

# WEB SCRAPING CON PYTHON

usando GitHub Actions para automatizarlo todo



# ANTONIO FEREGRINO

Ingeniero de MLOps. Edutuber,  
programador en Python,  
automatizador de tareas. Fan de  
LEGO y Yu-Gi-Oh!

**@feregri\_no** en Twitter... y en todas  
las otras redes sociales.  
**/thatcsharpguy** en YouTube

# ¿Qué vamos a hacer?

Descargar información  
automáticamente de un sitio web  
del gobierno y ponerla disponible  
en un repositorio de GitHub.

# CONTENIDO

01

## PREPARACIÓN

Los primeros bloques del proyecto

02

## WEB SCRAPING

Utilizando `requests` y `beautifulsoup`

03

## AUTOMATIZACIÓN

Usando GitHub Actions



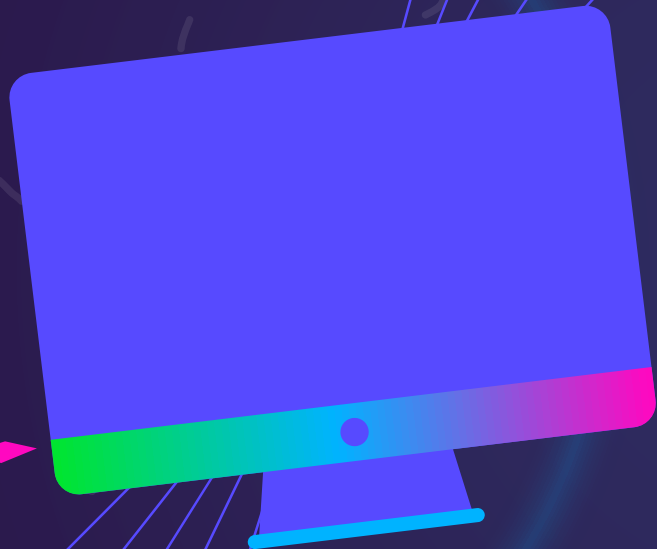
tcsq.dev/pycon22

ACCEDE AL CÓDIGO

Backup: [github.com/fferegrino/uk-blogs/tree/PyConLatam22](https://github.com/fferegrino/uk-blogs/tree/PyConLatam22)

01

# PREPARA CIÓN



# Primeros pasos



## REPOSITORIO

Automatización y  
almacenamiento



## ENTORNO VIRTUAL

Especificación y aislamiento  
de dependencias



## SITIO WEB OBJETIVO

...

# Dependencias

## requests

```
pip install requests
```

Realizar peticiones HTTP utilizando Python

## BeautifulSoup

```
pip install beautifulsoup4
```

Extraer información de archivos HTML.



# Sitio web objetivo



## Estático

No requiere JavaScript para mostrar el contenido



## Público

Contiene información pública



## Homogéneo

Las páginas siguen una misma estructura



## Interesante

Totalmente subjetivo, con que te guste está bien

# GOV.UK blogs



## GOV.UK blogs

### Find a post

You can search for posts that match your interests

Blog post contains:

2780 <sup>posts</sup>

#### [Enabling Networks](#)

5 August 2022 - [Government Science and Engineering](#)

Networking is a crucial skill for scientists and engineers in government, who need to collaborate widely to effectively perform public duties. However, networking can be a daunting prospect, with many factors influencing how comfortable we might feel when networking. Lucy ...

#### [Woodlands expanded for communities across England](#)

5 August 2022 - [Defra in the media](#)

Today we are looking at media coverage of Defra's announcement of funding for England's 13 Community Forests, as well as five regional woodland creation partners.

#### [Making learning part of the job at BEIS Digital](#)

5 August 2022 - [BEIS Digital](#)

A people and development manager talks about how we are making learning part of the job at BEIS Digital.

#### [Environment Agency taking action during prolonged dry weather](#)





# El índice y sus *URLs*

Concentra todo el contenido publicado en los blogs del gobierno

Páginas:

<https://www.blog.gov.uk/all-posts/page/1>

<https://www.blog.gov.uk/all-posts/page/2>

<https://www.blog.gov.uk/all-posts/page/1000>



Todas dirigen a una página válida



# El índice y sus *URLs*

Concentra todo el contenido publicado en los blogs del gobierno

Páginas:

`https://www.blog.gov.uk/all-posts/page/1`

`https://www.blog.gov.uk/all-posts/page/2`

`https://www.blog.gov.uk/all-posts/page/1000`



Todas dirigen a una página válida



02

# WEB SCRAPING

# *Inspect* es tu amiga

- Debes conocer la estructura de la página
- Usa *Inspeccionar* para interactuar con el contenido.

[Click contextual] > *Inspeccionar*

# Find a post

You can search for posts that match your interests

Blog post contains:

# 2780 posts

---

## [Enabling Networks](#)

5 August 2022 - [Government Science and Engineering](#)

Networking is a crucial skill for scientists and engineers in government, who need to collaborate widely to effectively perform public duties. However, networking can be a daunting prospect, with many factors influencing how comfortable we might feel when networking. Lucy ...

---

## [Woodlands expanded for communities across England](#)

5 August 2022 - [Defra in the media](#)

Today we are looking at media coverage of Defra's announcement of funding for England's 13 Community Forests, as well as five regional woodland creation partners.

---

## [Making learning part of the job at BEIS Digital](#)

5 August 2022 - [BEIS Digital](#)

A people and development manager talks about how we are making learning part of the job at BEIS Digital.

---

## [Environment Agency taking action during prolonged dry weather](#)

5 August 2022 - [Creating a better place](#)

In this blog post we explore the latest prolonged dry weather situation and how the Environment Agency has already been taking action.

---

## [Cyber Insider: Meet Ellie](#)

5 August 2022 - [Strategic Command](#)

The National Cyber Force (NCF). A partnership between Defence and Intelligence, and an



# Página con contenido



## GOV.UK blogs

### Find a post

You can search for posts that match your interests

Blog post contains:

2780 posts

#### [Enabling Networks](#)

5 August 2022 - [Government Science and Engineering](#)

Networking is a crucial skill for scientists and engineers in government, who need to collaborate widely to effectively perform public duties. However, networking can be a daunting prospect, with many factors influencing how comfortable we might feel when networking. Lucy ...

#### [Woodlands expanded for communities across England](#)

5 August 2022 - [Defra in the media](#)

Today we are looking at media coverage of Defra's announcement of funding for England's 13 Community Forests, as well as five regional woodland creation partners.



# Página con contenido

```
<ul class="blogs-list">
  <li>
    <h3 class="govuk-heading-m"><a class="govuk-link"
href="https://blog.gov.uk/2022/08/05/enabling-networks/">Enabling Networks</a></h3>
    <div class="meta">
      <span class="govuk-visually-hidden">Posted on: </span>
      <time class="updated" datetime="2022-08-05T18:01:38+01:00" pubdate="">5 August
2022</time>
      -
      <span class="govuk-visually-hidden">On blog: </span>
      <a class="govuk-link" href="https://blog.gov.uk">Government Science and
Engineering</a>
    </div>
    <p>Networking is a (...)</p>
  </li>
  <!-- More content -->
</ul>
```

# Página con contenido

```
<ul class="blogs-list">
  <li>
    <h3 class="govuk-heading-m"><a class="govuk-link"
href="https://blog.gov.uk/2022/08/05/enabling-networks/">Enabling Networks</a></h3>
    <div class="meta">
      <span class="govuk-visually-hidden">Posted on: </span>
      <time class="updated" datetime="2022-08-05T18:01:38+01:00" pubdate="">5 August
2022</time>
      -
      <span class="govuk-visually-hidden">On blog: </span>
      <a class="govuk-link" href="https://blog.gov.uk">Government Science and
Engineering</a>
    </div>
    <p>Networking is a (...)</p>
  </li>
  <!-- More content -->
</ul>
```

# Página con contenido

```
<ul class="blogs-list">
  <li>
    <h3 class="govuk-heading-m"><a class="govuk-link"
href="https://blog.gov.uk/2022/08/05/enabling-networks/">Enabling Networks</a></h3>
    <div class="meta">
      <span class="govuk-visually-hidden">Posted on: </span>
      <time class="updated" datetime="2022-08-05T18:01:38+01:00" pubdate="">5 August
2022</time>
      -
      <span class="govuk-visually-hidden">On blog: </span>
      <a class="govuk-link" href="https://blog.gov.uk">Government Science and
Engineering</a>
    </div>
    <p>Networking is a (...)</p>
  </li>
  <!-- More content -->
</ul>
```

# Página con contenido

```
<ul class="blogs-list">
  <li>
    <h3 class="govuk-heading-m"><a class="govuk-link"
href="https://blog.gov.uk/2022/08/05/enabling-networks/">Enabling Networks</a></h3>
    <div class="meta">
      <span class="govuk-visually-hidden">Posted on: </span>
      <time class="updated" datetime="2022-08-05T18:01:38+01:00" pubdate="">5 August
2022</time>
      -
      <span class="govuk-visually-hidden">On blog: </span>
      <a class="govuk-link" href="https://blog.gov.uk">Government Science and
Engineering</a>
    </div>
    <p>Networking is a (...)</p>
  </li>
  <!-- More content -->
</ul>
```

# Página con contenido

```
<ul class="blogs-list">
  <li>
    <h3 class="govuk-heading-m"><a class="govuk-link"
href="https://blog.gov.uk/2022/08/05/enabling-networks/">Enabling Networks</a></h3>
    <div class="meta">
      <span class="govuk-visually-hidden">Posted on: </span>
      <time class="updated" datetime="2022-08-05T18:01:38+01:00" pubdate="">5 August
2022</time>
      -
      <span class="govuk-visually-hidden">On blog: </span>
      <a class="govuk-link" href="https://blog.gov.uk">Government Science and
Engineering</a>
    </div>
    <p>Networking is a (...)</p>
  </li>
  <!-- More content -->
</ul>
```

# Página sin contenido



## GOV.UK blogs

### Find a post

You can search for posts that match your interests

Blog post contains:

0 posts

### No posts found

Please try:

- searching again using different words



# Página sin contenido

```
<ul class="blogs-list">
  <li class="noresults">
    <h3 class="govuk-heading-m">No posts found</h3>
    <p class="govuk-body">Please try:</p>
    <ul class="govuk-list govuk-list--bullet">
      <li>searching again using different words</li>
    </ul>
  </li>
</ul>
```

# Página sin contenido

```
<ul class="blogs-list">
  <li class="noresults">
    <h3 class="govuk-heading-m">No posts found</h3>
    <p class="govuk-body">Please try:</p>
    <ul class="govuk-list govuk-list--bullet">
      <li>searching again using different words</li>
    </ul>
  </li>
</ul>
```



# Página sin contenido

```
<ul class="blogs-list">
  <li class="noresults">
    <h3 class="govuk-heading-m">No posts found</h3>
    <p class="govuk-body">Please try:</p>
    <ul class="govuk-list govuk-list--bullet">
      <li>searching again using different words</li>
    </ul>
  </li>
</ul>
```

# Página sin contenido

```
<ul class="blogs-list">
  <li class="noresults">
    <h3 class="govuk-heading-m">No posts found</h3>
    <p class="govuk-body">Please try:</p>
    <ul class="govuk-list govuk-list--bullet">
      <li>searching again using different words</li>
    </ul>
  </li>
</ul>
```

The background is a dark purple gradient. On the left, several thin, light purple lines radiate from a single point towards the bottom right. Scattered across the background are various colorful geometric shapes: a small green rectangle, a yellow zigzag, a cyan triangle with a pink outline, a pink diamond, a cyan rectangle, a cyan square with a yellow outline, and another yellow zigzag. There are also faint, light purple dashed lines scattered throughout.

# Scraping!

Construyendo el índice

# Obteniendo urls de una página

```
def get_urls(page):
    final_url = f"https://www.blog.gov.uk/all-posts/page/{page}"
    page_response = requests.get(final_url)

    soup = BeautifulSoup(page_response.text)
    blog_list = soup.find("ul", {"class": "blogs-list"})
    entries = blog_list.find_all('li')

    if entries[0].get('class') != ['noresults']:
        anchors = [entry.select_one("h3 > a") for entry in entries ]
        hrefs = [a['href'] for a in anchors]
        return hrefs
    else:
        return None
```

# Obteniendo urls de una página

```
def get_urls(page):  
    final_url = f"https://www.blog.gov.uk/all-posts/page/{page}"  
    page_response = requests.get(final_url)  
  
    soup = BeautifulSoup(page_response.text)  
    blog_list = soup.find("ul", {"class": "blogs-list"})  
    entries = blog_list.find_all('li')  
  
    if entries[0].get('class') != ['noresults']:  
        anchors = [entry.select_one("h3 > a") for entry in entries ]  
        hrefs = [a['href'] for a in anchors]  
        return hrefs  
    else:  
        return None
```

# Obteniendo urls de una página

```
def get_urls(page):
    final_url = f"https://www.blog.gov.uk/all-posts/page/{page}"
    page_response = requests.get(final_url)

    soup = BeautifulSoup(page_response.text)
    blog_list = soup.find("ul", {"class": "blogs-list"})
    entries = blog_list.find_all('li')

    if entries[0].get('class') != ['noresults']:
        anchors = [entry.select_one("h3 > a") for entry in entries ]
        hrefs = [a['href'] for a in anchors]
        return hrefs
    else:
        return None
```

# Obteniendo urls de una página

```
def get_urls(page):
    final_url = f"https://www.blog.gov.uk/all-posts/page/{page}"
    page_response = requests.get(final_url)

    soup = BeautifulSoup(page_response.text)
    blog_list = soup.find("ul", {"class": "blogs-list"})
    entries = blog_list.find_all('li')

    if entries[0].get('class') != ['noresults']:
        anchors = [entry.select_one("h3 > a") for entry in entries ]
        hrefs = [a['href'] for a in anchors]
        return hrefs
    else:
        return None
```

# Obteniendo urls de una página

```
def get_urls(page):  
    final_url = f"https://www.blog.gov.uk/all-posts/page/{page}"  
    page_response = requests.get(final_url)  
  
    soup = BeautifulSoup(page_response.text)  
    blog_list = soup.find("ul", {"class": "blogs-list"})  
    entries = blog_list.find_all('li')  
  
    if entries[0].get('class') != ['noresults']:  
        anchors = [entry.select_one("h3 > a") for entry in entries ]  
        hrefs = [a['href'] for a in anchors]  
        return hrefs  
    else:  
        return None
```



The background is a dark purple gradient. On the left, several thin, light purple lines radiate from a single point towards the bottom right. Scattered across the background are various colorful geometric shapes: a small green rectangle, a yellow zigzag line, a blue triangle with a pink outline, a pink diamond, a blue parallelogram, a blue square with a pink outline, and another yellow zigzag line. There are also many small, faint, light purple dashes scattered throughout.

# DRY!

*Don't Repeat Yourself*  
No trabajos doble

# Mantén un registro

- Es importante saber qué *URLs* hemos procesado ya
  - Evitamos consumir recursos innecesariamente
- En un sistema más complejo, podría ser una base de datos
- En este caso es un solo archivo: `processed_urls.txt`



# Archivo processed\_urls.txt



Una *url* por línea:

```
https://food.blog.gov.uk/2022/08/04/chairs-stakeholder-update/
```

```
https://space.blog.gov.uk/2022/08/04/how-i-made-it-to-mission-control/
```

```
https://ruralpayments.blog.gov.uk/2022/08/04/rpa-is-supporting-farm-24/
```

```
https://forestrycommission.blog.gov.uk/2022/08/04/reducing-the-impact/
```

```
https://companieshouse.blog.gov.uk/2022/08/04/our-new-power/
```



# Interactuando con el registro

```
def get_scraped_urls():
    scraped_urls = []
    if os.path.exists("scraped_urls.txt"):
        with open("scraped_urls.txt") as url_file:
            for line in url_file:
                scraped_urls.append(line.strip())
    return scraped_urls

def append_scraped_urls(urls):
    with open("scraped_urls.txt", "a") as url_file:
        for url in urls:
            url_file.write(url + "\n")
```



**De vuelta  
al web  
scraping...**

# Obteniendo URLs

- Necesitamos obtener las URLs de los artículos para *scrapear*
- Ya tenemos un método para descargar las URLs de cualquier número de página
- Vamos a usarla para descargar todas las páginas desde la 1 hasta  $\infty$ , o
  - hasta que ya no haya urls, o
  - hasta que encontremos una url que ya hayamos procesado

# Creando el registro

```
existing_urls = set(get_scraped_urls())

urls_to_scrape = []
for current_page in range(1, 1_000_000):
    urls = get_urls(current_page)

    if not urls:
        break

    for url in urls:
        if url in existing_urls:
            break
        urls_to_scrape.append(url)
```

# Creando el registro

```
existing_urls = set(get_scraped_urls())

urls_to_scrape = []
for current_page in range(1, 1_000_000):
    urls = get_urls(current_page)

    if not urls:
        break

    for url in urls:
        if url in existing_urls:
            break
        urls_to_scrape.append(url)
```



# Creando el registro

```
existing_urls = set(get_scraped_urls())

urls_to_scrape = []
for current_page in range(1, 1_000_000):
    urls = get_urls(current_page)

    if not urls:
        break

    for url in urls:
        if url in existing_urls:
            break
        urls_to_scrape.append(url)
```

# Creando el registro

```
existing_urls = set(get_scraped_urls())

urls_to_scrape = []
for current_page in range(1, 1_000_000):
    urls = get_urls(current_page)

    if not urls:
        break

    for url in urls:
        if url in existing_urls:
            break
        urls_to_scrape.append(url)
```

# Creando el registro

```
existing_urls = set(get_scraped_urls())

urls_to_scrape = []
for current_page in range(1, 1_000_000):
    urls = get_urls(current_page)

    if not urls:
        break

    for url in urls:
        if url in existing_urls:
            break
        urls_to_scrape.append(url)
```

# Obteniendo URLs

Al terminar tendremos una lista de *urls* para  
*scrapear* más a detalle:

```
new_urls
```



**Descargando  
los posts**

# GOV.UK blogs

**GOV.UK**

[Blog](#)  
**BEIS Digital**



Organisations: [Department for Business, Energy & Industrial Strategy](#)

## Making learning part of the job at BEIS Digital

[Ashley Justice](#), 5 August 2022 • [Capability, People and skills](#)



**Making learning part of the job at BEIS Digital**

**About the BEIS Digital blog**

This blog is about the work of the digital, data and technology teams at the Department for Business, Energy and Industrial Strategy (BEIS).


[Find out more.](#)

**Categories**


Select Category ▾

**Work for us**

[Working in BEIS Digital.](#)

**GOV.UK**


[Blog](#)  
**Defra in the media**



Organisations: [Department for Environment, Food & Rural Affairs](#)

## Woodlands expanded for communities across England

[defrablogs](#), 5 August 2022 • [Weekly stories](#)



Tree Planting. Credit: Jill Jennings Forestry Commission.


**About Defra**


Defra in the Media is run by the Defra group Press Office and is our review of Defra group stories in the news.


The blog features a review of our leading media stories, rebuttal of inaccurate comment, and updates about our campaigns and stories.

[Find out more.](#)



**Sign up and manage updates**

 Email

 Atom

**GOV.UK**

[Blog](#)  
**Government Science and Engineering**



Organisations: [Government Science & Engineering Profession](#)

## Enabling Networks

[Lucy Hurford](#), 5 August 2022 • [Talent & Leadership](#)

Networking is a crucial skill for scientists and engineers in government, who need to collaborate widely to effectively perform public duties. However, networking can be a daunting prospect, with many factors influencing how comfortable we might feel when networking.

Lucy and Rohan joined the Fast Stream in 2021 and were slightly intimidated by the heavy emphasis on the importance of networking. Realising that they weren't alone in feeling overwhelmed at the prospect of networking, Lucy and Rohan spoke to 30 senior civil servants (SCS) working in STEM (science, technology, engineering and maths) roles from 15 departments, to draw upon their lived experiences of networking and try to level the playing field when it comes to networking skills. These conversations were used to write an article

**Become a Member**

Become a member of the GSE Profession to access our opportunities, network, events, and join the community of Scientists and Engineers in Government.

Register

**Membership Offer**

Learn more about the GSE Profession's Membership Offer by clicking on one of our information sections below.

# GOV.UK blogs



Blog

**BEIS Digital**

Organisations: [Department for Business, Energy & Industrial Strategy](#)

## Making learning part of the job at BEIS Digital

[Ashley Justins](#), 5 August 2022 • [Capability](#), [People and skills](#)



**Making learning part of the job at BEIS Digital**

**About the BEIS Digital blog**

This blog is about the work of the digital, data and technology teams at the Department for Business, Energy and Industrial Strategy (BEIS).


[Find out more.](#)

**Categories**

Select Category

**Work for us**

[Working in BEIS Digital.](#)



Blog

**for england**

Search blog

**About Defra**

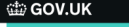
Defra in the Media is run by the Defra group Press Office and is our review of Defra group stories in the news.

The blog features a review of our leading media stories, rebuttal of inaccurate comment, and updates about our campaigns and stories.


[Find out more.](#)

**Sign up and manage updates**

[Email](#) [Atom](#)



Blog



**Government Science and Engineering**

Search blog

Organisations: [Government Science & Engineering Profession](#)

## Enabling Networks

[Lucy Hurford](#), 5 August 2022 • [Talent & Leadership](#)

Networking is a crucial skill for scientists and engineers in government, who need to collaborate widely to effectively perform public duties. However, networking can be a daunting prospect, with many factors influencing how comfortable we might feel when networking.

Lucy and Rohan joined the Fast Stream in 2021 and were slightly intimidated by the heavy emphasis on the importance of networking. Realising that they weren't alone in feeling overwhelmed at the prospect of networking, Lucy and Rohan spoke to 30 senior civil servants (SCS) working in STEM (science, technology, engineering and maths) roles from 15 departments, to draw upon their lived experiences of networking and try to level the playing field when it comes to networking skills. These conversations were used to write an article

**Become a Member**

Become a member of the GSE Profession to access our opportunities, network, events, and join the community of Scientists and Engineers in Government.

**Register**

**Membership Offer**

Learn more about the GSE Profession's Membership Offer by clicking on one of our information centres below.

# GOV.UK blogs



Blog

## BEIS Digital

Organisations: [Department for Business, Energy & Industrial Strategy](#)

**Making learning part of the job at BEIS Digital**

[Ashley Justin](#), 5 August 2022 • [Capability, People and skills](#)



**Making learning part of the job at BEIS Digital**

**About the BEIS Digital blog**

This blog is about the work of the digital, data and technology teams at the Department for Business, Energy and Industrial Strategy (BEIS).


[Find out more.](#)

**Categories**

Select Category ▾

**Work for us**

[Working in BEIS Digital.](#)




Blog

## Defra in the media

Organisations: [Department for Environment, Food & Rural Affairs](#)

**Woodlands expanded communities across England**

[defrablogs](#), 5 August 2022 • [Weekly stories](#)



Tree Planting. Credit: Jill Jennings Forestry Commission.



Blog

## Government Science and Engineering

Organisations: [Government Science & Engineering Profession](#)

**Enabling Networks**

[Lucy Hurford](#), 5 August 2022 • [Talent & Leadership](#)

Networking is a crucial skill for scientists and engineers in government, who need to collaborate widely to effectively perform public duties. However, networking can be a daunting prospect, with many factors influencing how comfortable we might feel when networking.

Lucy and Rohan joined the Fast Stream in 2021 and were slightly intimidated by the heavy emphasis on the importance of networking. Realising that they weren't alone in feeling overwhelmed at the prospect of networking, Lucy and Rohan spoke to 30 senior civil servants (SCS) working in STEM (science, technology, engineering and maths) roles from 15 departments, to draw upon their lived experiences of networking and try to level the playing field when it comes to networking skills. These conversations were used to write an article

**Become a Member**

Become a member of the GSE Profession to access our opportunities, network, events, and join the community of Scientists and Engineers in Government.

**Register**

**Membership Offer**

Learn more about the GSE Profession's Membership Offer by clicking on one of our information sections below.



Blog

# Defra in the media

Organisations: [Department for Environment, Food & Rural Affairs](#)

Search blog



## Woodlands expanded for communities across England

[defrablogs](#), 5 August 2022 - [Weekly stories](#)



Tree Planting, Credit Jill Jennings Forestry Commission.

### About Defra

Defra in the Media is run by the Defra group Press Office and is our review of Defra group stories in the news.

The blog features a review of our leading media stories, rebuttal of inaccurate comment, and updates about our campaigns and stories.

[Find out more.](#)

### Sign up and manage updates

[Email](#)

[Atom](#)

Blog

### BEIS Digital

Organisations: [Department for Business, Energy & Industrial Strategy](#)

### Making learning part of the job at BEIS Digital

[Ashley Justice](#), 5 August 2022 - [Capability, People and Skills](#)



Search blog



### Become a Member

Become a member of the GSE Profession to access our opportunities, network, events, and join the community of Scientists and Engineers in Government.

[Register](#)

### Membership Offer

Learn more about the GSE Profession's Membership Offer by clicking on one of our information centres below.

# Inspeccionando los posts

```
<article>
  <header>
    <h1>Woodlands expanded for communities across England</h1>
    <div>
      <a href="#" rel="author">defrablogs</a>,
      <time datetime="2022-08-05T16:19:06+01:00">5 August 2022</time>
      <a href="#" rel="category tag">Weekly stories</a>
    </div>
  </header>
  <div class="entry-content">
    <p>There has been widespread coverage today in the...</p>
    <p>The £44.2 million of funding will support ...</p>
  </div>
</article>
```

# Inspeccionando los posts

```
<article>
  <header>
    <h1>Woodlands expanded for communities across England</h1>
    <div>
      <a href="#" rel="author">defrablogs</a>,
      <time datetime="2022-08-05T16:19:06+01:00">5 August 2022</time>
      <a href="#" rel="category tag">Weekly stories</a>
    </div>
  </header>
  <div class="entry-content">
    <p>There has been widespread coverage today in the...</p>
    <p>The £44.2 million of funding will support ...</p>
  </div>
</article>
```

# Inspeccionando los posts

```
<article>
  <header>
    <h1>Woodlands expanded for communities across England</h1>
    <div>
      <a href="#" rel="author">defrablelogs</a>,
      <time datetime="2022-08-05T16:19:06+01:00">5 August 2022</time>
      <a href="#" rel="category tag">Weekly stories</a>
    </div>
  </header>
  <div class="entry-content">
    <p>There has been widespread coverage today in the...</p>
    <p>The £44.2 million of funding will support ...</p>
  </div>
</article>
```

# Inspeccionando los posts

```
<article>
  <header>
    <h1>Woodlands expanded for communities across England</h1>
    <div>
      <a href="#" rel="author">defrablelogs</a>,
      <time datetime="2022-08-05T16:19:06+01:00">5 August 2022</time>
      <a href="#" rel="category tag">Weekly stories</a>
    </div>
  </header>
  <div class="entry-content">
    <p>There has been widespread coverage today in the...</p>
    <p>The £44.2 million of funding will support ...</p>
  </div>
</article>
```

# Inspeccionando los posts

```
<article>
  <header>
    <h1>Woodlands expanded for communities across England</h1>
    <div>
      <a href="#" rel="author">defrablogs</a>,
      <time datetime="2022-08-05T16:19:06+01:00">5 August 2022</time>
      <a href="#" rel="category tag">Weekly stories</a>
    </div>
  </header>
  <div class="entry-content">
    <p>There has been widespread coverage today in the...</p>
    <p>The £44.2 million of funding will support ...</p>
  </div>
</article>
```

# Estructura de los posts

- Hay sitios web mejores estructurados que otros
- Debemos revisar múltiples versiones para tomar en cuenta diferentes variaciones:
  - Artículos con múltiples autores, múltiples categorías, o con contenido multimedia

# División de tareas

```
def process_header(header):
    title = header.find("h1").text
    authors = [a.text for a in header.find("div").find_all("a", {"rel": "author"})]
    cats = [a.text for a in header.find("div").find_all("a", {"rel": "category tag"})]
    time = header.find("div").find("time")["datetime"]
    header_content = {
        "title": title, "authors": authors,
        "categories": cats, "pub_date": time,
    }
    return header_content

def process_content(content):
    article_content = []
    for child in content.find_all(recursive=False):
        if not child.text: continue
        if child.name == "p":
            article_content.append({"text": child.text})
        elif child.name.startswith("h"):
            article_content.append({"heading": int(child.name[1:]), "text": child.text})
    return {"content": article_content}
```



# Descargando un solo artículo

```
def get_article(url):  
    article_response = requests.get(url)  
    article_soup = BeautifulSoup(article_response.text)  
    article = article_soup.find("article")  
  
    header = article.find("header")  
    header_content = process_header(header)  
  
    content = article.find("div", {"class": "entry-content"})  
    article_content = process_content(content)  
  
    return {"url": url, **header_content, **article_content}
```

# Descargando un solo artículo

```
def get_article(url):  
    article_response = requests.get(url)  
    article_soup = BeautifulSoup(article_response.text)  
    article = article_soup.find("article")  
  
    header = article.find("header")  
    header_content = process_header(header)  
  
    content = article.find("div", {"class": "entry-content"})  
    article_content = process_content(content)  
  
    return {"url": url, **header_content, **article_content}
```

# Descargando un solo artículo

```
def get_article(url):  
    article_response = requests.get(url)  
    article_soup = BeautifulSoup(article_response.text)  
    article = article_soup.find("article")  
  
    header = article.find("header")  
    header_content = process_header(header)  
  
    content = article.find("div", {"class": "entry-content"})  
    article_content = process_content(content)  
  
    return {"url": url, **header_content, **article_content}
```

# Descargando un solo artículo

```
def get_article(url):  
    article_response = requests.get(url)  
    article_soup = BeautifulSoup(article_response.text)  
    article = article_soup.find("article")  
  
    header = article.find("header")  
    header_content = process_header(header)  
  
    content = article.find("div", {"class": "entry-content"})  
    article_content = process_content(content)  
  
    return {"url": url, **header_content, **article_content}
```

# Descargando un solo artículo

```
def get_article(url):  
    article_response = requests.get(url)  
    article_soup = BeautifulSoup(article_response.text)  
    article = article_soup.find("article")  
  
    header = article.find("header")  
    header_content = process_header(header)  
  
    content = article.find("div", {"class": "entry-content"})  
    article_content = process_content(content)  
  
    return {"url": url, **header_content, **article_content}
```

# Resultado de llamar get\_article

```
{
  "url": "https://naturalengland.blog.gov.uk/...",
  "title": "Natural England's ...",
  "authors": ["Ginny Swaile"],
  "categories": ["Biodiversity", "Wildlife"],
  "pub_date": "2022-08-04T16:22:57+01:00",
  "content": [
    {
      "text": "By Ginny Swaile, Deputy Director Science - Sustainable land
and sea use..."
    }
  ]
}
```

# Descargando los archivos

Con el método `get_article` podemos iterar y descargar todos los posts necesarios.

Luego hay que almacenarlos en una carpeta – preferentemente siguiendo un orden:

```
./data/[YEAR]/[MONTH]/[DAY]/[NAME].json
```

# Guardando el progreso

- Debemos guardar el progreso una vez descargados los posts – DRY!
- Ya tenemos un método para hacerlo,  
`append_scrapped_urls`



# Procesamiento inicial

# Procesamiento inicial


- La primera vez que ejecutes el proceso se va a tardar
- Es necesario para que la automatización actúe de forma incremental
- Debes crear un commit con los archivos descargados en el procesamiento inicial

# 2,769 archivos

Initial data scrape


Browse files

main

 **fferegrino** committed yesterday

1 parent 4cf489f

commit caf33ee8d31d628eb0c41b0e85bbdf3474695bc2

 Showing 2,769 changed files with 208,215 additions and 1 deletion.

Split Unified



03

# AUTOMATI ZACIÓN

# GitHub Actions

- Ejecución de flujos de trabajo en servidores virtuales en la nube
- Provocados por cambios en nuestro repositorio, petición manual o bajo una agenda
- Los *workflows* se definen en un archivo *YAML* dentro del directorio *.github/workflows*

# Workflow scrape.yaml

```
name: "Daily scrape"

on:
  schedule:
    - cron: "0 19 * * *"
```

# Workflow scrape.yaml

```
name: "Daily scrape"

on:
  schedule:
    - cron: "0 19 * * *"
```

# Workflow scrape.yaml

```
jobs:
  scrape:
    steps:
      - uses: actions/checkout@v2

      - name: Set up Python 3.9
        uses: actions/setup-python@v2
        with:
          python-version: "3.9"
```



# Workflow scrape.yaml

- `name:` Install dependencies  
`run:` |  
    `pip install --upgrade pip`  
    `pip install -r requirements.txt`
- `name:` Execute scrape.py  
`run:` `python scrape.py`
- `name:` Commit changes  
`run:` |  
    `git config --global user.email "cosme.fulanito@gmail.com"`  
    `git config --global user.name "Antonio Feregrino"`  
    `git add data/ scraped_urls.txt`  
    `git diff --staged --quiet || git commit -m 'New blog posts'`  
    `git push`

# Workflow scrape.yaml

- name: Install dependencies  
run: |  
 pip install --upgrade pip  
 pip install -r requirements.txt
- **name:** Execute scrape.py  
 **run:** python scrape.py
- name: Commit changes  
run: |  
 git config --global user.email "cosme.fulanito@gmail.com"  
 git config --global user.name "Antonio Feregrino"  
 git add data/ scraped\_urls.txt  
 git diff --staged --quiet || git commit -m 'New blog posts'  
 git push

# Workflow scrape.yaml

```
- name: Install dependencies
  run: |
    pip install --upgrade pip
    pip install -r requirements.txt

- name: Execute scrape.py
  run: python scrape.py

- name: Commit changes
  run: |
    git config --global user.email "cosme.fulanito@gmail.com"
    git config --global user.name "Antonio Feregrino"
    git add data/ scraped_urls.txt
    git diff --staged --quiet || git commit -m 'New blog posts'
    git push
```

# Workflow scrape.yaml

```
- name: Install dependencies
  run: |
    pip install --upgrade pip
    pip install -r requirements.txt

- name: Execute scrape.py
  run: python scrape.py

- name: Commit changes
  run: |
    git config --global user.email "cosme.fulanito@gmail.com"
    git config --global user.name "Antonio Feregrino"
    git add data/ scraped_urls.txt
    git diff --staged --quiet || git commit -m 'New blog posts'
    git push
```

# Workflow scrape.yaml

```
- name: Install dependencies
  run: |
    pip install --upgrade pip
    pip install -r requirements.txt

- name: Execute scrape.py
  run: python scrape.py

- name: Commit changes
  run: |
    git config --global user.email "cosme.fulanito@gmail.com"
    git config --global user.name "Antonio Feregrino"
    git add data/ scraped_urls.txt
    git diff --staged --quiet || git commit -m 'New blog posts'
    git push
```

# Workflow scrape.yaml

```
- name: Install dependencies
  run: |
    pip install --upgrade pip
    pip install -r requirements.txt

- name: Execute scrape.py
  run: python scrape.py

- name: Commit changes
  run: |
    git config --global user.email "cosme.fulanito@gmail.com"
    git config --global user.name "Antonio Feregrino"
    git add data/ scraped_urls.txt
    git diff --staged --quiet || git commit -m 'New blog posts'
    git push
```

# Workflow scrape.yaml

```
- name: Install dependencies
  run: |
    pip install --upgrade pip
    pip install -r requirements.txt

- name: Execute scrape.py
  run: python scrape.py

- name: Commit changes
  run: |
    git config --global user.email "cosme.fulanito@gmail.com"
    git config --global user.name "Antonio Feregrino"
    git add data/ scraped_urls.txt
    git diff --staged --quiet || git commit -m 'New blog posts'
    git push
```

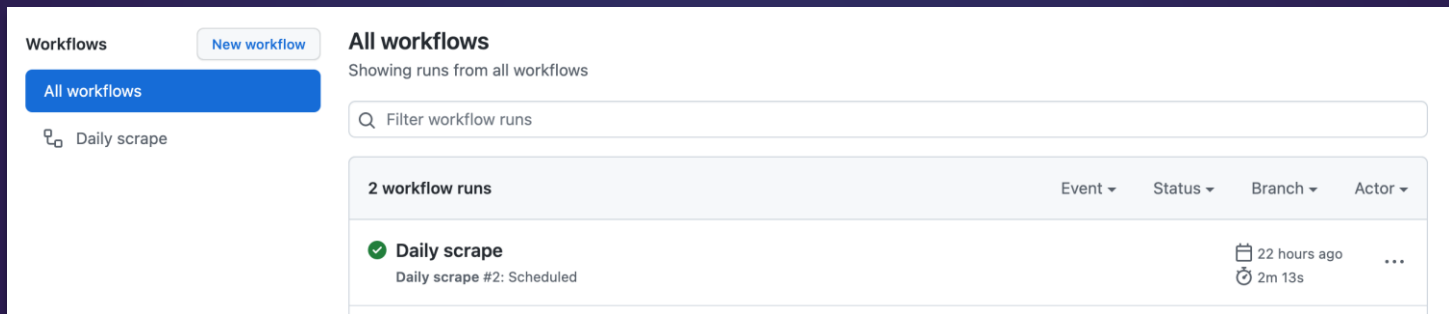
# Workflow scrape.yaml

- `name:` Install dependencies  
`run:` |  
    `pip install --upgrade pip`  
    `pip install -r requirements.txt`
- `name:` Execute scrape.py  
`run:` `python scrape.py`
- `name:` Commit changes  
`run:` |  
    `git config --global user.email "cosme.fulanito@gmail.com"`  
    `git config --global user.name "Antonio Feregrino"`  
    `git add data/ scraped_urls.txt`  
    `git diff --staged --quiet || git commit -m 'New blog posts'`  
    `git push`



# Para finalizar...

- Hacemos *commit* y *push* del archivo `scrape.yml` y...



The screenshot displays the GitHub Actions interface. On the left, the 'Workflows' sidebar shows 'All workflows' selected. The main area is titled 'All workflows' and indicates 'Showing runs from all workflows'. A search bar is present with the placeholder 'Filter workflow runs'. Below this, a table lists '2 workflow runs'. The first run, 'Daily scrape', is marked with a green checkmark and shows a status of 'Daily scrape #2: Scheduled'. It was executed '22 hours ago' and took '2m 13s' to complete. The table has columns for 'Event', 'Status', 'Branch', and 'Actor'.

Event	Status	Branch	Actor
✓	Completed	main	...

# CONSEJOS

## EXPERIMENTA



Web scraping es iterativo y requiere de experimentación

## AUTOMATIZA



Un poco de esfuerzo para que no te vuelvas a preocupar

## PRUEBA



Para encontrar errores antes de que sea tarde

## RESPETA



A los sitios web que consultes, es mejor para todos

## MODULARIZA



Es más fácil encontrar y corregir errores

## COMPARTE



Otros se pueden beneficiar de los datos

# Proyecto similar

[github.com/fferegrino/mananeras](https://github.com/fferegrino/mananeras)

[kaggle.com/datasets/ioexception/mananeras](https://kaggle.com/datasets/ioexception/mananeras)

[tcsg.dev/pycon22](https://tcsg.dev/pycon22)

# ¡GRACIAS!

¿Preguntas, comentarios? ¿quieres  
contarme qué dataset estás generando?

*Email:* antonio@feregri.no

*Asunto:* PyConLatam22



@feregri\_no en cualquier red social

/thatcsharpguy en YouTube

Antonio Feregrino wants to teach  
you some Python!