WEB
CRAWLER
Scrapy Spider

MARCOS FOLGUERA RIVERA

Código Python

```
import scrapy
import csv

class RtveSpider(scrapy.Spider):
    name = 'rtve_spider'
    start_urls = ['https://www.rtve.es/noticias/']

def parse(self, response):
    # Seleccionar las noticias usando XPath
    news = response.xpath('//a[span[@class="maintitle"]]')

# Escribir las noticias en un archivo CSV
    with open('rtve_news.csv', 'w', newline=") as file:
    writer = csv.writer(file)
    for new in news:
        title = new.xpath('span[@class="maintitle"]/text()').get()
        url = new.xpath('@href').get()
        writer.writerow([title, url])
```

```
import scrapy
import csv

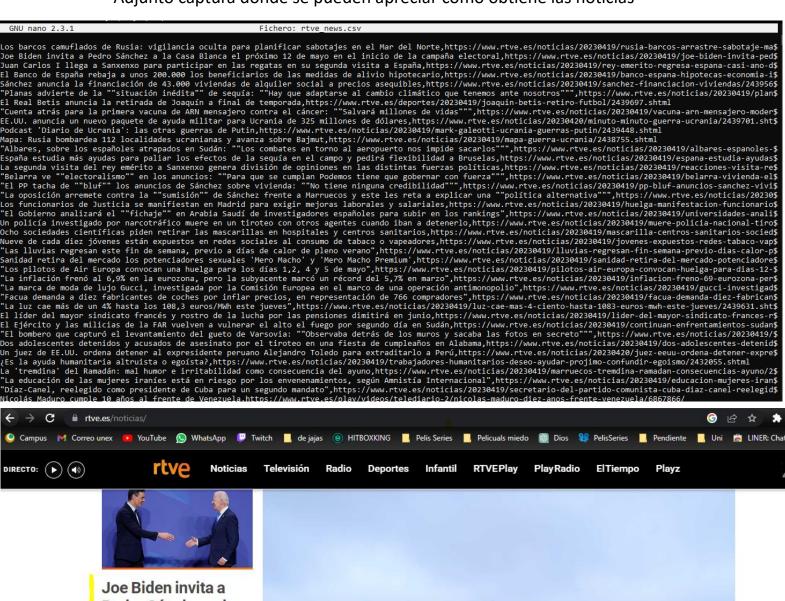
class RtveSpider(scrapy.Spider):
    name = 'rtve_spider'
    start_urls = ['https://www.rtve.es/noticias/']

def parse(self, response):
    # Seleccionar las noticias usando XPath
    news = response.xpath('//a[span[@class="maintitle"]]')

# Escribir las noticias en un archivo CSV
with open('rtve_news.csv', 'w', newline='') as file:
    writer = csv.writer(file)
    for new in news:
        title = new.xpath('span[@class="maintitle"]/text()').get()
        url = new.xpath('@href').get()
        writer.writerow([title, url])
```

Nuestro programa recopila los títulos y su url de las noticias de rtve.es/noticias y los almacena en un fichero .csv

Adjunto captura donde se pueden apreciar como obtiene las noticias



Joe Biden invita a Pedro Sánchez a la Casa Blanca el próximo 12 de mayo en el inicio de la campaña electoral



Juan Carlos I llega a Sanxenxo para participar en las regatas en su



Los barcos camuflados de Rusia: vigilancia oculta para planificar sabotajes en el Mar del Norte

Una investigación de cuatro cadenas públicas nórdicas revela actividades de vigilancia encubiertas de Rusia.

Webgrafía:

https://docs.scrapy.org/en/latest/intro/tutorial.html

https://chat.openai.com/

https://elmundodelosdatos.com/extraccion-datos-sitios-web-productos-zara/