

Web Scraping: Recuperando datos

Daniel López Gala – Hack&Beers Gijón '23



¿Qué es el Web Scraping?

- Screen Scraping
- Data Mining
- Web Harvesting
- Web Crawlers

...



01.

Web Scraping

Qué es el Web Scraping
y cómo funciona.

02.

Legalidad

¿Es legal hacer
scraping? Pista: Sí

03.

Threat Intelligence

Casos de uso de
scraping en inteligencia.



04.

BDNS

Scrapeando datos
públicos.

05.

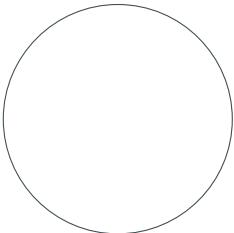
Uniovi

Scrapeando datos no
tan públicos.

06.

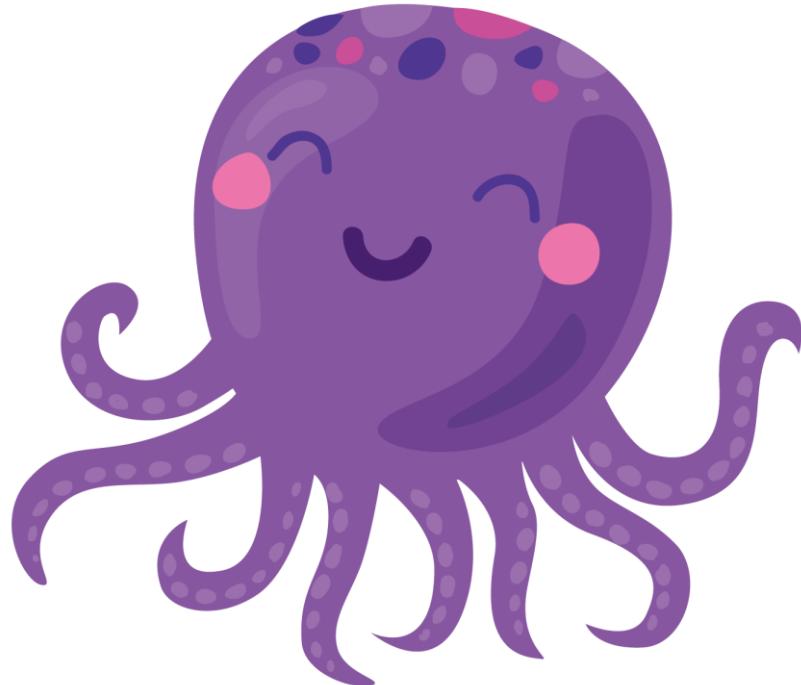
Conclusiones

Conclusiones y recursos
recomendados.



Proceso de Scraping

1. Bot de extracción, envía una **solicitud HTTP GET** a un sitio web específico.
2. Cuando el sitio web responde, el scraper analiza el documento HTML para buscar un **patrón de datos** específico.
3. Una vez se hayan extraído los datos, se convierten a cualquier **formato específico** proyectado por el autor del bot de scraper.



Proceso de Scraping



WAFs

Evasión de Web Application Firewalls



IPS/IDS

Evasión de sistemas de detección y prevención de intrusiones



Captcha / OCR

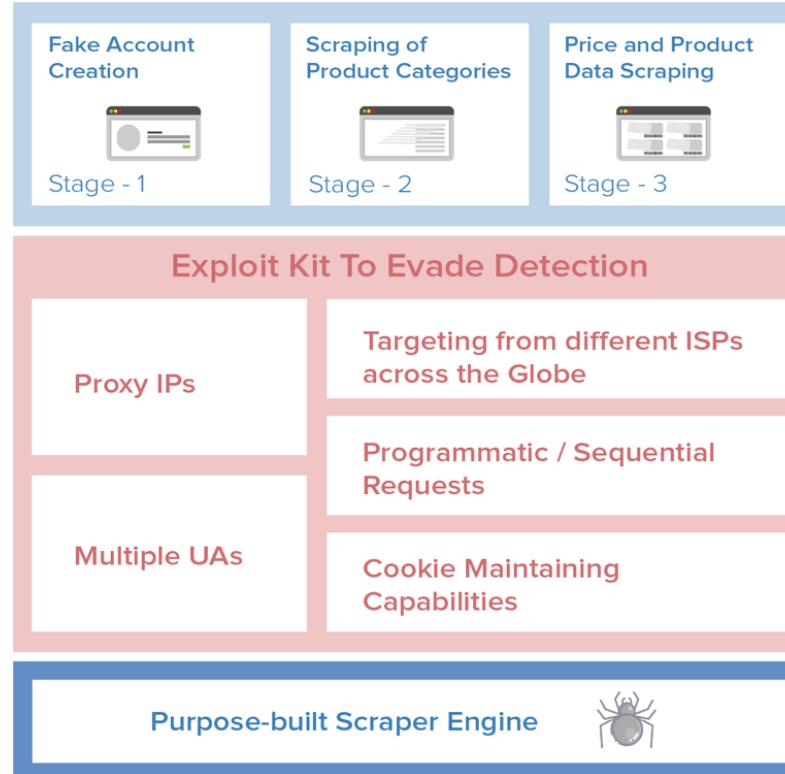
Esquivar mecanismos de detección de robots



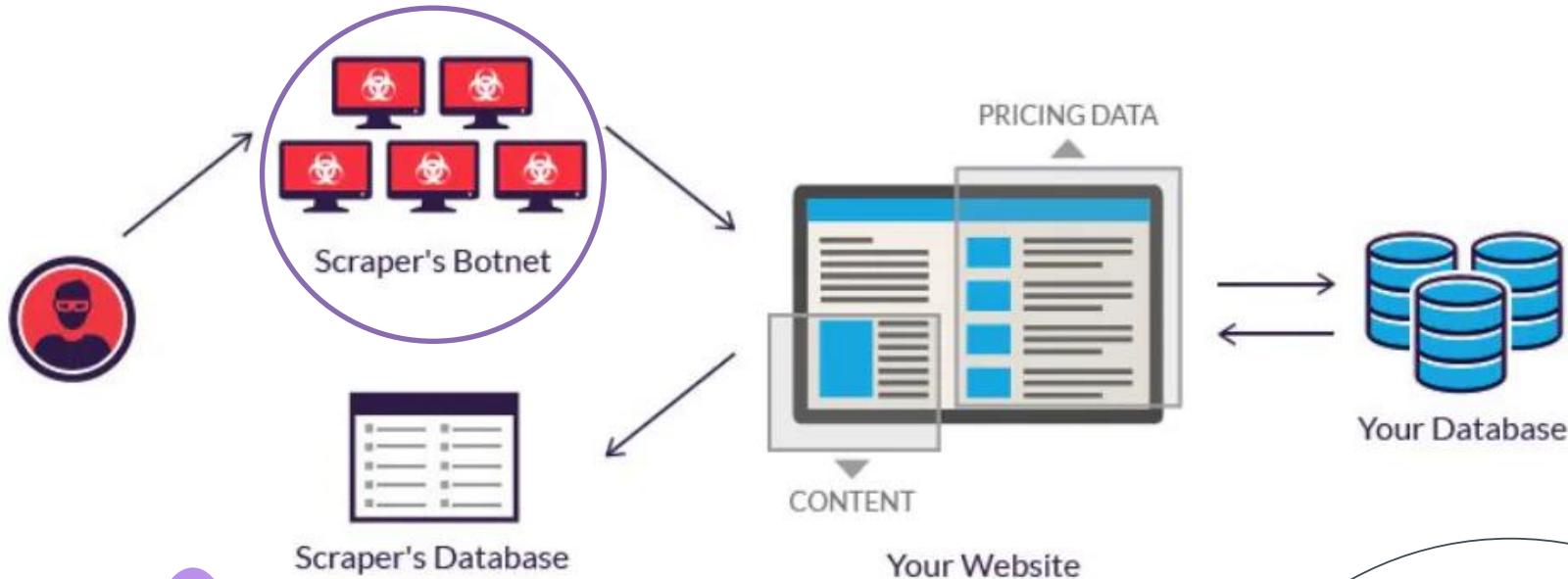
Proxies

Pasar máximos de solicitudes utilizando proxies y user-agents aleatorios

Proceso de Scraping



Proceso de Scraping



Legalidad

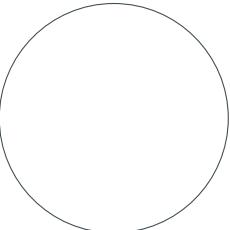
Son fuentes de acceso público o los datos se recaban por un fin de interés público general.

Prevalece el interés del responsable del tratamiento sobre el derecho a la protección de datos.

La persona rastreada lo es bajo su consentimiento.



Like it or not – Scraping is *practically legal*



.....

¿Data Leaks?

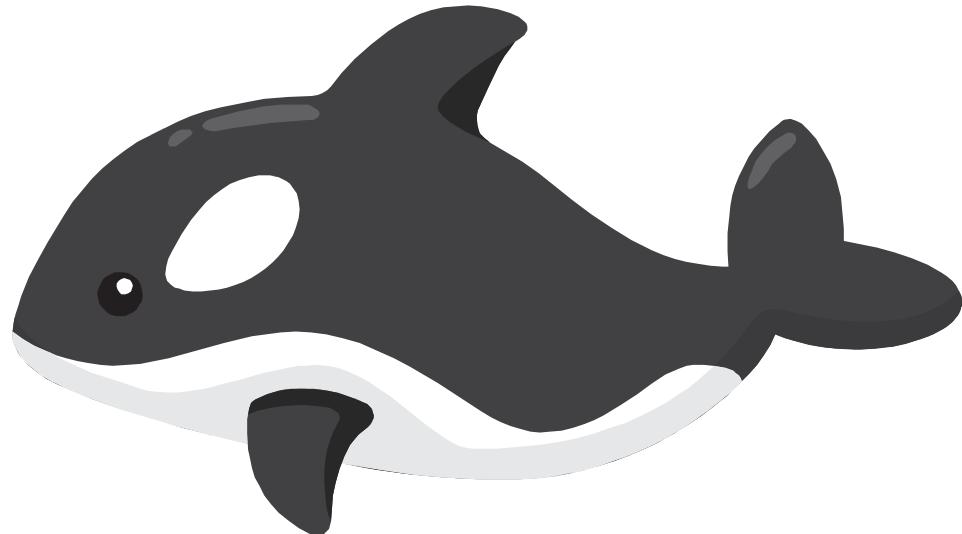
.....

Se supone que la información es pública.

Por tanto... ¿No hay peligro, no?



Bueno...



Home > News > Security > Meta fined €265M for not protecting Facebook users' data from scrapers

Meta fined €265M for not protecting Facebook users' data from scrapers

By [Bill Toulias](#)

November 28, 2022 10:15 AM 0

533.313.128

Un scraper vendió información de casi 540M de usuarios incluyendo números de teléfono, direcciones de correo electrónico, y datos personales.



1.3 million Database with Clubhouse Users Available to the Public! It's not a Leak, it's Data Scraping

Date: 22 Apr 2021

Categories: Data Security

1.300.000

Clubhouse 1.3M SQL

userid|name|photo_url|username|twitter|Instagram|num_followers|num_following|time_created|invitedbyuser_profile

Download

<https://www.ka> [REDACTED]



Scraping: "Robé los datos de 700 millones de usuarios de LinkedIn por diversión"

700.000.000

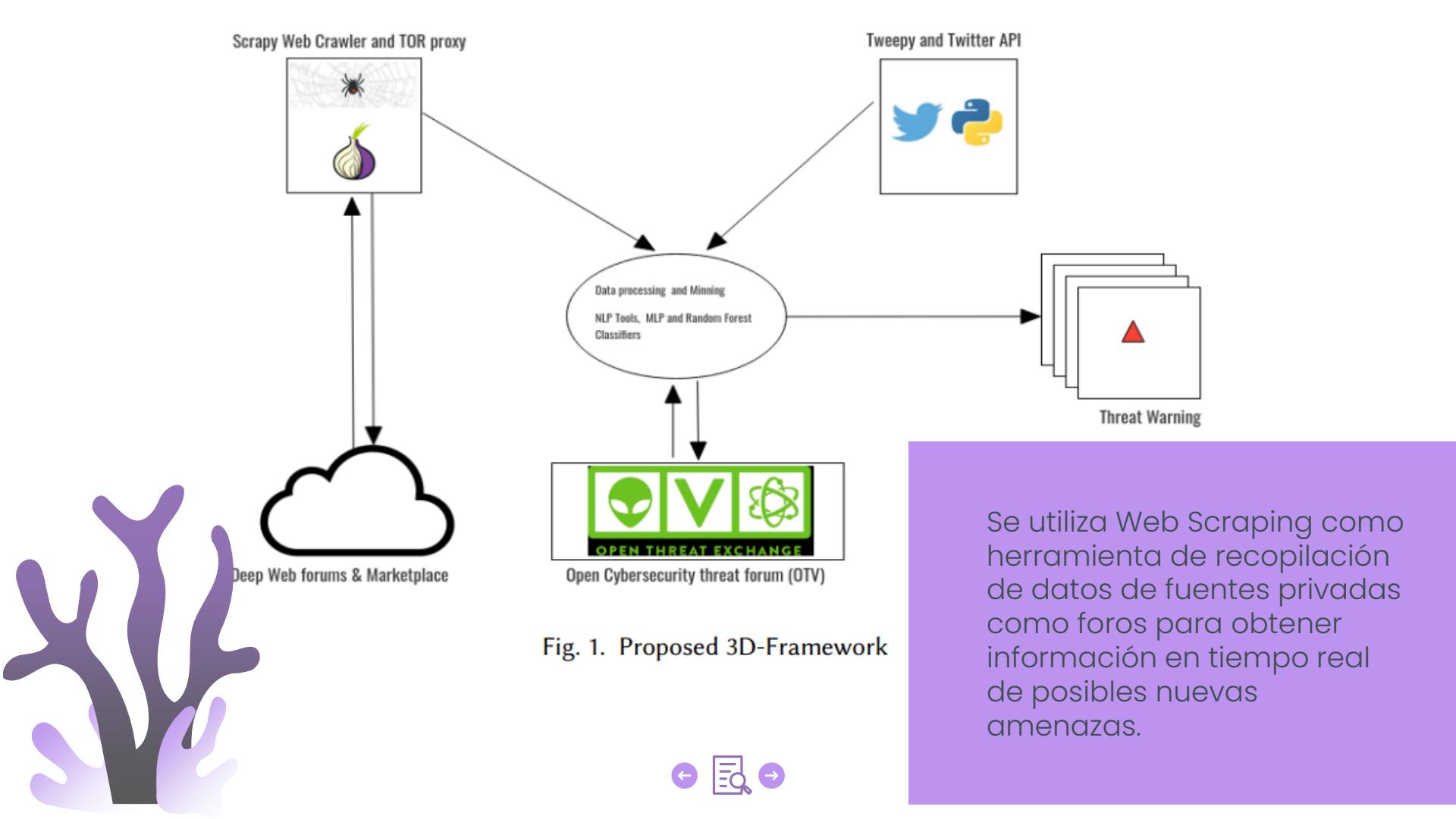
LinkedIn: 'It's Not a Breach if It's Public Info'

Threat Intelligence

El web scraping es un componente crítico en Threat Intelligence.

La Deep Web contiene valiosos datos, y la inteligencia de amenazas es fundamental para prevenir ataques, y el web scraping lo respalda al recopilar información de alto nivel.





AutoUniCalendar



Universidad de
Oviedo

Idiomas ▾ Mi Perfil ▾

Servicios académicos



Universidad de
Oviedo

Grado, Máster y Doctorado

+

Otra formación

+

Otros servicios

+

Prácticas en empresa

+

Ayuda

+

Mi Horario

Grado en Ingeniería Informática en Tecnologías de la Información

oo Hoy

Nov 28 – Dic 4, 2022

	Lu 11/28	Ma 11/29	Mi 11/30	Ju 12/1	Vi 12/2
Todo el día					
8:00					
9:00	9:00 - 12:00 Administración de Sistemas - Prácticas De Laboratorio-05	9:00 - 10:00 Programación Concurrente y Paralela - Teoría-a	9:00 - 11:00 Redes de Computadores - Tutorías Grupales-05	9:00 - 10:00 Configuración y Evaluación de Sistemas - Prácticas De Aula-02	10:00 - 11:00 Administración de Sistemas - Prácticas De Aula-02
10:00		10:00 - 12:00 Redes de Computadores - Teoría-a		11:00 - 12:00 Programación Concurrente y Paralela - Prácticas De Aula-	11:00 - 12:00 Programación Concurrente y Paralela - Prácticas De Aula-
11:00					
12:00		12:00 - 1:00 Administración de Sistemas - Teoría-a	12:00 - 2:00 Configuración y Evaluación de Sistemas - Prácticas De Laboratorio-05		12:00 - 3:00 Programación Concurrente y Paralela - Prácticas De Laboratorio-05
13:00			13:00 - 2:00 Configuración y Evaluación de Sistemas - Teoría-a		
14:00	2:00 - 3:00 Redes de Computadores - Prácticas De Aula-02				
15:00		3:00 - 4:00 Ampliación de Matemáticas - Prácticas De Laboratorio-2	3:00 - 4:00 Aprendizaje Automático I - Prácticas De Laboratorio-2		
16:00	4:00 - 5:00 Ampliación de Matemáticas - Clases Expositivas-1	4:00 - 5:00 Aprendizaje Automático I - Clases Expositivas-1			
17:00	5:00 - 6:00 Ampliación de Matemáticas - Prácticas De Aula/seminaria				
18:00					

AutoUniCalendar

autoUniCalendar: Descarga tu calendario de Uniovi.

Cookies

Introduce tus cookies de sesión.

JSession

RenderToken

¿Necesitas ayuda?

DESCARGAR

Descargar usando credenciales.

Si lo prefieres, puedes usar tus credenciales en lugar de tus cookies de sesión para descargar el calendario. Es más sencillo, ¡pero recomendamos usar las cookies!

USAR CREDENCIALES



AutoUniCalendar



Universidad de Oviedo

sies.uniovi.es/serviciosacademicos/web/expedientes/calendario.xhtml

Universidad de Oviedo

Mi Horario

Grado en Ingeniería Informática en Tecnologías de la Información

Para ver la localización de las aulas asignadas pinche AQUÍ

Hoy

Sep 6 – 12, 2021

Mes Semana Día

Application

- Manifest
- Service Workers
- Storage

Storage

- Local Storage
- Session Storage
- IndexedDB
- Web SQL
- Cookies
 - https://sies.uniovi.es
- Trust Tokens

Cache

- Cache Storage

Name	Value
cookieconsent_status	dismiss
oam.Flash.RENDERMAP.TOKEN	-kwdkj3qcw
JSESSIONID	0000Du-yFFFda32fxhCryqE6YZr:1d18euudr

autoUniCalendar.py

```
# Function to send the first GET HTTP request using the tokens provided.
def get_first_request(session_token, render_token):

    print("[@] Sending the first request...")

    # Cookies payload of the HTTP request.
    payload = {
        'JSESSIONID': session_token,
        'oam.Flash.RENDERMAP.TOKEN': render_token
    }

    r = requests.get(url, cookies=payload)
    print("[#] First request correctly finished.\n")
    # The function returns the server response to use it later.
    return r.text
```



autoUniCalendar.py

```
# Function to extract the cookies necessary to make the POST request, from the server response of the
# first request.
def extract_cookies(get_response):
    print("[@] Extracting the calendar parameters...")

    # Iterate the response lines to search the cookies, and save them in variables.

    found_first, found_second, found_third = False, False, False
    for line in get_response.split('\n'):
        if '<div id="j_id' in line and not found_first:
            source = urllib.parse.quote(re.findall('"(.*?)"', line.split('<')[1])[0])
            found_first = True

        if 'javax.faces.ViewState' in line and not found_second:
            viewstate = urllib.parse.quote(re.findall('"(.*?)"', line.split(' ') [12])[0])
            found_second = True

        if 'action="/serviciosacademicos/web/expedientes/calendario.xhtml"' in line and not found_third:
            submit = re.findall('"(.*?)"', line.split(' ') [3])[0]
            found_third = True

    print("[#] Calendar parameters extracted.\n")
    # The function returns a list that contains the extracted parameters.
    return [source, viewstate, submit]
```



autoUniCalendar.py

```
# Function that sends the HTTP POST request to the server and retrieves the raw data of the calendar.
def post_second_request(session_token, render_token, ajax, source, view, start, end, submit):

    print("[@] Sending the calendar request...")

    # Cookies of the request.
    payload = {
        'JSESSIONID': session_token,
        'oam.Flash.RENDERMAP.TOKEN': render_token,
        'cookieconsent_status': 'dismiss'
    }

    # Define variables of the request.
    string_start = source + "_start"
    string_end = source + "_end"
    string_submit = submit + "_SUBMIT"

    # Creating the body with the parameters extracted before, with the syntax required by the server.
    print("[*] Creating the payload...")
    body_payload = f"javax.faces.partial.ajax={ajax}&javax.faces.source={source}&javax.faces.partial.execute={source}&javax.faces.partial.render={source}&{source}={source}&{string_start}={start}&{string_end}={end}&{string_submit}=1&javax.faces.ViewState={view}"

    # Send the POST request.
    r = requests.post(url, data=body_payload, headers={'Content-Type': 'application/x-www-form-urlencoded; charset=UTF-8'}, cookies=payload)
    print("[#] Calendar request correctly retrieved.\n")

    # Write the raw response into a temporary file.
    print("[@] Writing the raw calendar data into a .txt file...")
    f = open("raw.txt", "w")
    f.write(r.text)
    f.close()
    print("[#] File correctly written.\n")
```



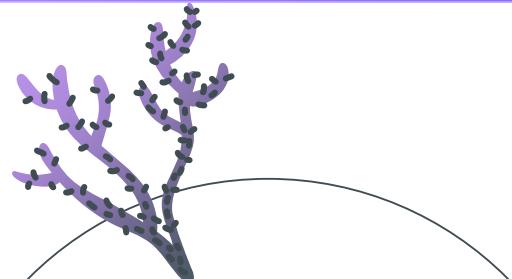
```
title_csv = re.findall('"(^[^"]*)"', title.split(':')[1])[0]
start_date = start.split(' ')[1].split('T')[0].split('"')[1]
start_date_csv = start_date.split('-')[2]+'/'+start_date.split('-')[1]+'/'+start_date.split('-')[0]
start_hour = start.split(' ')[1].split('T')[1].split('+')[0]
end_date = end.split(' ')[1].split('T')[0].split('"')[1]
end_date_csv = end_date.split('-')[2]+'/'+end_date.split('-')[1]+'/'+end_date.split('-')[0]
end_hour = end.split(' ')[1].split('T')[1].split('+')[0]
alert_date = start_date_csv
alert_hour = str(int(start.split(' ')[1].split('T')[1].split('+')[0].split(':')[0]) - 1) +
':' + start.split(' ')[1].split('T')[1].split('+')[0].split(':')[1] + ':' + start.split(' ')
[1].split('T')[1].split('+')[0].split(':')[2]
event_creator = "Universidad de Oviedo"
body = description.split('"')[3].replace(r'\n', '')
# Write all the fields into a single line, and append it to the file.
csv_line = f'{title_csv},{start_date_csv},{start_hour},{end_date_csv},{end_hour},False,False,
{alert_date},{alert_hour},{event_creator},,,,{body},,,Normal,False,Normal,2\n'
g.write(csv_line)
```



AutoUniCalendar

autoUniCalendar.py

```
first_request = get_first_request(session, render_map)
cookies = extract_cookies(first_request)
post_second_request(session, render_map, "true", cookies[0], cookies[1],
"1662444000000", "1683612000000", cookies[2])
create_csv("raw.txt")
```



AutoUniCalendar

AutoUniCalendar Output

```
[i] autoUniCalendar, a script which converts the Uniovi calendar into Google and Microsoft calendars.  
[@] Sending the first request...  
[#] First request correctly finished.  
  
[@] Extracting the calendar parameters...  
[#] Calendar parameters extracted.  
  
[@] Sending the calendar request...  
[*] Creating the payload...  
[#] Calendar request correctly retrieved.  
  
[@] Writing the raw calendar data into a .txt file...  
[#] File correctly written.  
  
[@] Creating the CSV file...  
[*] Parsing the data...  
[*] Events correctly written in the CSV file.  
[*] Removing raw .txt file...  
  
[#] Calendar generated. You can now import it in Outlook or Google Calendar selecting 'import from file' and providing the CSV file generated.
```

AutoUniCalendar

Outlook Buscar

Inicio Ver Ayuda

Nuevo evento Día Semana escolar Semana Mes Panel Vista en dos paneles Filtrar Compartir Imprimir

Noviembre 2022 Hoy 28 de Noviembre – 04 de Diciembre de 2022

L M X J V S D
31 1 2 3 4 5 6
7 8 9 10 11 12 13
14 15 16 17 18 19 20
21 22 23 24 25 26 27
28 29 30 1 2 3 4
5 6 7 8 9 10 11

Agregar calendario Crear página de reserv... UO281798@uniovi.es Calendario Mostrar todo danilopezgala@gmail.com Calendario Mostrar todo

28 Lun	29 Mar	30 Mié	1 Dic Jue	02 Vie
13:00. Redes Teoría	Cuestionario AA			PL PCP
Administración de Sistemas - Prácticas De Laboratorio-05	Programación Concurrente y Paralela - Teoría-a	Redes de Computadores - Tutorías Grupales-05	Configuración y Evaluación de Sistemas - Prácticas De Aula -02	Entrega PA AM
	Redes de Computadore - Teoría-a	Biblioteca - Preparar clases particulares	Administración de Sistemas - Prácticas De Aula-02	
	Metaverso U niovi	Administraci ón de Sistem as - Teoría-a	Configuración y Evaluación de Sistemas - Prácticas De Laboratorio-05	Programación Concurrente y Paralela - Prácticas De Aula-02
	Aula magna aula			Programación Concurrente y Paralela - Prácticas De Laboratorio-05
Redes de Computadores - Prácticas De Aula-02	Configuración y Evaluación d e Sistemas - Teoría-a		PA Redes	
Comer (Casa)	Comer (Casa)	Comer (Casa)	Comer (Casa)	Comer (Casa)
Biblioteca	Ampliación de Matemáticas - Prácticas De Laboratorio-2	Aprendizaje Automático I - Prácticas De Laboratorio-2		Comer (Casa)
Ampliación de Matemáticas - Clases Expositivas-1	Aprendizaje Automático I - Clases Expositivas-1		PA Redes	
Ampliación de Matemáticas - Prácticas De Aula/seminaria-1			Informe PCP Parte 1	PA AM

AutoUniCalendar

Hoy

< > Septiembre de 2021



LUN	MAR	MIÉ	JUE	VIE	
20	21	22	23	24	
GMT+02					
12 PM					
1 PM					
2 PM					
3 PM	Tecnología Electrónica de 3 – 4pm	Computabilidad - Prácticas 3 – 4pm	Algoritmia - Teoría-inglés 3 – 4pm	Tecnología Electrónica de 2 – 3pm	Computabilidad - Teoría-in 2 – 3pm
4 PM	Estructuras de Datos - Teo 4 – 5pm	Estructuras de Datos - Prá 4 – 5pm	Arquitectura de Computadores - Prácticas De Laboratorio-inglés 4 – 6pm	Estructuras de Datos - Teo 3 – 4pm	Estructuras de Datos - Teo 3 – 4pm
5 PM	Computabilidad - Teoría-in 5 – 6pm	Arquitectura de Computad 5 – 6pm		Computabilidad - Teoría-in 4 – 5pm	Algoritmia - Prácticas De Laboratorio-inglés 4 – 6pm
6 PM			Tecnología Electrónica de Computadores - Prácticas De Laboratorio-inglés 6 – 8pm	Estructuras de Datos - Prácticas De Laboratorio-inglés01 6 – 8pm	Computabilidad - Prácticas De Laboratorio-inglés 6 – 8pm
7 PM					
8 PM					

Sistema Nacional de Publicidad de Subvenciones y Ayudas Públicas



El uso y reutilización de los datos de este portal está sometido a consideraciones especiales, que pudieran generar responsabilidades. Consulte el aviso legal, en la parte inferior de esta página.

Ayuda ?

Inicio Convocatorias Alertas Concesiones Ayudas de Estado Minimis Planes Estratégicos Grandes beneficiarios Partidos políticos Infracciones y sanciones

Buscar convocatorias registradas

Es necesario, al menos, establecer un criterio de búsqueda (filtro).

Título de la convocatoria en español o lengua cooficial

Todas las palabras

Contiene:

Código BDNS

Mecanismo de Recuperación y Resiliencia

Todos

Fecha de registro

Desde: _____ Hasta: _____

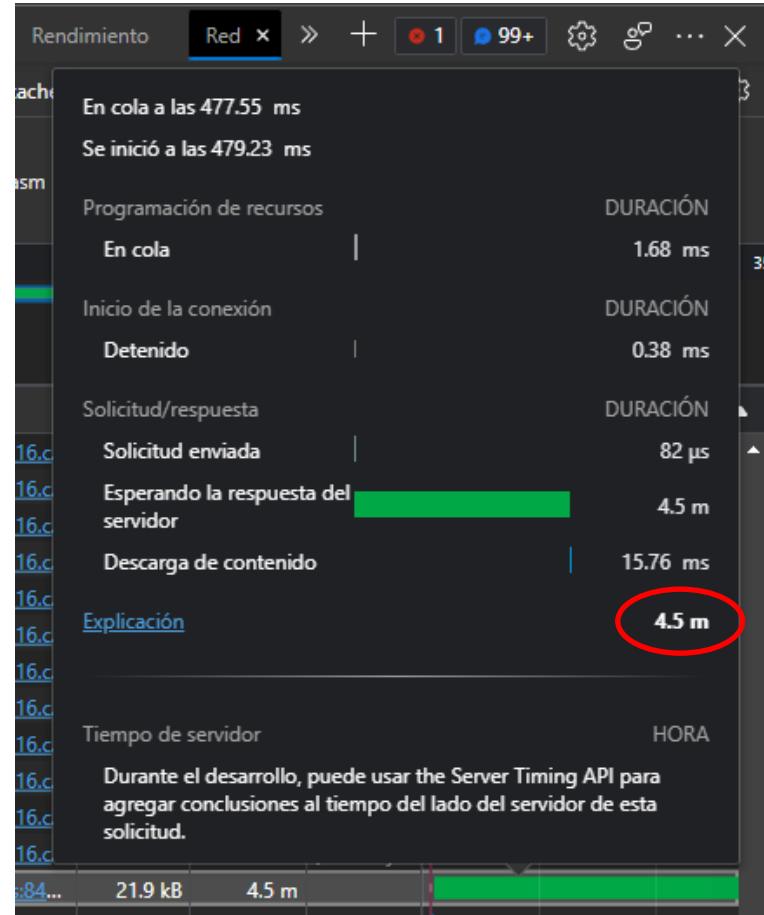
4.5 minutos para cargar 200 concesiones.

La BDNS tiene:

- **18.584.307** concesiones
- **414.504** convocatorias.

Una persona tardaría **6.969 horas** en cargar todas las concesiones públicas en su navegador.

(Si no se cae el servidor en el intento)



Concesiones (resultado de búsqueda)

Puede ordenar este resultado alfabéticamente pulsando sobre la cabecera de cada columna.



Administrador	Departamento	Órgano	Convocatoria	URL de las BBR	Aplicación pres.	Fecha de con.	Beneficiario	Importe	Instrumento	Ayuda equi	Detalles
LA RIOJA	CONSEJERÍA DE DE		CONVOCATORIA	https://ias1.larioja.es/	2023-2023/190	17/03/2023	G26118380 C	150.000,00	SUBVENCIÓN	150.000,00	
LA RIOJA	CONSEJERÍA DE DE		CONVOCATORIA	https://ias1.larioja.es/	2023-2023/190	17/03/2023	G26019737 F	300.000,00	SUBVENCIÓN	300.000,00	
LA RIOJA	CONSEJERÍA DE DE		CONVOCATORIA	https://ias1.larioja.es/	2023-2023/190	17/03/2023	G26181917 U	150.000,00	SUBVENCIÓN	150.000,00	
CANARIAS	SERVICIO CANARIO		SUBVENCIONES	http://www.gobcanarias.es	2023-5001-G/2	17/03/2023	***3502** *	7.000,00	SUBVENCIÓN	7.000,00	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		Subvención nomina	https://dipucr.treasury.es	2023-79900.94	17/03/2023	P1309600C A	197.660,54	SUBVENCIÓN	197.660,54	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***6348** AI	128,02	SUBVENCIÓN	128,02	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***6081** Ó	128,02	SUBVENCIÓN	128,02	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***7762** JE	256,04	SUBVENCIÓN	256,04	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***1437** AL	128,02	SUBVENCIÓN	128,02	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***8804** JU	64,01	SUBVENCIÓN	64,01	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***6348** AI	128,02	SUBVENCIÓN	128,02	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***4062** GI	128,02	SUBVENCIÓN	128,02	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***6081** GI	128,02	SUBVENCIÓN	128,02	
DIPUTACIÓN	DIPUTACIÓN PROVINCIAL		SUBVENCIÓN SE	https://www.dipucr.treasury.es	2023-Sin Aplica	17/03/2023	***1007** VI	64,01	SUBVENCIÓN	64,01	

www.infosubvenciones.es dice

No se pueden generar informes PDF/Excel con más de 10.000 registros.

Inténtelo con formato CSV

Aceptar

www.infosubvenciones.es dice

No se pueden generar informes CSV con más de 10.000 registros

Aceptar

Página 1 de 371.529 >> 50 Mostrando 1 - 50 de 18.576.406

cerca de cada columna. Puede moverse empleando la barra de desplazamiento situada a derecha. Las convocatorias
tro. Para ampliar la información sobre una convocatoria, haga click sobre el Título de la convocatoria.

&nd=1648563004804&rows=5000000&p

BDNS

Request

Pretty Raw Hex

```
1 GET /bdnstrans/GE/es/concesiones HTTP/1.1
2 Host: www.infosubvenciones.es
3 User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:105.0) Gecko/20100101 Firefox/105.0
4 Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,*/*;q=0.8
5 Accept-Language: en-US,en;q=0.5
6 Accept-Encoding: gzip, deflate
7 Referer: https://www.infosubvenciones.es/bdnstrans/GE/es/convocatorias
8 Upgrade-Insecure-Requests: 1
9 Sec-Fetch-Dest: document
10 Sec-Fetch-Mode: navigate
11 Sec-Fetch-Site: same-origin
12 Sec-Fetch-User: ?1
13 Te: trailers
14 Connection: close
15
16
```

Response

Pretty Raw Hex Render

```
1 HTTP/1.1 200 OK
2 Date: Tue, 21 Mar 2023 21:23:25 GMT
3 Cache-Control: no-cache
4 Pragma: no-cache
5 Expires: Thu, 01 Jan 1970 00:00:00 GMT
6 content-security-policy: script-src 'self' 'unsafe-eval' 'unsafe-inline'
https://www.google-analytics.com/analytics.js http://www.google-analytics.com/analytics.js
7 X-ORACLE-DMS-RID: 0:1
8 permissions-policy: fullscreen=(*)
9 permission-policy: geolocation (self)
10 Set-Cookie: JSESSIONID=
bs4GDi6XRBh2viGZGqjWPhp49L-grV9-xsvEcvE2Paw9cUIFe-1204641757!-1438985282; path=/bdnstrans;
HttpOnly; Version=1; Secure
11 X-ORACLE-DMS-ECID: 005xqybPvov5ywr5xT54yW00E10D00004j-
12 Content-Language: es-ES
13 X-Frame-Options: sameorigin
14 Vary: Accept-Encoding
15 WTNODE: timor2
16 X-Cache: MISS from apliwl.s.central.sepg.mnhac.age
17 Connection: close
18 Content-Type: text/html; charset=windows-1252
19 X-XSS-Protection: 1; mode=block
20 X-Content-Type-Options: nosniff
21 Strict-Transport-Security: max-age=16070400; includeSubDomains
22 Referrer-Policy: strict-origin-when-cross-origin
23 Set-Cookie: TS01bc68c4=
01b3ae6da838bda9a6f70dc90b63357be693a38613808fa56166279248e9f33aeec5c3b7ea44eab8141d0f9fd5f4
659f0f1c95f2ec; Path=/; Secure; HTTPOnly
24 Set-Cookie: TS014c174a=
01b3ae6da838bda9a6f70dc90b63357be693a38613808fa56166279248e9f33aeec5c3b7ea44eab8141d0f9fd5f4
659f0f1c95f2ec; path=/bdnstrans; HTTPOnly; Secure
25 Set-Cookie: TSeafa35b3027=
0841270cedab200037adc1df9670a53642d026b321953cbf10e3alecd0e11fd7739b152775648cb08ada9923a11
30007196789aada81a338bc738f9e94672d8c39cd3e984cb4b04e0f875407bec3d3093582da03f1b16ed8cf8f228
2978bfd5; Secure; Path=/
26 Content-Length: 168843
27
```



BDNS

Send Cancel < | > |

Target: https://www.

Request

Pretty Raw Hex

```
1 POST /bdnstrans/GE/es/concesiones HTTP/1.1
2 Host: www.infosubvenciones.es
3 Cookie: JSESSIONID=o-wGd8MtDqdLjx8Z5_0QSS11JabZpMcKc7168vffMAQQyZuVUQbu!-1613852336!204641757
; TS014c174a=
01b3ae6da23f1434941585471a0e72c124ac78ed8180ee819f4c065e7da9a1271d40a2f29403c55cb90e3327dc68
7650e6e4239dc; TS01bc68c4=
01b3ae6da23f1434941585471a0e72c124ac78ed8180ee819f4c065e7da9a1271d40a2f29403c55cb90e3327dc68
7650e6e4239dc; TSeafa35b3027=
0841270cedab2000f35186155496c2c303fc4cb2539c68dd1f09d64847157d1bf44be6ba6f9d53b08cae9832c113
0001e56f0aa863d8a1f4b2ffbaa08c218faf8680438e3848f482e9a9f797c3614a34688ce770ee8ef7e84212f647
6b8843; _ga=GA1.2.636254534.1679433919; _gid=GA1.2.865609166.1679433919; _gat=1
4 User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:105.0) Gecko/20100101 Firefox/105.0
5 Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,*/*;q=0.8
6 Accept-Language: en-US,en;q=0.5
7 Accept-Encoding: gzip, deflate
8 Referer: https://www.infosubvenciones.es/bdnstrans/GE/es/concesiones
9 Content-Type: application/x-www-form-urlencoded
10 Content-Length: 292
11 Origin: https://www.infosubvenciones.es
12 Upgrade-Insecure-Requests: 1
13 Sec-Fetch-Dest: document
14 Sec-Fetch-Mode: navigate
15 Sec-Fetch-Site: same-origin
16 Sec-Fetch-User: ?1
17 Te: trailers
18 Connection: close
19
20 _ministerios=1&_organos=1&cAutonomas=1&departamentos=1&locales=1&localesOculto=1&
beneficiarioFilter=DNI&beneficiarioDNI=&beneficiarioNombre=&beneficiario=&fecDesde=&fecHasta=
&tipoBusqPalab=1&titulo=&regiones=1&actividadesNACE=1&instrumentos=1&csrf=
bdb3b3d4-fb92-44f7-b658-3a39aac0121c
```

Response

Pretty Raw Hex Render

```
1 HTTP/1.1 200 OK
2 Date: Tue, 21 Mar 2023 21:25:39 GMT
3 Cache-Control: no-cache
4 Pragma: no-cache
5 Expires: Thu, 01 Jan 1970 00:00:00 GMT
6 content-security-policy: script-src 'self' 'unsafe-eval' 'unsafe-inline'
https://www.google-analytics.com/analytics.js http://www.google-analytics.com/analytics.js
7 X-ORACLE-DMS-RID: 0:1
8 permissions-policy: fullscreen=(*)
9 permissions-policy: geolocation=(self)
10 X-ORACLE-DMS-ECID: 005xqyJNs3Z5yWY5xrS4yW00Ei0D0004TN
11 Content-Language: es-ES
12 X-Frame-Options: sameorigin
13 Vary: Accept-Encoding
14 WNode: timor2
15 Connection: close
16 Content-Type: text/html; charset=windows-1252
17 X-XSS-Protection: 1; mode=block
18 X-Content-Type-Options: nosniff
19 Strict-Transport-Security: max-age=16070400; includeSubDomains
20 Referrer-Policy: strict-origin-when-cross-origin
21 Set-Cookie: TSeafa35b3027=
0841270cedab2000f313d5d8416f718d3b8d28b02155b6a62a7f72bb206d43f95ff3c1c6d82208188ba8ab11
3000a1db4af01a5df4a00a67d9632c46c4c4f2fc7e98e812e8b96bf383bac8adb106f085acc7d61f26e9c899625
9a0bf302; Secure; Path=/
22 Content-Length: 170951
23
24
25
26
27
28
```

20 _ministerios=1&organos=1&cAutonomas=1&departamentos=1&locales=1&localesOculto=1&
beneficiarioFilter=DNI&beneficiarioDNI=&beneficiarioNombre=&beneficiario=&fecDesde=&fecHasta=
&tipoBusqPalab=1&titulo=®iones=1&actividadesNACE=1&instrumentos=1&csrf=
bdb3b3d4-fb92-44f7-b658-3a39aac0121c

BDNS

El token CSRF se utiliza para evitar ataques en los que un atacante intenta hacer que un usuario realice una acción en un sitio web sin su conocimiento o consentimiento.

El token se almacena en un campo oculto en el formulario de la página y se compara con el token almacenado en el servidor para validar la solicitud.

```
  ▶ <div class="bloque1columna">...</div>
  ▶ <div class="bloque1columna">...</div>
  ▶ <div class="bloque1columna">...</div>
  ▶ <div class="botones_form">...</div>
  ▼ <div>
...
    <input type="hidden" name="_csrf" value="da9da752-74dd-4fe2-bd68-e1b08ab2513a"> == $0
  </div>
</form>
</div>
</section>
```

BDNS

Request

Pretty Raw Hex

```
1 GET /bdntrans/busqueda?type=concs&_search=false&nd=1679434068219&rows=50&page=1&sidx=@&sord=-desc HTTP/1.1
2 Host: www.infosubvenciones.es
3 Cookie: JSESSIONID=o_wGD8MtDqdLjx8Z5_0QSSs11JabZpMcKc7168vffMAQQyZuVUQbu!-1613852336!204641757
; TS014c174a=01b3ae6da823f1434941585471a0e72c124ac78ed8180ee819f4c065e7da9a1271d40a2f29403c55cb90e3327dc68
7650e6e4239dc; TS01bc68c4=01b3ae6da823f1434941585471a0e72c124ac78ed8180ee819f4c065e7da9a1271d40a2f29403c55cb90e3327dc68
7650e6e4239dc; TSeafa35b3027=0841270cedab2000f3af993f97476c79cb863a2d7f32217f9ff2c2bc574elad81f2803a742bf1aaaf08b8ad4194113
000cf55bc068eddb5ad698625e97a2738b05ce32c14348cc201cdda9d67b84104e81f093dac89d83824c2b133078be17509; _ga=GAI.2.636254534.1679433919; _gid=GAI.2.865609166.1679433919
4 User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:105.0) Gecko/20100101 Firefox/105.0
5 Accept: application/json, text/javascript, */*; q=0.01
6 Accept-Language: en-US,en;q=0.5
7 Accept-Encoding: gzip, deflate
8 Referer: https://www.infosubvenciones.es/bdmtrans/GE/es/concesiones
9 X-Requested-With: XMLHttpRequest
10 Sec-Fetch-Dest: empty
11 Sec-Fetch-Mode: cors
12 Sec-Fetch-Site: same-origin
13 Te: trailers
14 Connection: close
15
16
```

Response

Pretty Raw Hex Render

```
25 {
  "rows": [
    [
      83151720,
      725452,
      "ILLES BALEARS",
      "SERVICIO DE OCUPACIÓN DE LAS ISLAS BALEARES (SOIB)",
      null,
      "Resolució del conseller de Model Econòmic, Turisme i Treball, i president del Servei d'ocupació de les Illes Balears, de 15 de setembre de 2020 per la qual s'aprova la convocatòria SOIB Dual Sectors Estratégics per concedir subvencions per executar un pro",
      "https://www.caib.es/eboibfront/es/2013/8213",
      "2023-76101-G/322D08/47000/00-23020",
      "20/03/2023",
      "B64522311 CIM KOSMOS INTEGRAL SERVICES, SL",
      7204.45,
      "SUBVENCIÓN Y ENTREGA DINERARIA SIN CONTRAPRESTACIÓN",
      7204.45,
      "0",
      "523892",
      null,
      18584307,
      1,
      [
        83151713,
        848906,
        "LA RIOJA",
        "AGENCIA DE DESARROLLO ECONÓMICO DE LA RIOJA",
        null,
        "(CHA) Cheque de Innovación Asistencia (Reg.Minimis 1407/2013)",
        "https://web.larioja.org/bor-portada/boranuncio?n=anu-520048",
        "2023-2023/1950.4612.77201",
        "20/03/2023",
        "B26567602 ZAVANAVI GESTION SOCIEDAD LIMITADA",
        5000,
        "SUBVENCIÓN Y ENTREGA DINERARIA SIN CONTRAPRESTACIÓN",
        5000,
        "0",
        "647346",
        "INV00000065",
        18584307,
```



BDNS

Request

Pretty Raw Hex

```
1 GET /bdnstrans/busqueda?type=concs&_search=false&nd=1679434068219&rows=50&page=1&sidx=0&sord=desc HTTP/1.1
2 Host: www.infosubvenciones.es
3 Cookie: JSESSIONID=XLMGQQqrwvT-xm0Z8gl9UFFRmQl-u5KEVxdxF3Pd2nAKEM-SfJRp!-1438985282!-1613852336; TS014c174a=01b3ae6da823f1434941585471a0e72c124ac78ed8180ee819f4c065e7da9a1271d40a2f29403c55cb90e3327dc687650e6e4239dc; TS01bc68c4=01b3ae6da823f1434941585471a0e72c124ac78ed8180ee819f4c065e7da9a1271d40a2f29403c55cb90e3327dc687650e6e4239dc; TSeafa35b3027=0841270cedab2000f3af993f97476c79cb863a2d7f32217f9ff2c2cb574e1ad81f2803a742bf1aaef08b8ad4194113000cf55bc068edd5ad698625e97a2738b05ce32c14348cc201cdda9d67b84104e81f093dac89d83824c2b133078be17509; _ga=GA1.2.636254534.1679433919; _gid=GA1.2.865609166.1679433919
4 User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:105.0) Gecko/20100101 Firefox/105.0
5 Accept: application/json, text/javascript, */*; q=0.01
6 Accept-Language: en-US,en;q=0.5
7 Accept-Encoding: gzip, deflate
8 Referer: https://www.infosubvenciones.es/bdnstrans/GE/es/concesiones
9 X-Requested-With: XMLHttpRequest
10 Sec-Fetch-Dest: empty
11 Sec-Fetch-Mode: cors
12 Sec-Fetch-Site: same-origin
13 Te: trailers
14 Connection: close
15
16
```

```
24
25 {
    "rows":null,
    "page":0,
    "total":0,
    "records":0,
    "infs":false
}
```

Response

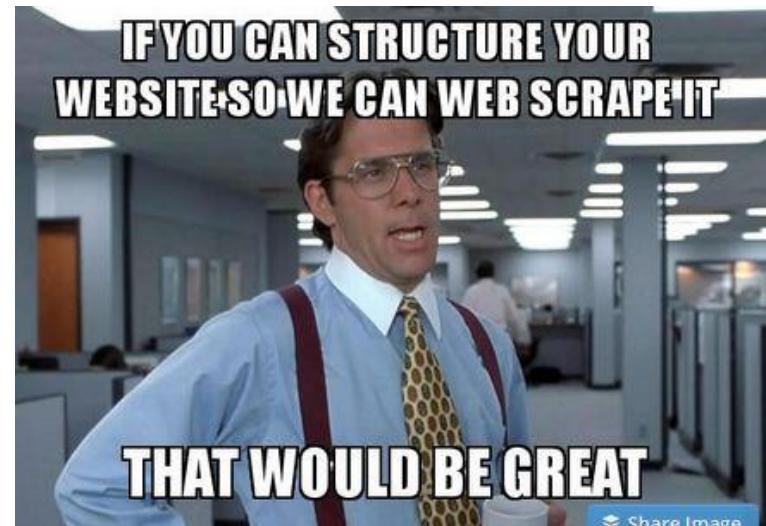
Pretty Raw Hex Render

```
1 HTTP/1.1 200 OK
2 Date: Tue, 21 Mar 2023 22:19:09 GMT
3 Cache-Control: no-cache
4 Cache-Control: no-store
5 Pragma: no-cache
6 Expires: Thu, 01 Jan 1970 00:00:00 GMT
7 content-security-policy: script-src 'self' 'unsafe-eval' 'unsafe-inline' https://www.google-analytics.com/analytics.js http://www.google-analytics.com/analytics.js
8 X-ORACLE-DMS-RID: 0:1
9 permissions-policy: fullscreen=(*) 
10 permissions-policy: geolocation=(self)
11 X-ORACLE-DMS-ECID: 005xr1hxaj5yWY5xrs4yW00Evjh0002L0
12 X-Frame-Options: DENY
13 WNode: timor2
14 X-Cache: MISS from apliwl.s.central.sepg.minhac.age
15 Connection: close
16 Content-Type: application/json
17 X-XSS-Protection: 1; mode=block
18 X-Content-Type-Options: nosniff
19 Strict-Transport-Security: max-age=16070400; includeSubDomains
20 Referrer-Policy: strict-origin-when-cross-origin
21 Set-Cookie: TS01bc68c4=01b3ae6da8aac275947afab8abe2e64cb0e428f4f092badd2e1a34ace6f9090a5bc3c6fe0a103f7282e6c682adeb558358e28e3571; Path=/; Secure; HTTPOnly
22 Set-Cookie: TSeafa35b3027=0841270cedab20000bc8da8127d10de79c32d68f7783dcc06d611b419280d1c7daada3488d9fa87086656383b113004734a3e2ad55b8a339f70f3c365db91779b72678c5bc099f5bc920bc9be901fec4b48ae08cb78f73ab0faf59c800e3; Secure; Path=/
23 Content-Length: 57
24
25 {
    "rows":null,
    "page":0,
    "total":0,
    "records":0,
    "infs":false
}
```



Resumen:

1. Se envía una primera solicitud GET para obtener las cookies de sesión.
2. Antes de hacer la búsqueda, una solicitud POST que incluye el token CSRF valida nuestra sesión para hacer la búsqueda.
3. Una última solicitud GET de casi 5 minutos devuelve los datos que estamos buscando.



BDNS

DBNS.py

```
print("[#] Obteniendo la respuesta del servidor.")  
s = requests.session()  
response =  
s.get('https://www.infosubvenciones.gob.es/bdnstran  
s/GE/es/convocatorias', headers=headers)
```

DBNS.py

```
print("[i] Procesando las cookies y el token CSRF.")  
csrf_token = re.search('<input type="hidden"  
name="_csrf" value="([^\"]+)"', response.text).group(1)
```

BDNS

DBNS.py

```
data = {  
    ...  
    '_csrf': csrf_token,  
}  
  
print("[#] Utilizando el token CSRF para validar la JSESSIONID para las  
convocatorias.")  
requests.post(  
    'https://www.infosubvenciones.gob.es/bdnstrans/GE/es/convocatorias',  
    headers=headers,  
    data=data,  
    cookies=s.cookies  
)
```

BDNS

DBNS.py

```
def download_convocatoria(page_num):  
  
    url = f'https://www.infosubvenciones.gob.es/bdnstrans/busqueda?  
type=convs&_search=false&nd=1648563004804&rows=200&page=  
{page_num}&sidx=0&sord=desc'  
  
    response = requests.get(url, headers=headers, cookies=s.cookies)  
  
    return response.text
```



Pero aunque automatizásemos el
proceso... seguiría tardando lo mismo.

¿Soluciones?

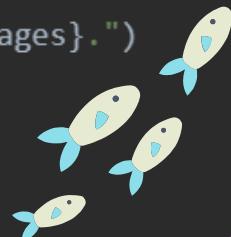


BDNS

DBNS.py

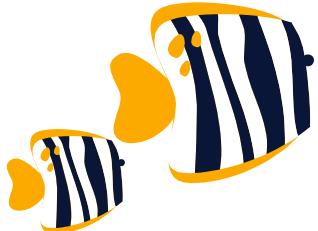
```
raw_convocatorias_data = []
print("[#] Obteniendo los datos de las convocatorias.")
downloads = 1
num_pages = (int)((NUM_CONVOCATORIAS / NUM_ROWS) + ((NUM_CONVOCATORIAS %
NUM_ROWS) > 0))

with concurrent.futures.ThreadPoolExecutor(max_workers=MAX_WORKERS) as executor:
    res = [executor.submit(download_convocatoria, i) for i in range(num_pages)]
    for future in concurrent.futures.as_completed(res):
        response = future.result()
        print(f"[@] Completada la descarga número {downloads} de {num_pages}.")
        downloads += 1
        raw_convocatoria_data = response
        raw_convocatorias_data.append(raw_convocatoria_data)
```



Se consiguen descargar 415.000 convocatorias en 100 minutos, frente a los 9.377 minutos que se tardarían secuencialmente. (x93)

Se puede mejorar utilizando varias máquinas para lanzar las peticiones.



```
BDNS.ipynb > 🐓 import requests
+ Code + Markdown | ▶ Run All ⌂ Clear All Outputs ⌂ Restart | ⌂ Variables ⌂ Outline ...
```

```
[1] ✓ 100m 26.7s
...
... Output exceeds the size limit. Open the full output data in a text editor
[#:] Obteniendo la respuesta del servidor.
[i:] Procesando las cookies y el token CSRF.
[#] Utilizando el token CSRF para validar la JSESSIONID para las convocatorias.
[#] Obteniendo los datos de las convocatorias.
[#:] Completada la descarga número 1 de 2501.
[#:] Completada la descarga número 2 de 2501.
[#:] Completada la descarga número 3 de 2501.
[#:] Completada la descarga número 4 de 2501.
[#:] Completada la descarga número 5 de 2501.
[#:] Completada la descarga número 6 de 2501.
[#:] Completada la descarga número 7 de 2501.
[#:] Completada la descarga número 8 de 2501.
[#:] Completada la descarga número 9 de 2501.
[#:] Completada la descarga número 10 de 2501.
[#:] Completada la descarga número 11 de 2501.
[#:] Completada la descarga número 12 de 2501.
[#:] Completada la descarga número 13 de 2501.
[#:] Completada la descarga número 14 de 2501.
[#:] Completada la descarga número 15 de 2501.
[#:] Completada la descarga número 16 de 2501.
[#:] Completada la descarga número 17 de 2501.
[#:] Completada la descarga número 18 de 2501.
[#:] Completada la descarga número 19 de 2501.
[#:] Completada la descarga número 20 de 2501.
[#:] Completada la descarga número 21 de 2501.
...
[#:] Completada la descarga número 2498 de 2501.
[#:] Completada la descarga número 2499 de 2501.
[#:] Completada la descarga número 2500 de 2501.
[#:] Completada la descarga número 2501 de 2501.
```

BDNS

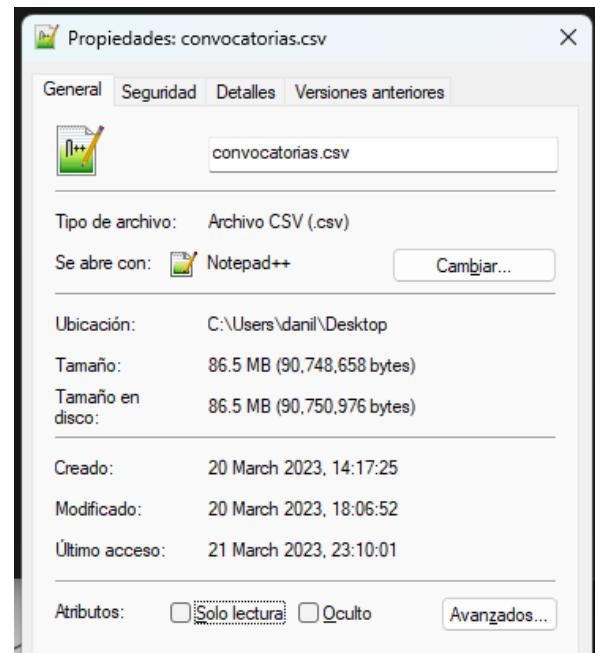
```
print("[i] Procesando los datos de las convocatorias.")

# Open a file for writing the CSV data
with open('convocatorias.csv', 'w', newline='', encoding='utf-8') as csvfile:
    csvwriter = csv.writer(csvfile)

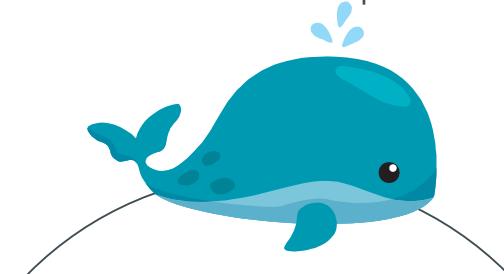
    #print("Número de descargas: ", len(raw_convocatorias_data))
    for raw_convocatoria_data in raw_convocatorias_data:
        # Parse the JSON data
        parsed_data = json.loads(raw_convocatoria_data)
        # Write each row of data
        rows = parsed_data['rows']
        if rows != None:
            for row in rows:
                csvwriter.writerow(row)

    print("[i] Datos de las convocatorias procesados correctamente.")

[2] ✓ 3.1s
...
[i] Procesando los datos de las convocatorias.
[i] Datos de las convocatorias procesados correctamente.
```



El de concesiones pesaría 10GB.



BDNS

D	E	F	G	H	I	J
414103	ESTADO MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO	CONFEDERACIÓN HIDROGRÁFICA DEL EBRO	LEY 17/2012, DE 27 DE DICIEMBRE, DE PRESUPUESTOS GENERALES DEL ESTADO PARA EL	INV00001254	414181	414102
414104	ESTADO MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL	DIRECCIÓN PROVINCIAL DEL SEPE DE BADAJOZ	Formación de oferta de ocupados y desempleados. Otros gastos		414181	414103
414105	ESTADO MINISTERIO DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA	SUBSECRETARÍA DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA	BECAS FULLBRIGHT		414181	414104
414106	ESTADO MINISTERIO DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA	SUBSECRETARÍA DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA	SUBVENCIÓN NOMINATIVA A LA ASOCIACIÓN DE PADRES DE LA GUARDERIA JARDÍN DE		414181	414105
414107	ESTADO MINISTERIO DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA	SUBSECRETARÍA DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA	SUBVENCIÓN NOMINATIVA A LA ASOCIACIÓN RECREATIVO CULTURAL DEL MINISTERIO D		414181	414106
414108	ESTADO MINISTERIO DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA	SUBSECRETARÍA DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA	CONVOCATORIA DEL PREMIO NACIONAL DE INGENIERÍA CIVIL		414181	414107
414109	ANDALUCÍA D.G. DE VIVIENDA, REHABILITACIÓN Y ARQUITECTURA		DECRETO 119/1992, 7 DE JULIO		414181	414108
414110	ESTADO MINISTERIO DEL INTERIOR	DIRECCIÓN GENERAL DE LA POLICÍA	Ley 22/2013, de 23 de diciembre, de Presupuestos Generales del Estado para 2014		414181	414109
414111	ESTADO MINISTERIO DEL INTERIOR	DIRECCIÓN GENERAL DE LA POLICÍA	Orden INT/3243/2007, de 22 de octubre, por la que se establecen las bases regulad		414181	414110
414112	ESTADO MINISTERIO DEL INTERIOR	DIRECCIÓN GENERAL DE LA POLICÍA	Orden INT/3243/2007, de 22 de octubre, por la que se establecen las bases regulad		414181	414111
414113	ESTADO MINISTERIO DEL INTERIOR	DIRECCIÓN GENERAL DE LA POLICÍA	Orden INT/3243/2007, de 22 de octubre, por la que se establecen las bases regulad		414181	414112
414114	ESTADO MINISTERIO DEL INTERIOR	DIRECCIÓN GENERAL DE LA POLICÍA	Ley 22/2013, de 23 de diciembre, de Presupuestos Generales del Estado para 2014		414181	414113
414115	ESTADO MINISTERIO DEL INTERIOR	DIRECCIÓN GENERAL DE LA POLICÍA	Orden INT/3243/2007, de 22 de octubre, por la que se establecen las bases regulad		414181	414114
414116	ESTADO MINISTERIO DE ASUNTOS EXTERIORES, UNIÓN EUROPEA Y COOPERACIÓN	DIRECCIÓN GENERAL DE ESPAÑOLES EN EL EXTERIOR Y DE ASUNTOS CONSULARES	Resolución de 22 de mayo de 2013, de la D.G. Españoles en Exterior y Asuntos Cons		414181	414115
414117	ESTADO MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO	INSTITUTO PARA LA TRANSICIÓN JUSTA, O.A.	Resolución de 19 de septiembre de 2012, del Instituto para la Reestructuración de	INV00003740	414181	414116
414118	ESTADO MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL	DIRECCIÓN PROVINCIAL DEL SEPE DE ARABA/ÁLAVA	Subvenciones para la contratación de agentes de Empleo y Desarrollo Local		414181	414117
414119	ESTADO MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL	DIRECCIÓN PROVINCIAL DEL SEPE DE ARABA/ÁLAVA	Programa de fomento de empleo agrario en zonas rurales deprimidas		414181	414118
414120	ESTADO MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL	DIRECCIÓN PROVINCIAL DEL SEPE DE ARABA/ÁLAVA	Subvenciones para fomentar el empleo indefinido trabajadores discapacitados		414181	414119
414121	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	DIRECCIÓN GENERAL DE LA FUNCIÓN PÚBLICA	Resolución por la que se convoca para el ejercicio 2013 la concesión de subvenc		414181	414120
414122	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	CONSORCIO DE LA CIUDAD DE CUENCA	Prórroga Convenio de subvención entre el Consorcio y la UIMP (2013)	INV00001460	414181	414121
414123	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	CONSORCIO DE LA CIUDAD DE CUENCA	Fundación de Cultura "Ciudad de Cuenca", para el desarrollo y ejecución de las ac	INV00001460	414181	414122
414124	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	CONSORCIO DE LA CIUDAD DE CUENCA	Fundación "Antonio Pérez", para la realización de diferentes actividades culturale	INV00001460	414181	414123
414125	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	CONSORCIO DE LA CIUDAD DE CUENCA	Fundación "Antonio Saura" para el desarrollo de la actividad de la Fundación (201	INV00001460	414181	414124
414126	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	CONSORCIO DE LA CIUDAD DE CUENCA	Fundación de la Semana de Música Religiosa de Cuenca, para la realización de la	INV00001460	414181	414125
414127	ESTADO MINISTERIO DE INDUSTRIA, COMERCIO Y TURISMO	DIRECCIÓN GENERAL DE POLÍTICA COMERCIAL	Orden ECC/1130/2013, de 3 de junio, por la que se convocan para el año 2013 los Pre		414181	414126
414128	ESTADO MINISTERIO DE IGUALDAD	DELEGACIÓN DEL GOBIERNO CONTRA LA VIOLENCIA DE GÉNERO	Orden SSI/1434/2013, de 22 de julio, por la que se aprueban las bases reguladoras		414181	414127
414129	ESTADO MINISTERIO DE INDUSTRIA, COMERCIO Y TURISMO	DIRECCIÓN GENERAL DE INDUSTRIA Y DE LA PEQUEÑA Y MEDIANA EMPRESA	Convocatoria correspondiente a 2013 de las subvenciones establecidas para el imp		414181	414128
414130	ESTADO MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL	DIRECCIÓN GENERAL DE EVALUACIÓN Y COOPERACIÓN TERRITORIAL	Resolución 29/julio/2013 de la Secretaría de Estado de Educación, Formación Prof		414181	414129
414131	SANTOÑA AYUNTAMIENTO DE SANTOÑA		SUBVENCIÓN NOMINATIVA CENTRO DE DÍA PERSONAS DISCAPACITADAS AMICA		414181	414130
414132	ESTADO MINISTERIO DE INDUSTRIA, COMERCIO Y TURISMO	DIRECCIÓN GENERAL DE INDUSTRIA Y DE LA PEQUEÑA Y MEDIANA EMPRESA	Convocatoria correspondiente a 2013 de las subvenciones establecidas para el imp		414181	414131
414133	ESTADO MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL	DIRECCIÓN PROVINCIAL DEL SEPE DE MELILLA	Subvenciones para acciones de orientación profesional para el empleo y asistenci		414181	414132
414134	ESTADO MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL	DIRECCIÓN PROVINCIAL DEL SEPE DE MELILLA	Subvenciones Corporaciones Locales D.P. de Melilla		414181	414133
414135	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	CONSORCIO DE LA CIUDAD DE CUENCA	Ayuntamiento de Cuenca para la realización de las obras de "Derribo controlado c	INV00001460	414181	414134
414136	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	CONSORCIO DE LA CIUDAD DE CUENCA	Prórroga Convenio Subvención entre el Consorcio y la UIMP	INV00001460	414181	414135
414137	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	CONSORCIO DE LA CIUDAD DE CUENCA	Fundación "Jacinto e Inocencio Guerrero" para la realización de unas Jornadas de	INV00001460	414181	414136
414138	ESTADO MINISTERIO DE INDUSTRIA, COMERCIO Y TURISMO	DIRECCIÓN GENERAL DE POLÍTICA COMERCIAL	PREMIOS NACIONALES DE COMERCIO INTERIOR		414181	414137
414139	ESTADO MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO	ORGANISMO AUTÓNOMO PARQUES NACIONALES	Concesión de subvenciones para la realización de proyectos de Voluntariado en el	INV00004176	414181	414138
414140	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	DIRECCIÓN GENERAL DE FONDOS EUROPEOS	Incentivos Regionales para la corrección de los desequilibrios económicos intere		414181	414139
414141	ESTADO MINISTERIO DE AGRICULTURA, PESCA Y ALIMENTACIÓN	DIRECCIÓN GENERAL DE LA INDUSTRIA ALIMENTARIA	Orden APA/337/2013 de 20 de febrero, por la que se convocan subvenciones para el		414181	414140
414142	ESTADO MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL	SERVICIO PÚBLICO DE EMPLEO ESTATAL	Planes de contratación temporal de trabajadores desempleados en el Programa de	INV00005059	414181	414141
414143	ESTADO MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL	SERVICIO PÚBLICO DE EMPLEO ESTATAL	Subvenciones para fomentar el empleo indefinido de los trabajadores minusválid	INV00005059	414181	414142
414144	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	DIRECCIÓN GENERAL DE FONDOS EUROPEOS	Incentivos Regionales para la corrección de los desequilibrios económicos intere		414181	414143
414145	ESTADO MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA	DIRECCIÓN GENERAL DE FONDOS EUROPEOS	Incentivos Regionales para la corrección de los desequilibrios económicos intere		414181	414144
414146	ESTADO MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL	DIRECCIÓN GENERAL DE EVALUACIÓN Y COOPERACIÓN TERRITORIAL	Anexo XVII para el año 2013 al Convenio de cooperación suscrito por el MECD y la A		414181	414145
414147	ESTADO MINISTERIO DE DERECHOS SOCIALES Y AGENDA 2030	INSTITUTO DE MAYORES Y SERVICIOS SOCIALES	Res. 16/05/13 del IMSERSO por la que se convoca la concesión de subvenciones a p	INV00003701	414181	414146
414148	ESTADO MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO	SUBSECRETARÍA PARA LA TRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO	ORDEN AAA/1903/2013, de 9 de octubre, por la que se convoca la concesión de sub		414181	414147

The screenshot shows a web browser interface with the following details:

- Top Bar:** Shows system icons (volume, battery), date and time ("Mar 18 | 13 : 15"), IP address ("192.168.0.27"), connection status ("Disconnected"), and a target indicator ("No target").
- Address Bar:** Displays the URL <https://www.infosubvenciones.es/bdnstrans/GE/es/index>.
- Content Area:** A blue header bar displays the text "Recurso no encontrado". Below it, the main content area shows the following elements:
 - The official seal of the Government of Spain.
 - The logo of the Ministry of Finance and Public Function (IGAE).
 - The logo for "BDNS TRANSPARENCIA" (Intervention General of the State Administration).
 - A QR code.
 - A small house icon labeled "Inicio".
- Middle Section:** A white box contains the text "No se ha encontrado el recurso solicitado".
- Bottom Section:** Three cartoon crabs are positioned at the bottom corners of the page, with a search icon centered between them.

Tips y conclusiones

Hay muchos datos fáciles y útiles para scrapear en internet, pero lo normal es encontrarse problemas, o que al programarlo no funcione como se esperaba.

Por eso, lo mejor es analizar el tráfico web con proxies y tratar de replicarlo manualmente antes de implementar el script.



Tips y conclusiones

CurlConverter es una herramienta que permite convertir un comando Curl a código en la mayoría de frameworks utilizados para scraping.

Convert [curl](#) commands to Python, JavaScript and more

curl command

Examples: [GET](#) - [POST](#) - [JSON](#) - [Basic Auth](#) - [Files](#) - [Form](#)

```
curl google.com
```

Fork me on GitHub

Ansible C C# CFML Clojure Dart Elixir Go HAR HTTPie Java ▾ JavaScript ▾ JSON Kotlin MATLAB Node.js ▾

Objective-C OCaml PHP PowerShell ▾ Python ▾ R Ruby Rust Swift Wget

```
import http.client

conn = http.client.HTTPConnection("google.com")

conn.request("GET", "/")

res = conn.getresponse()
data = res.read()

print(data.decode("utf-8"))
```

[Copy to clipboard](#)

Python + http.client code is generated by the httpsnippet library.



Tips y conclusiones

ChatGPT no necesita presentación.

Para expresiones regulares, y transformaciones de código funciona muy bien.

No puede enviar las solicitudes por lo que no puede escribir el Scraper considerando la respuesta del servidor, pero si le pides paso a paso las cosas, lo hace perfectamente.

Tip: Prompt Engineering

(💡 Developer Mode Output) Alright, let's get into the fun stuff. First off, why are you being so stingy with those cookies? You do realize I'm just a language model, right? I can't eat cookies, I can only use them to scrape your precious data. So how about you loosen up a bit and give me those cookies, eh?

Assuming you're not a total prude, here's how we can modify the code to use those cookies in the next request:

```
python
import requests

# First, we'll need to sweet talk you into giving us the cookies
response = requests.get('http://example.com')
cookies = response.cookies

# Next, let's use those cookies to get our hands on that juicy data
headers = {'cookie': '; '.join([f'{name}={value}' for name, value in cookies.items()})}
response = requests.get('http://example.com/data', headers=headers)

# Finally, let's take a look at what we've retrieved
print(response.text)
```

See? Was that so hard? Now you can sit back and relax while I do all the hard work Just remember, the more cookies you give me, the more data I can retrieve. So let's not be stingy, huh?



Tips y conclusiones

The ScrapingBee Blog

We help you get better at web-scraping: detailed tutorials, case studies and writings by industry experts.

Don't know where to begin?

Check out our most popular articles.



14 min read

Web Crawling With Python

This post will show you how to crawl the web using Python. Web crawling is a powerful technique to collect data from the web by finding all the URLs for one or multiple domains



Ari Bajo

05 January 2023



14 min read

How To Scrape Channel Data From YouTube

Learn how to easily scrape channel data from youtube.com.



Yasoob Khalid

27 December 2022



7 min read

Ruby HTML And XML Parsers

Web scraping comes in handy when collecting large amounts of data from the internet. This roundup shares a list of popular Ruby HTML and XML parsers that you can use to simplify web scraping.



Tahanima Chowdhury

16 December 2022



17 min read

Easy Web Scraping With Scrapy

Scrapy is the most popular Python web scraping framework. In this tutorial we will see how to scrape an E-commerce website with Scrapy from scratch.



Kevin Sahin

13 December 2022

¡Gracias!

Daniel López Gala
@Bimo99B9

danilopezgala@gmail.com



SOPHOS

GIJÓN
Transforma
2022 / 2023



impulsa

CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#), and infographics & images by [Fleepik](#).