare.com>, por varias razones:

- Una, la página tiene una estructura amigable y, leyendo el archivo robots.txt, no deniega el uso de robots.
- Dos, desde el punto del análisis de la plataforma, por el contenido que aloja y por su modelo de negocio (acceso a todos los cursos a través de una suscripción anual), nos pareció interesante. De esta forma, pretendíamos crear una base de datos, con el contenido de la información de los principales cursos alojados en la plataforma (número de alumnos, horas a dedicar al curso, información de la persona docente, etcétera).

Sin embargo, tras la obtención de las categorías disponibles en esta web, mediante un simple scraper con Selenium, nos surgen problemas con métodos de bloqueo de robots, específicamente con reCAPTCHA.

Aunque habríamos podido seguir con la misma web, e implementar soluciones a dicho problema como el uso y rotación de Proxies y User-Agents, el límite de la ratio de acceso, también probando con headless Selenium o incluso con servicios de *captcha-solving*; en cualquier caso, hubiéramos estado rompiendo los acuerdos de licencia de Skillshare. Los cuales, prohíben explícitamente el uso de robots, spiders o similares.

Así que, en vista de esto, decidimos elegir otra página web.

La inquietud por obtener información sobre las ofertas de empleo público de distintas administraciones públicas nos movió a elegir como ejemplo la de la Junta de Andalucía, donde el Instituto Andaluz de Administración Pública (IAAP) muestra la situación actual de la Oferta de Empleo Pública (OEP), que puede encontrarse en esta web. La dificultad de acceso a los datos de forma intuitiva, y la necesidad de consultarlos de forma frecuente hace que sea candidata para realizar *web scraping* y poder contar con esos datos, por ejemplo, para preparadores particulares, academias de formación y los propios interesados.

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

El título elegido ha sido: **Convocatorias Públicas de la Junta de Andalucía**, y el nombre del archivo csv: *JAconvocatoriasPublicas.csv*

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.

Tal y como se comenta en la primera pregunta, el tema de la formación nos ha movido desde primera hora. Decantarnos por la oferta pública de empleo de una administración pública, las

convocatorias y los resultados de sus pruebas selectivas nos pareció interesante. El dataset extraído, en formato csv (*JAconvocatoriasPublicas.csv*), ofrece información sobre:

- Convocatorias con proceso ya cerrado
- Convocatorias en proceso, indicando en qué fase se encuentran
- Convocatorias futuras

Si bien la página web establece la posibilidad de filtrar resultados, nuestro dataset ofrece la totalidad de la oferta, ordenada desde la fecha más reciente a la más antigua. La cabecera de nuestro dataset incluye más columnas que las que ofrece la tabulación de datos en la web. Datos como el nombre de la convocatoria, la fecha de publicación de la convocatoria, la url para la consulta en el boletín oficial de la Junta de Andalucía, si la convocatoria ha sufrido modificaciones, la url de acceso a esa convocatoria en concreto y la fecha de extracción de los datos, que permite dar validez (o no) a lo consultado.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

A continuación, se muestra el diagrama de flujo del proyecto:

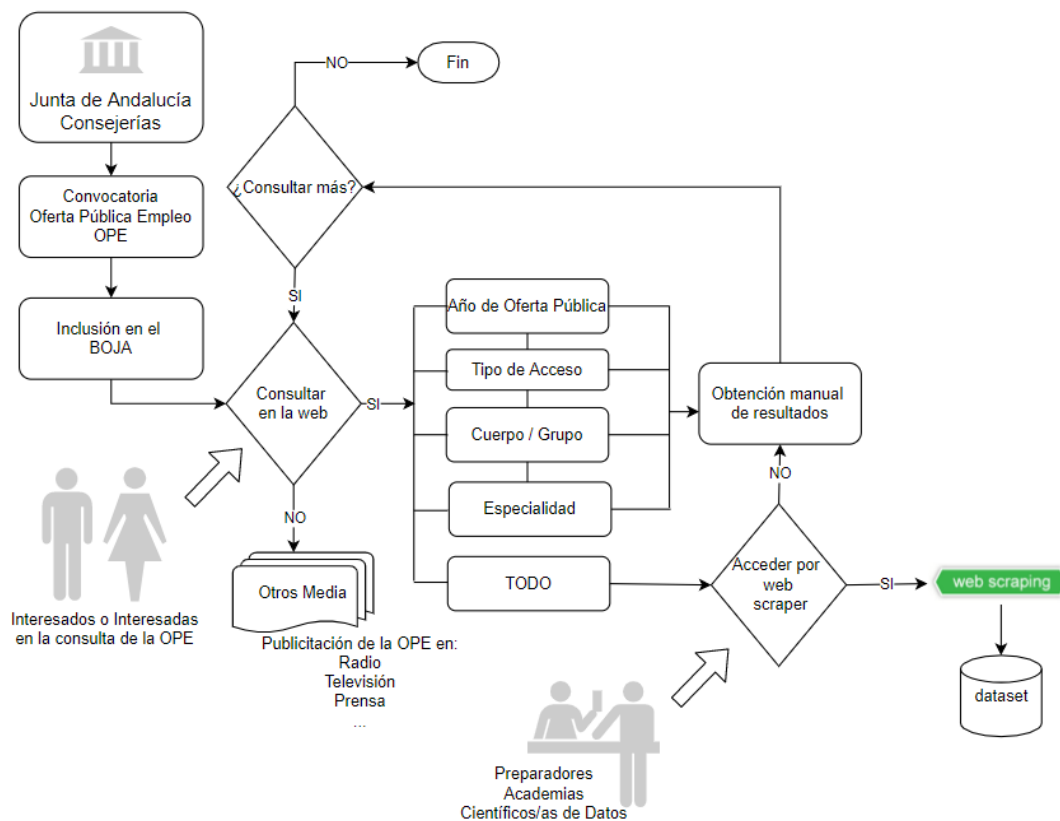


Ilustración 1. Diagrama de flujo general

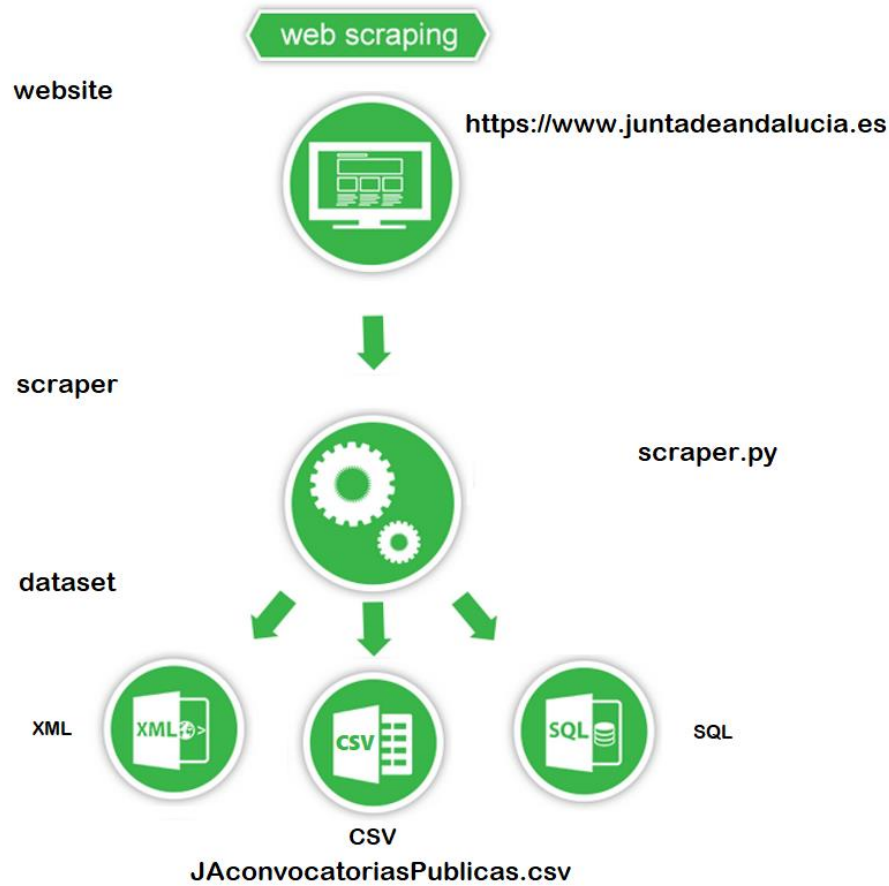


Ilustración 2. Proceso de web scraping de esta práctica

También, la forma de proceder en el momento de extraer y elaborar el dataset:

Esquema de la obtención del dataset: Convocatorias Públicas de la Junta de Andalucía

- Nombre de la convocatoria
- Cuerpo/Especialidad o Categoría
- Año de Oferta Pública
- Tipo de Acceso
- Número de plazas
- Fin de plazo
- Publicación de la convocatoria
- Url BOJA
- ¿Ha habido modificaciones en la convocatoria? (Yes/No)
- Estado
- Url de acceso (El vínculo que nos da acceso directo a la convocatoria)
- Fecha de extracción (Fecha cuando los datos fueron extraídos)

Ilustración 3. Identificación gráfica de los datos del dataset

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

Tal y como se indicó en la pregunta 3, el dataset obtenido para esta práctica contiene más información de la que se muestra en la página web de la oferta pública de empleo de la Junta de Andalucía. Nuestra estructura es la que se muestra en la siguiente tabla.

Campo	Tipo	Ejemplo
Año de la oferta pública	Numérico	2019
Nombre de la convocatoria	Texto	A1.1100 Administradoras/Administradores Generales (Estabilización)
Tipo de acceso	Texto	Personal funcionario - Acceso libre
Cuerpo/Especialidad o Categoría	Texto	A1.1 Cuerpo Superior de Administradoras y Administradores
Número de plazas	Numérico	98
Fin de plazo	Data	21/11/2019
Publicación de la convocatoria (BOJA)	Texto	BOJA nº 205 de 23/10/2019
Url Boja	Texto	https://www.juntadeandalucia.es/institutodeadministracionpublica/servlet/descarga?up=134883
¿Ha habido modificaciones en la convocatoria?	Texto	Yes
Estado de la convocatoria	Texto	Convocatoria 3º ejercicio
Url de acceso	Texto	https://www.juntadeandalucia.es/institutodeadministracionpublica/publico/seleccionjunta.filter?step=read&cp=1&id=1&chm=-1&ca=-1&cu=15&v=1&cdp=-1&ch=50&cd=221883
Fecha de extracción	Data	05/11/2021

Se excluye del dataset los campos “url modificación” y “url modificación 2” por no estar rellenos en ninguna de las convocatorias desde el año 2009.

El período de tiempo de los datos obtenidos comprende desde las convocatorias realizadas en el año 2009 hasta las últimas, que se corresponden con el año 2019, y de las que aún hay un número elevado que aún no han concluido, y que han sufrido la unificación con el proceso selectivo del año anterior (2018).

La web de consulta muestra el siguiente aspecto:

Función Pública J.A.

A través del siguiente formulario obtendrá información sobre las distintas convocatorias publicadas y el estado en que se encuentra cada una.

Año de Oferta Pública (*)

Tipo de Acceso (*)

Cuerpo/Grupo

Especialidad

Año de Oferta Pública	Tipo de Acceso	Cuerpo/Especialidad ó Categoría Profesional	URL Modificación	URL Modificación 2	Plazas	Plazo de solicitud	Estado
-----------------------	----------------	---	------------------	--------------------	--------	--------------------	--------

Ilustración 4. Detalle de la cabecera de datos en la web

La imagen de nuestro dataset, una vez formateado y llevado a una hoja de cálculo Excel, es el siguiente

	A	B	C	D	E	F	G	H	I	J	K	L
	Año de Oferta Pública	Nombre de la Convocatoria	Tipo de Acceso	Cuerpo/Especialidad ó Categoría	Número de Plazas	Fin de plazo	Publicación de la convocatoria	Url BOJA	¿Ha habido modificaciones en la convocatoria?	Estado	Url de acceso	Fecha de extracción
1		A1.1100	Personal	A1.1 Cuerpo			BOJA nº 205 de	https://www.juntadeandalucia.es	Si	Convocatoria 3ª	https://www.juntadeandalucia.es	06-11-21
2	2019	Administradoras/Administradores	Funcionario -	Superior de	98	21-11-19	23/10/2019	https://www.juntadeandalucia.es	Si	Convocatoria 3ª	https://www.juntadeandalucia.es	06-11-21
3	2019	Administradoras/Administradores	Funcionario -	Superior de	12	21-11-19	23/10/2019	https://www.juntadeandalucia.es/boja	Si	Convocatoria 3ª	https://www.juntadeandalucia.es	06-11-21

Ilustración 5. Detalle de la cabecera del dataset obtenido, en formato Excel

La forma en la que han sido recopilados los datos ha sido la siguiente:

1. Obtención de la url base sobre la que trabajar. La web cuenta con distintas páginas para las convocatorias, esta url sería la base que nos permitiría acceder a cada una de las páginas dónde se encuentran las convocatorias.
2. Obtención de la totalidad de los links de las convocatorias. Desde la url base, solamente cambiando su número final. Así, por ejemplo, url base + 10, nos permitiría acceder a la página 10.
3. Se determina la ruta *XPATH**, la cual contiene el bloque del link, y de este bloque se extrae atributo *"href"* que contiene el link. También, se gestionan esperas explícitas (con el elemento *"contenido"*) y posibles errores.
4. Dado el gran número de urls de convocatorias, para una mejor gestión, se guardan en un archivo *pickle*, para evitar iteraciones innecesarias.
5. A partir del archivo *pickle* que guarda una línea por convocatoria, iteramos, y con *BeautifulSoup* extraemos el bloque de contenido definido por la *class "ficha"* (que llamaremos *whole_section*).
6. Ahora ya podemos obtener los siguientes datos de interés:
 - a. Nombre de la convocatoria: *whole_section.h3*
 - b. Para las siguientes variables, definimos una función que busque específicamente los elementos bajo su nombre y el elemento *"p"*.
 - i. Año de Oferta Pública.
 - ii. Cuerpo.
 - iii. Tipo de Acceso.
 - iv. Número de plazas.
 - v. Finaliza plazo de solicitud.
 - vi. Publicación de la convocatoria.
 - vii. Modificación de la convocatoria.
 - viii. Estado.
7. Se añade la información de cada fila leída a un diccionario.
8. Se guarda en el fichero *JAconvocatoriasPublicas.csv* todo el dataframe obtenido.

```
driver = webdriver.Chrome('./chromedriver/chromedriver', options=option)
47%|██████████| 9/19 [01:33<01:42, 10.29s/it]
```

```
driver = webdriver.Chrome('./chromedriver/chromedriver', options=option)
100%|██████████| 19/19 [03:12<00:00, 10.14s/it]
progress bar: 4%|██████| 17/429 [00:16<06:38, 1.03it/s]
```

```
driver = webdriver.Chrome('./chromedriver/chromedriver', options=option)
100%|██████████| 19/19 [03:12<00:00, 10.14s/it]
progress bar: 45%|███████| 193/429 [03:13<03:38, 1.08it/s]
```

```
driver = webdriver.Chrome('./chromedriver/chromedriver', options=option)
100%|██████████| 19/19 [03:12<00:00, 10.14s/it]
progress bar: 100%|██████████| 429/429 [07:13<00:00, 1.01s/it]
```

Ilustración 6. Relación de imágenes que muestran el progreso del Web Scraping.

Índice	Año de Oferta Pública	Nombre de la Convocatoria	Tipo de Acceso	Superficie Construida o Categoría Profesional	Número de Plazas	Fin de plazo	Publicación de la convocatoria	URL BOLA	Ha habido modificaciones en la convocatoria	Estado	URL de acceso	Fecha de actualización
0	2019	A1.1100 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	98	21/11/2019	BOLA nº 205 de 23/10/2019	https://www.bola.es	Yes	Conv.	https://www.bola.es	2021-11-06
1	2019	A1.1200 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	12	21/11/2019	BOLA nº 205 de 23/10/2019	https://www.bola.es	Yes	Conv.	https://www.bola.es	2021-11-06
2	2018	A1.1100 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	288	nan	BOLA nº 84 de 6 de Mayo de 2019	https://www.bola.es	No	Ofer.	https://www.bola.es	2021-11-06
3	2018	A1.1200 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	104	nan	BOLA nº 84 de 6 de Mayo de 2019	https://www.bola.es	No	Proc.	https://www.bola.es	2021-11-06
4	2018	A1.1100 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	105	21/11/2019	BOLA nº 205 de 23/10/2019	https://www.bola.es	Yes	Conv.	https://www.bola.es	2021-11-06
5	2018	A1.1200 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	85	21/11/2019	BOLA nº 205 de 23/10/2019	https://www.bola.es	Yes	Conv.	https://www.bola.es	2021-11-06
6	2017	A1.1100 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	nan	nan	nan	nan	No	Proc.	https://www.bola.es	2021-11-06
7	2017	A1.1200 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	nan	nan	nan	nan	No	Proc.	https://www.bola.es	2021-11-06
8	2017	A1.1100 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	40	nan	nan	nan	No	Proc.	https://www.bola.es	2021-11-06
9	2017	A1.1200 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	46	nan	nan	nan	No	Proc.	https://www.bola.es	2021-11-06
10	2017	A1.1200 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	3	nan	nan	nan	No	Proc.	https://www.bola.es	2021-11-06
11	2017	A1.1200 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	40	nan	nan	nan	No	Proc.	https://www.bola.es	2021-11-06
12	2016	A1.1100 Administrador	Personal Fun.	A1.1 Cuerpo Superior de Administradoras y Administradores	168	21/10/2016	BOLA n.º 180, de 19 de septiembre de 2016	http://www.bola.es	No	Proc.	https://www.bola.es	2021-11-06

Ilustración 7. Detalle del dataset obtenido

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Agradecemos al Instituto Andaluz de Administración Pública (IAAP), por el alojamiento y cesión de los datos. El Instituto Andaluz es una agencia administrativa, creada por la [Ley 6/85 de 28 de noviembre de Ordenación de la Función Pública de la Junta de Andalucía](#), y tanto el instituto, al pertenecer a un organismo público, como sus estatutos se encuentran regulados por el [Decreto 277/ 2009 de 16 de junio](#).

La búsqueda y gestión de Ofertas de Empleo Público es una temática de interés para las entidades que tienen su nicho de mercado en las personas que salen al mercado laboral, o que quieren mejorar sus condiciones laborales acercándose a la administración. Páginas como [savia.net](https://www.savia.net/) (<https://www.savia.net/>), que ha creado la herramienta *convoca* (<https://convoca.online/funcionalidad/>) donde se da información desde la publicación de la convocatoria hasta el cierre del proceso selectivo, y se encamina tanto a candidatos como a gestores, es un ejemplo de un análisis similar.

El proyecto no va dirigido contra propiedades intelectuales ajenas o marcas registradas, no viola los derechos de autor, no obtiene los datos con fines comerciales y por tanto no realiza competencia desleal. Cumple fielmente tanto la *Ley Orgánica de Protección de Datos de Carácter Personal* (LOPD), como el reglamento *Europeo de General Data Protection Regulation* (RGPD). Tampoco en ningún momento sobrecarga los servicios del sitio objetivo.

Por tanto, en vistas de los puntos expuestos, podemos determinar que el proyecto cumple los principios éticos y legales.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Dado que en marzo de 2020 se declaró el confinamiento por la pandemia de la COVID 19, el retraso en todos los procesos administrativos afecta por supuesto a las convocatorias de empleo público. Este confinamiento ha cambiado hábitos en las actividades de formación, tomando un gran protagonismo las plataformas on-line y la comunicación entre academias y estudiantes por medios telemáticos.

Pero el interés de este conjunto de datos va más allá de todo lo provocado por el confinamiento. Realizar scraping en la web de esta administración permite tener una base de datos actualizada de las distintas convocatorias para dejar claro al estudiante los plazos a cumplir, saber cuáles están en marcha, así como las futuras. Las convocatorias pasadas pueden servir como base para hacer previsiones de cuáles estarán próximas a salir, así como el número de plazas que suelen ser ofertadas.

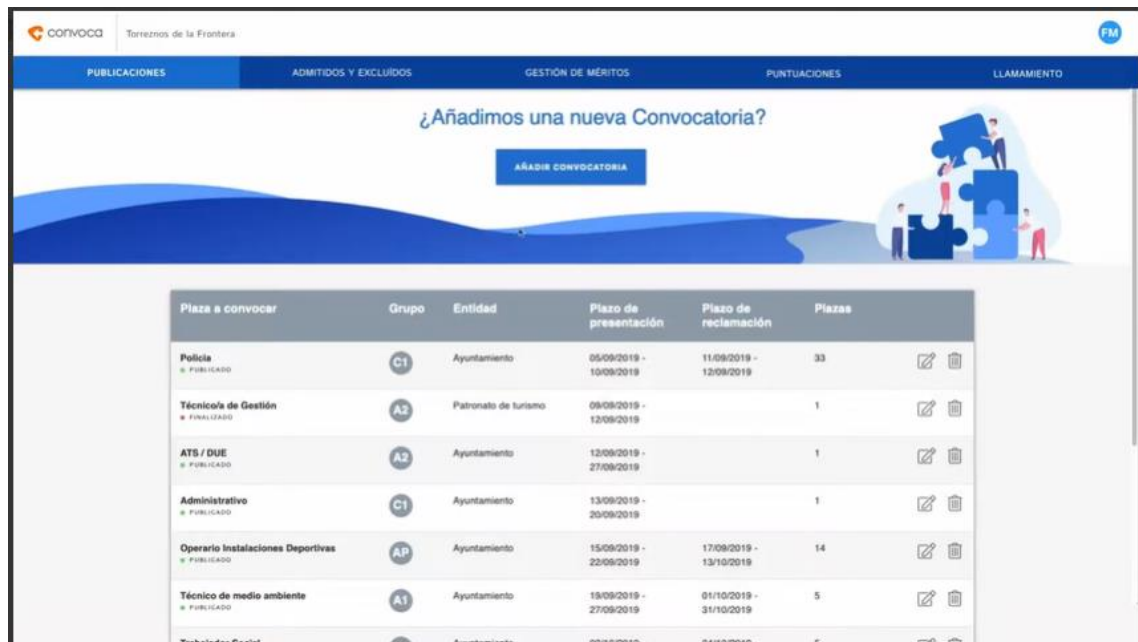
En el caso de *savia.net*, y su herramienta *convoca*, podemos ver cómo esta herramienta de pago gestiona las convocatorias correspondientes a una Oferta de Empleo Pública, con datos como los obtenidos en nuestro web scraping: bases, número de plazas, condiciones...

Esta herramienta comercial está ideada para centralizar, digitalizar y automatizar todo el ciclo de selección de empleados públicos, y en su primera fase, está la introducción de los datos de la convocatoria o convocatorias.¹



Ilustración 8. Entrada a la herramienta *convoca*, de *savia.net*

¹ El Mundo. (2019, 19 noviembre). *Savia lanza el primer portal de empleo para ayuntamientos*. ELMUNDO. Recuperado 7 de noviembre de 2021, de <https://www.elmundo.es/tecnologia/innovacion/2019/11/19/5dd3cafffc6c83a0348b4752.html>



The screenshot shows the CONVOCO website interface. At the top, there's a navigation bar with links: PUBLICACIONES, ADMITIDOS Y EXCLUIDOS, GESTIÓN DE MÉRITOS, PUNTUACIONES, and LLAMAMIENTO. Below this, a banner asks '¿Añadimos una nueva Convocatoria?' with a button 'AÑADIR CONVOCATORIA'. The main content is a table of job vacancies.















Plaza a convocar	Grupo	Entidad	Plazo de presentación	Plazo de reclamación	Plazas	
Policia PUBLICADO	C1	Ayuntamiento	05/09/2019 - 10/09/2019	11/09/2019 - 12/09/2019	33	 
Técnico/a de Gestión PUBLICADO	A2	Patronato de turismo	09/09/2019 - 12/09/2019		1	 
ATS/ DUE PUBLICADO	A2	Ayuntamiento	12/09/2019 - 27/09/2019		1	 
Administrativo PUBLICADO	C1	Ayuntamiento	13/09/2019 - 20/09/2019		1	 
Operario Instalaciones Deportivas PUBLICADO	AP	Ayuntamiento	15/09/2019 - 22/09/2019	17/09/2019 - 31/10/2019	14	 
Técnico de medio ambiente PUBLICADO	A1	Ayuntamiento	19/09/2019 - 27/09/2019	01/10/2019 - 31/10/2019	5	 
Trabajador/a Social		Ayuntamiento	20/10/2019 -	24/10/2019 -	6	 

Ilustración 9. Detalle de las ofertas que convoca tiene en su base de datos

8. Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Other (specified above).
- Unknown License.

Se selecciona **CC BY-NC-SA 4.0**. Licenciando el proyecto bajo esta licencia, se permite que otros usuarios sean libres de usar, cambiar y distribuir el software, pero tiene unas particularidades que creemos que son las más adecuadas para nuestro caso:

- Los usuarios son libres de compartir, copiar y redistribuir el material. También lo pueden adaptar, transformar y construir sobre este.
- No pueden hacer uso comercial del material licenciado.
- Han de otorgar crédito correspondiente, proporcionando un enlace e indicar si se han realizado cambios.

Siendo un proyecto de carácter académico, realizado sobre un conjunto de datos público, no consideramos apropiado que se haga un uso comercial de este. Además, la distribución de estas obras derivadas se debe hacer con una licencia igual a la que regula la obra original, con lo cual, nos aseguramos de que se nos dé el crédito por el conjunto de datos obtenido.

Dicha licencia se puede leer detenidamente en el repositorio que contiene el proyecto realizado.

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

El repositorio de Github se encuentra en

[UOCPgarcia/IAAP-OEP-WebScraping: Práctica WebScraping UOC M2.851 \(github.com\)](https://github.com/UOCPgarcia/IAAP-OEP-WebScraping)

El código con el que se ha generado el dataset se encuentra en la carpeta `./code/scrapper.py` del repositorio Github.

La estructura completa elegida para el repositorio es la que sigue

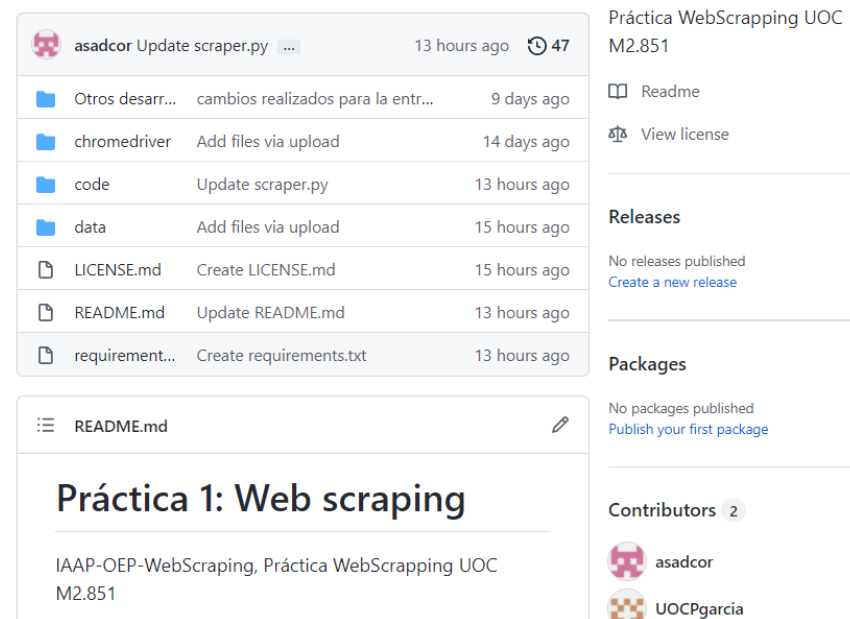


Ilustración 10. Detalle del repositorio

- Fichero *LICENSE.md*: bajo licencia CC BY-NC-SA 4.0
- Fichero *README.md*: información completa del repositorio, estructura y enlace a Zenodo.
- Fichero *requirements.txt*: contiene las librerías instaladas durante la ejecución del proyecto, y la versión bajo las cuales el web scraper se ha ejecutado correctamente, logrando así la generación exitosa el dataset.
- Fichero *./data/practica1_tipologia.pdf* memoria del proyecto, con las respuestas a las preguntas planteadas.
- Fichero *./data/IAconvocatoriasPublicas.csv* dataset obtenido.
- Fichero *./data/urls_convocatorias.pickle* archivo pickle que aloja los links de cada convocatoria.
- Fichero *./code/scrapper.py* archivo Python que contiene el código del web scraper.
- Fichero *./code/robots.py* archivo Python que contiene el código para el análisis de robots y propietario.
- Fichero *./chromedriver/chromedriver.exe* contiene el **driver** de la versión de Chrome específica, necesaria para Selenium.
- Carpeta *./otros desarrollos/* contiene proyectos descartados, pero que mantenían cierta funcionalidad.

10. Dataset. Publicar el dataset obtenido (*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

El DOI correspondiente a la base de datos es:
<https://doi.org/10.5281/zenodo.5651108>

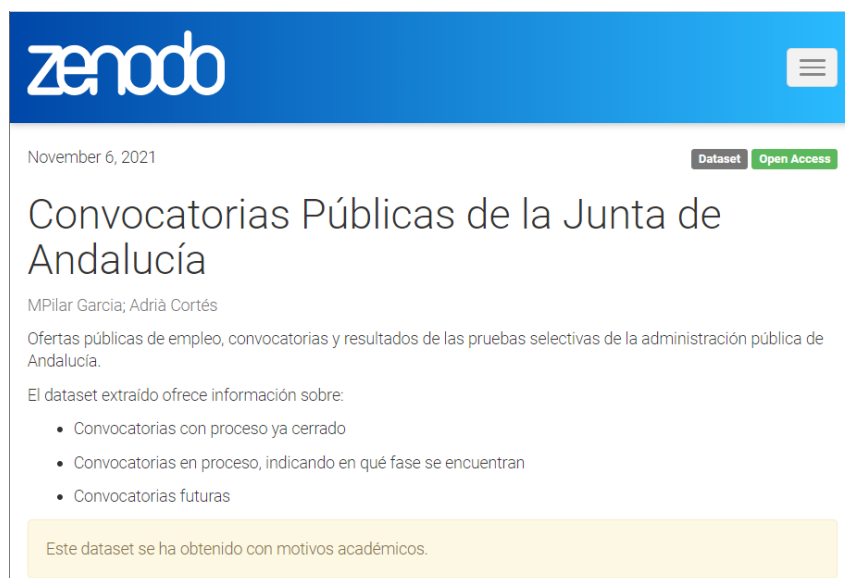


Ilustración 11. Detalle de la publicación del dataset obtenido

Tabla de contribuciones al trabajo

Contribuciones	Firma
Investigación previa	M.P.G.R, A.C.A
Redacción de las respuestas	M.P.G.R, A.C.A
Desarrollo código	M.P.G.R, A.C.A