UOC – UNIVERSITAT OBERTA CATALUNYA
Asignatura: Tipología y ciclo de vida de los datos
Actividad: Práctica 1

Alumno: José Cortés Novales

Contexto

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

La información se saca de una de las páginas web de venta de vinos nacional, llamada EnCopa de Balón, cuya url es https://www.encopadebalon.com/. La información a extraer son los productos y precios de los vinos de Rioja (aunque podría haber sido los de otra zona o denominación como el cava), debido a que parte de mi trabajo transcurre como controller de una empresa donde uno de sus negocios significativos es la venta de vino. De esta manera se pueden obtener los precios de venta de los productos de la competencia en uno de los portales de venta más conocidos de la ciudad de Madrid.

Como comentario adicional, mencionar que he explorado otros lugares de venta de vinos de otras página web y las estructuras son parecidas a la encontrada. Son marcos en los que se añade la imagen del producto, más descripciones que pueden comprender nombre, precio y otras características del producto, incluyendo pequeñas descripciones.

Definir un título para el dataset

Elegir un título que sea descriptivo.

El título elegido para el dataset es PreciosVinoRioja, puesto que vamos a sacar los datos de los productos/vinos que se venden en este portal web de esta Denominación de Origen de la Rioja, y aunque, tal vez, sería necesario añadir el nombre de la tienda al título para que sea más descriptivo, podremos añadir este detalle a la descripción que se haga de los datos/repositorio de datos.

Descripción del dataset

Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

Los datos recopilados recogen la descripción de los productos de la Denominación Calificada Rioja que se venden en la página Encopadebalon.com, durante el momento de la extracción. Y junto al nombre, se incluye el dato de la añada o la calidad, si están incluidos en la descripción

del producto. Además, se recogen los precios de estos productos, como medio para hacer comparativas entre productos de esa región y entre vinos de las mismas características de crianza.

Representación gráfica

Presentar una imagen o esquema que identifique el dataset visualmente

La idea sería conseguir un fichero .csv que contuviera una columna con nombres y otra con precios, de forma que quedase algo parecido a la imagen que se adjunta,

	A	B
1	nombres	precios
2	Familia Comenge 2016	12,60 €
3	Figuero 4 2019	51,50 €
4	Artazu Pasos de San Martín 2016	11,80 €
5	Valduero Una Cepa 2016	18,85 €
6	Ramón Bilbao Crianza 2017	8,80 €
7	Tilenus La Florida 2015	89 €

Contenido

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

Aunque ya se ha explicado anteriormente, los campos que incluye el dataset son el nombre, que incluye otras características adicionales, y los precios de los productos de Rioja, que es la denominación que vende más vinos en España y una de las más antiguas en cuanto a tradición y producción.

El momento en el que se extraen los datos es el que se ejecuta el código del programa. Si la estructura de la página no cambia, podría servir para extraer los datos en cualquier momento y hacer comparativos entre ellos, para ver como cambian, o para analizar los mismos.

La recogida se realiza mediante la técnica del web scraping.

Agradecimientos

Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

El propietario de los datos es la empresa dueña de la página web de los que se obtienen:

“INFORMACIÓN

Distribuidora de Bebidas Europa S.A., Calle Trueno, Nº78, C.P. 28918, Leganés, Madrid, España | atencionalcliente@encopadebalon.com Aprender De Vinos

INFORMACIÓN SOBRE LA TIENDA

- Encuentra vinos y licores al mejor precio o disfruta de las mejores especialidades en nuestros restaurantes de Madrid
- Llámanos ahora: +34 91 612 74 64
- Email: Atencionalcliente@Encopadebalon.Com"

Y respecto a las cookies, suele aparecer en el código de la página y en la página web lo siguiente:

<td class="td_provider">Propias</td>
<td class="td_description">
Estas cookies son propias y nos ayudan a guardar la configuración del usuario para el correcto funcionamiento de la web.
Estas cookies son propias y nos ayudan a guardar la configuración del usuario para el correcto funcionamiento de la web.
</td>

Inspiración

Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

En mi caso es interesante, porque como he comentado, trabajo en una bodega que también distribuye vinos y que tiene un club privado, con lo cual se pueden comparar datos de las marcas competidoras y ver como están posicionadas en ele mercado y, si se hace un seguimiento, ver como van cambiando los mismos y ver que ofertas se presentan en función del tiempo y las circunstancias, por ejemplo, ver si en Navidades hay ofertas especiales o si en otras épocas del año, por ejemplo, en verano, hay otro tipo de ofertas.

También ver si se promocionan más un tipo de vino que se vende en un tipo de establecimientos, por ejemplo, CVNE o Cáceres, propios de bodegas grandes, frente a otros de bodegas minoritarias que se venden en hostelería. En ese sentido el operador del cual se extraen los datos es interesante, puesto que hace venta directa a cliente, pero también tiene un restaurante propio en el que venden todos esos vinos. Y los precios, tanto en uno, como en otro canal, son bastante competitivos, muchas veces, incluso mejores que los que se ofrecen al público en grandes cadenas, que se supone que tienen un poder de negociación mejor con el proveedor.

Para la inspiración no he recurrido a nadie. Es algo que ya había pensado en desarrollar dentro de mis funciones en el trabajo, aunque de una forma más manual, así que ha sido una gran sorpresa encontrarse con otra forma de obtener los datos de otra forma más automatizada.

Licencia

Seleccione una de estas licencias para su dataset y explique el motivo de su selección:
○ Released Under CC0: Public Domain License
○ Released Under CC BY-NC-SA 4.0 License
○ Released Under CC BY-SA 4.0 License
○ Database released under Open Database License, individual contents under Database Contents License
○ Other (specified above) ○ Unknown License

Les he echado un vistazo a todas, aunque no las conocía, pese a haber visto los símbolos que las identifican en muchas ocasiones. Por ejemplo,

Tendría que investigar un poco más para ver cuál sería la más adecuada según el caso, aunque muchas parecen compatibles y que pueden ser aplicables al caso.

Al ser un database, de una primera opinión, y con la información que tengo actualmente, parece que la más indicada podría ser la específica de los Database, es decir, la cuarta opción, que siempre recogerá aspectos más relacionados con el campo en el que nos estamos moviendo.

Código

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

He intentado diferentes alternativas de código, sin encontrar la que me satisficiera. Como ya comenté con el profesor de la asignatura, no tenía conocimientos previos de web scrapping y mis conocimientos de Python eran bastante básicos, así que eso, unido a la falta de tiempo, han supuesto pequeñas dificultades a la hora de realizar el código de la práctica.

De los libros recomendamos en la bibliografía, he podido conseguir el de Lawson, aunque me ha servido más consultar vídeos en internet, donde se dan pequeñas nociones de web scraping, para poder intentar pequeñas propuestas de resolución. Uno de ellos ha sido el que hacía un scrapy https://www.youtube.com/watch?reload=9&v=ViOFqeRgu5s&feature=emb_title , con el cual he podido intentar, sin éxito el código

```
import scrapy
from scrapy.item import item, Field
from scrapy.spiders import CrawlSpider
from scrapy.spiders import Rule
from scrapy.linkextractors import LinkExtractor
from scrapy.loader.processors import Join
from bs4 import BeatifulSoup

#Definimos la clase y los campos que queremos guardar de cada uno de
```

```
los artículos que son objetos del tipo Field
class EnCopaBalonItem(Item):
    nombre = Field()
    precio = Field()

#Definimos la clase del CrawlSpider
class EnCopaBalonCrawler(CrawlSpider):
    name = 'encopabaloncrawler'
    allowed_domains = ['www.encopadebalon.com']
    start_urls = ['https://www.encopadebalon.com/es/32-rioja']

#Montamos tupla con la variable Rules para las decisiones
rules (
    Rule(LinkExtractor(allow = r'/32-rioja#/page-\d+'), follow=True),
    Rule(LinkExtractor(allow = r'/es/32-rioja'), follow=True,
callback='parse_items')
)

def parse_items(self, response):
        item = scrapy.loader.ItemLoader (EnCopaBalonItem(), response)
        item.add_xpath('nombre', '//h5/title()')

        soup = BeautifulSoup(response.body)
        price = soup.find(id="price")
        precio = price.text
        item.add_value('nombre', nombre)

        yield item.load_item()
```

No obstante, lo que he empleado es beautifulSoup y, aunque no he conseguido limpiar los campos extraídos, si creo que iba en el buen camino,

```
import requests
import pandas as pd

from bs4 import BeautifulSoup
url = 'https://www.encopadebalon.com/es/32-rioja'
html = requests.get(url)
soup = BeautifulSoup(html.content, 'html.parser')

for tag in soup.findAll('h5'):
    nombre = soup.find_all('a', class_='product-name')
    nombres = list()
    for i in nombre:
            nombres.append(i.text)
    print(nombres)
precio = soup.find_all('span', class_='price product-price')
precios = list()
for i in precio:
        precios.append(i.text)
print(precios)

df = pd.DataFrame(('nombres' ,'precios'), index=list())
df.drop_duplicates()
print(df)

df.to_csv('PreciosVionoRioja.csv', index=False)
```

Dataset

Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

Me ha dado tiempo a ver lo que era el DOI https://www.doi.org/, pero ya no me ha dado tiempo a aplicarlo.