

Tipología y ciclo de vida de los datos

Práctica 1

Autor: Antonio Arencibia Guerra (aarenc)

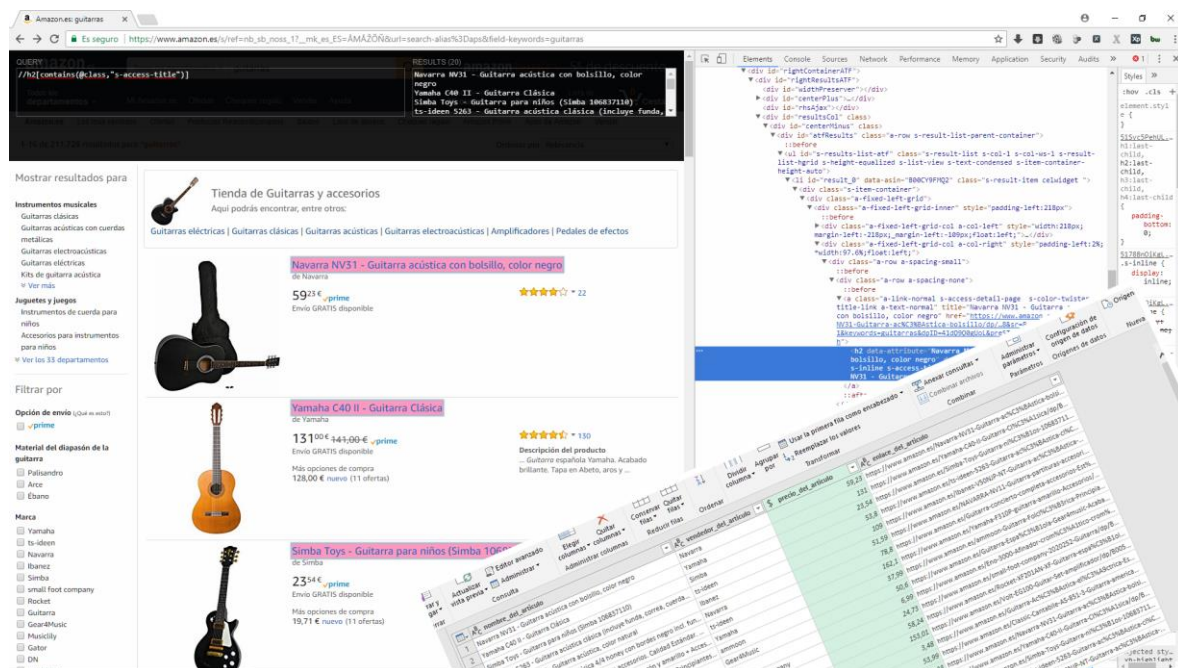
1. Título del dataset. Poned un título que sea descriptivo.

Guitarras & Accesorios Amazon España

2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.

Extracción de artículos basados en una keyword sobre un e-commerce. En este caso el dataset consta de guitarras y accesorios para guitarras actuales que podemos encontrar en Amazon España.

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente



4. Contexto. ¿Cuál es la materia del conjunto de datos?

El seguimiento o reutilización de los datos de una plataforma e-commerce para nuevas aplicaciones o servicios. Siempre teniendo en cuenta que no perjudique al propietario de los datos, en este caso Amazon.

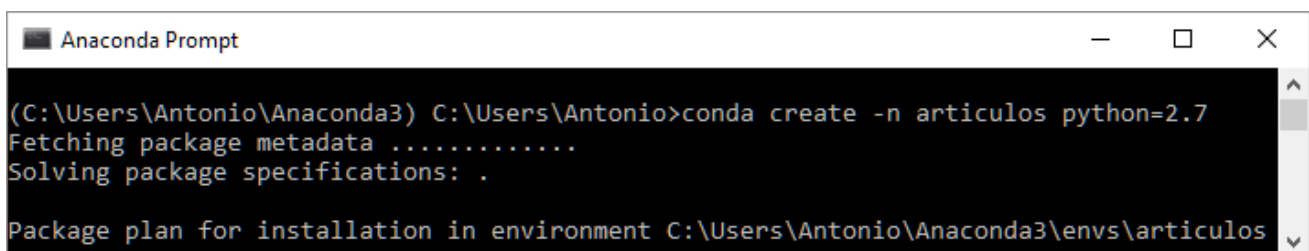
5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?

El conjunto de datos esta formado por un conjunto de articulos, filtrados en este caso mediante la keyword “guitarras” previamente, en el E-commerce “Amazon”. Este ejemplo de Web Scraping se almacenan los campos del nombre del artículo, el nombre del vendedor, el precio del artículo y el enlace al propio articulo en Amazon junto al enlace a su imagen relacionada.

Para la extracción de los datos se ha realizado un proyecto con la librería Python para web scraping Scrapy. Se realiza la extracción vertical de la página donde se inicia la ejecución para despues ir realizando lo mismo en la siguiente página automáticamente, y así por todas las primeras 5 páginas de resultados que se muestren. 5 páginas es el limite que le he puesto al scraper para el scraping horizontal y se puede ampliar o reducir dependiendo de lo requerido.

El proceso para la construcción del proyecto web scrapper para la extracción de articulos de Amazon han sido:

- Estudio previo
 - o Valorar las diferentes alternativas para la realización de la práctica.
 - Se decide realizarlo con Python y su librería Scrapy.
- Preparacion del entorno de trabajo
 - o Instalación de Anaconda para el desarrollo sobre Python
 - o Creación del entorno virtual en Anaconda para la realización del proyecto.
 - Debido a compatibilidades se crea el entorno virtual con python 2.7

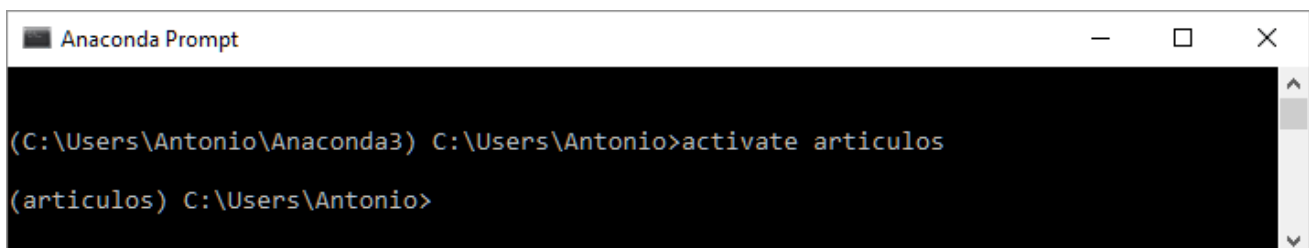


```

Anaconda Prompt
(C:\Users\Antonio\Anaconda3) C:\Users\Antonio>conda create -n articulos python=2.7
Fetching package metadata .....
Solving package specifications: .

Package plan for installation in environment C:\Users\Antonio\Anaconda3\envs\articulos
  
```

- o Activación el entorno virtual



```

Anaconda Prompt
(C:\Users\Antonio\Anaconda3) C:\Users\Antonio>activate articulos
(articulos) C:\Users\Antonio>
  
```

- Instalación del paquete Scrapy en el entorno virtual

```
Anaconda Prompt

(C:\Users\Antonio\Anaconda3) C:\Users\Antonio>activate articulos
(articulos) C:\Users\Antonio>conda install -c conda-forge scrapy_
```

- Creación del proyecto Scrapy

```
Anaconda Prompt

attrs-17.3.0-p 100% |#####| Time: 0:00:00 188.10 kB/s
(articulos) C:\Users\Antonio>scrapy startproject articuloscrawl
```

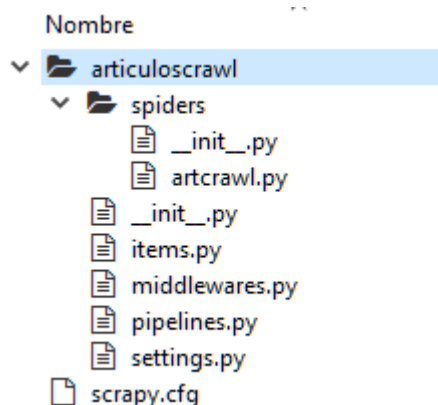
- Proyecto en Python con Scrapy.

- Creación del spider dentro del proyecto Scrapy. Para ello nos movemos hasta el directorio raíz del proyecto.

```
Anaconda Prompt

(articulos) C:\Users\Antonio>cd articuloscrawl
(articulos) C:\Users\Antonio\articuloscrawl>cd articuloscrawl
(articulos) C:\Users\Antonio\articuloscrawl\articuloscrawl>scrapy genspider artcrawl www.amazon.es
```

- Comprobamos la estructura del proyecto y el spider recién creado



- scrapy.cfg: archivo de configuración del proyecto.
- ./: módulo python de nuestro proyecto, después incluiremos aquí nuestro código.
- ./items.py: archivo donde definimos los items que queremos extraer.
- ./pipelines.py: definimos los pipelines o flujos del proyecto.
- ./settings.py: archivo de ajustes del proyecto.
- ./spiders/: directorio donde luego pondremos nuestros spiders

- Análisis previo de la web de origen de los datos en Scrapy shell

```
Anaconda Prompt - scrapy shell

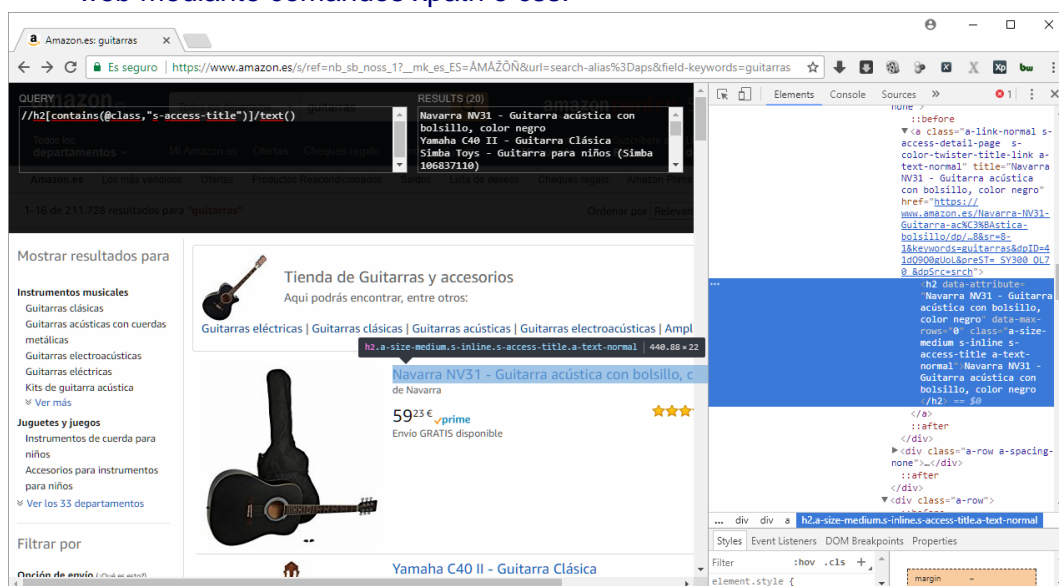
owed)
[s] fetch(req)          Fetch a scrapy.Request and update local objects
[s] shell()             Shell help (print this help)
[s] view(response)      View response in a browser
>>> fetch("https://www.amazon.es/s/ref=nb_sb_noss_1?__mk_es_ES=%C3%85M%C3%85C5BD%C3%95C3%91&url=search-alias%3Daps&field-keywords=guitarras")
```

- Con view(response) en el shell de Scrapy obtenemos una copia de la weben local para su estudio. Se abre automáticamente en el navegador.

```
Anaconda Prompt - scrapy shell

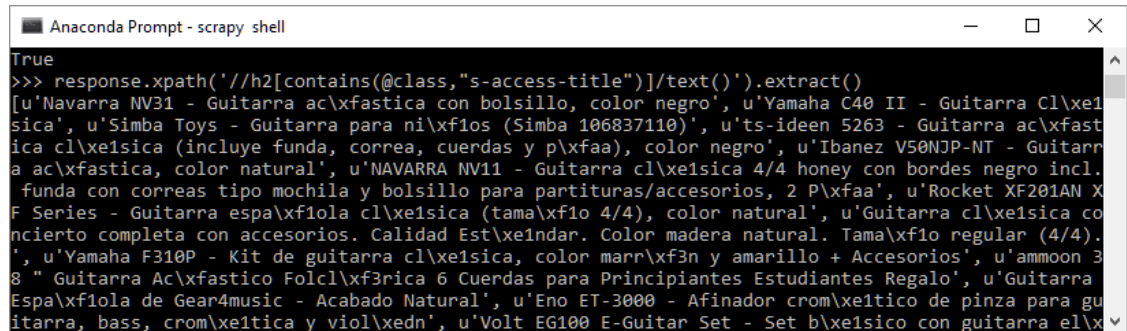
s=guitarras>
2017-11-13 20:04:17 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.amazon.es/guitarras/s?ie=UTF8&page=1&rh=1%3Aaps%2Ck%3Aguitarras> (referer: None)
>>> view(response)
True
>>>
```

- Mediante el Plugin Xpath Helper se pueden analizar como capturar los datos de la web mediante comandos xpath o css.



- En el shell de Scrapy podemos probar dichos comandos para probar su respuesta. Por ejemplo: Vamos a seleccionar todos los nombres de los artículos mostrados en la primera página.

```
response.xpath('//h2[contains(@class,"s-access-title")]/text()').extract()
```



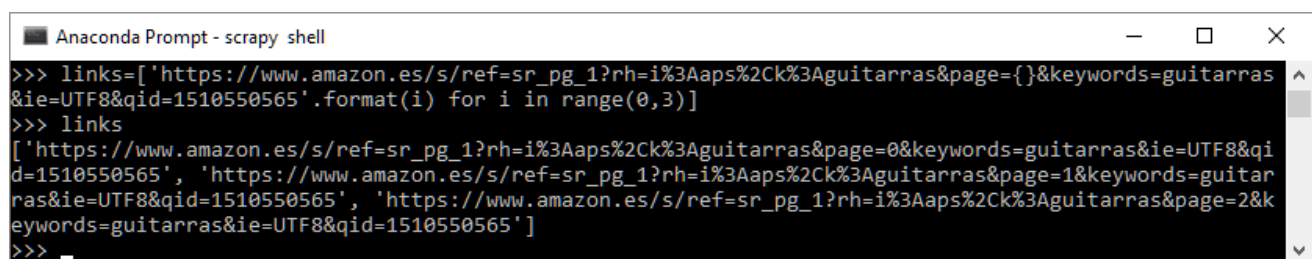
```
Anaconda Prompt - scrapy shell
True
>>> response.xpath('//h2[contains(@class,"s-access-title")]/text()').extract()
[u'Navarra NV31 - Guitarra ac\xfastica con bolsillo, color negro', u'Yamaha C40 II - Guitarra Cl\xeisica', u'Simba Toys - Guitarra para ni\xf1os (Simba 106837110)', u'ts-ideen 5263 - Guitarra ac\xfastica cl\xeisica (incluye funda, correa, cuerdas y p\xfaa), color negro', u'Ibanez V50NP-NT - Guitarra ac\xfastica, color natural', u'NAVARRA NV11 - Guitarra cl\xeisica 4/4 honey con bordes negro incl. funda con correas tipo mochila y bolsillo para partituras/accesorios, 2 P\xfaa', u'Rocket XF201AN X Series - Guitarra espa\xf1ola cl\xeisica (tama\xf1o 4/4), color natural', u'Guitarra cl\xeisica concierto completa con accesorios. Calidad Est\xandar. Color madera natural. Tama\xf1o regular (4/4).', u'Yamaha F310P - Kit de guitarra cl\xeisica, color marr\xf3n y amarillo + Accesorios', u'ammoon 38" Guitarra Ac\xfastico Folcl\xf3rica 6 Cuerdas para Principiantes Estudiantes Regalo', u'Guitarra Espa\xf1ola de Gear4music - Acabado Natural', u'Eno ET-3000 - Afinador crom\xeico de pinza para guitarra, bass, crom\xeico y viol\xeddn', u'Volt EG100 E-Guitar Set - Set b\xeisico con guitarra el\xectrica']
```

- Proceso de desarrollo del crawler con crawl vertical y horizontal.

Con la intención de realizar un crawl vertical en la primera página y luego pasar a la siguiente página de las muestras, crawl horizontal, realizar de nuevo crawling vertical y así sucesivamente tenemos que estudiar la url de la página de origen y encontrar un patrón que nos sirva para que las recorra nuestro crawler. Esto lo podemos analizar también en el shell de scrapy.

```
['https://www.amazon.es/s/ref=sr_pg_1?rh=i%3Aaps%2Ck%3Aguitarras&page={}&keywords=guitarras&ie=UTF8&qid=1510550565'.format(i) for i in range(0,3)]
```

El valor que está entre corchetes en rojo nos dará el número de paginación. Lo chequeamos en el shell.



```
Anaconda Prompt - scrapy shell
>>> links=['https://www.amazon.es/s/ref=sr_pg_1?rh=i%3Aaps%2Ck%3Aguitarras&page={}&keywords=guitarras&ie=UTF8&qid=1510550565'.format(i) for i in range(0,3)]
>>> links
['https://www.amazon.es/s/ref=sr_pg_1?rh=i%3Aaps%2Ck%3Aguitarras&page=0&keywords=guitarras&ie=UTF8&qid=1510550565', 'https://www.amazon.es/s/ref=sr_pg_1?rh=i%3Aaps%2Ck%3Aguitarras&page=1&keywords=guitarras&ie=UTF8&qid=1510550565', 'https://www.amazon.es/s/ref=sr_pg_1?rh=i%3Aaps%2Ck%3Aguitarras&page=2&keywords=guitarras&ie=UTF8&qid=1510550565']
>>>
```

Esta instrucción alimentará a la variable “start_urls”, fundamental en nuestro spider..

Ejecutaremos nuestro spider con salida csv con la instrucción:

```
Scrapy crawl artcrawl -o guitarras_amazon.csv
```

Cambiando csv por json obtendríamos un fichero de salida json.

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

El propietario de los datos es Amazon. Los datos que existen en Amazon estan siendo continuamente investigados y analizados por lo que hay innumerables analisis previos.

Un análisis previo claro puede ser el que hacen, sobre los comentarios de sus productos publicados en Amazon, las marcas comerciales. Para ellos es un feedback de incalculable valor y muy sensible para las ventas futuras.

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

He visto que se realizan continuamente extracciones de datos de tiendas o empresas online mediante técnicas de Web Scraping para el control de la competencia, ajuste de precios, reconocer demandas, como feedback o simplemente por casos de estudio. Todas ellas son legítimas pero hecho mi visión del potencial de los datos es otro, quizás más en el sentido de utilizar dichos datos para un bien comun en vez de para el lucro o beneficio propio. El objetivo que planteo es utilizar estos datos extraidos para crear un valor añadido que resulte rentable tanto al propietario de los datos, en este caso Amazon, el que extrae los datos y al cliente final. Existen numerosas formas de crear valor a partir de los datos y que, de esta manera , pueden resultar útiles.

Crear valor a partir de los datos, esto es cierto para cualquier dataset. Con una cantidad de datos inmensa contenida en Amazon resulta evidente la cantidad de respuestas que se pueden resolver a traves del análisis de los mismos, creando así, una capa más entre la tienda y el cliente. Ejemplos de uso de los datos podrian ser usarlos, dependiendo del keyword utilizado y otros filtros para la página origen, para crear una aplicación móvil, orientada para personas con dificultad visual, que pudiera explicar los productos seleccionados de manera que el cliente pueda informarse de ellos e incluso poder encargarlos mediante comandos de voz preestablecidos.

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:

- Released Under CC0: Public Domain License



He elegido esta licencia Creative Commons debido a que deseo que el ejemplo de web scraping que se ha desarrollado sea de dominio público y a que no deseo ningún derecho sobre la propiedad intelectual sobre el, liberándome así de todos ellos.

Las licencias Creative Commons informan a terceras personas de los derechos de uso de los productos elaborados. Este tipo de licencia, Released Under CC0: Public Domain License, permite copiar, modificar, distribuir la obra y hacer comunicación pública, incluso para fines comerciales sin la necesidad de pedir permiso ni de reconocer la propiedad o autoría de la obra original. Es la opción más abierta de todas las que ofrece el modelo Creative Commons. En este modelo ni siquiera se requiere reconocer la autoría de la obra original.

A continuación se describen el resto de licencias propuestas:

- Released Under CC BY-NC-SA 4.0 License

- **Atribución – No Comercial – Compartir Igual**

Esta licencia permite a otros re-mezclar, extraer fragmentos y construir a partir de tu obra, sin fines de lucro, siempre y cuando te den crédito por la creación original e inscriban sus nuevas creaciones bajo una licencia en los mismos términos.

- Released Under CC BY-SA 4.0 License

- **Atribución-Compartir Igual**

Esta licencia permite a otros distribuir, re-mezclar, extraer fragmentos y construir a partir de tu obra, incluso comercialmente, siempre y cuando te den crédito por la creación original e inscriban las nuevas obras en una licencia con los mismos términos.

- Database released under Open Database License

- **Atribucion licence ODC-ODBL**

Se puede compartir, copiar, distribuir y usar la base de datos. También crear a partir de la base de datos original así como adaptar modificar, transformar y construir sobre la base de datos original. Siempre compartiendo igual y manteniendo la licencia abierta.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset

El acceso al mismo en Github es:

https://github.com/aarenc/P1_TyCdVdID

10. Dataset: Dataset en formato CSV

Se adjunta el Dataset generado por el crawler en su última ejecución.