Asignatura: M2.851 – Tipología y ciclo de vida de los datos

Práctica: Práctica 1

Alumno: Lorenzo Mesa Morales

Fecha: 12/11/2018

Contenido

Enlace a Github

https://github.com/lmesamo/web-scraping

Pregunta 1

Título del dataset. Poned un título que sea descriptivo.

Recetas para todos.

Pregunta 2

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

Recopilación de recetas para todos los gustos y sabores.

Pregunta 3

Imagen. Agregad una imagen que identifique vuestro dataset visualmente

Pregunta 4

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

Este conjunto de datos contiene una recopilación de recetas estructuradas según la información correspondiente sobre alérgenos, calorías e ingredientes con sus cantidades.

Pregunta 5

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

Los campos que incluye son nombre de la receta, autor, nº de personas para las que se expresan las cantidades, alérgenos, calorías y la lista de ingredientes junto a sus cantidades. El periodo de tiempo es toda la historia d ela web y se han recogido mediante su página web https://www.saboresdehoy.com/

Pregunta 6

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

El propietario de los datos es la empresa RESERVIA INTERNET SL según podemos ver en https://www.saboresdehoy.com/aviso-legal. Agradecemos el trabajo realizado durante este tiempo en recopilar todas estas recetas así como estructurar la información correspondiente a calorías y alérgenos.

Pregunta 7

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

Este conjunto de datos puede ser interesante en diferentes ámbitos:

- Nutrición: puede ser una fuente de información interesante para nutricionistas o profesionales del área que quieran obtener recetas con determinadas características para sus pacientes (nº de calorías, ingredientes, alérgenos).
- Usuario normal: puede ser una fuente de información atractiva para personas que deseen buscar recetas en función de los ingredientes de los que dispongan en ese momento en casa.
- Personas con alergias o intolerancias: puede ser una fuente de información importante para personas que buscan recetas que no incluyan determinados alérgenos.

Entre las preguntas podríamos incluir:

- Análisis de recetas por calorías incluyendo ingredientes que son comunes a los platos más calóricos.
- Análisis de ingredientes comunes en alérgenos de modo que se pueda inferir qué ingredientes suelen estar vinculados a determinadas alergias.

Pregunta 8

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

- **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**

Seleccionamos la licencia Released Under CC BY-SA 4.0 License para cumplir con lo expresado en el apartado 3. CONDICIONES DE ACCESO Y UTILIZACIÓN de la web https://www.saboresdehoy.com/aviso-legal.

Con esta licencia se debe cumplir que:

- El beneficiario de la licencia tiene el derecho de copiar, distribuir, exhibir y representar la obra y hacer obras derivadas siempre y cuando reconozca y cite la obra de la forma especificada por el autor o el licenciante.
- El beneficiario de la licencia tiene el derecho de distribuir obras derivadas bajo una licencia idéntica a la licencia que regula la obra original.

Pregunta 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.

Disponible en

https://github.com/lmesamo/web-scraping/blob/master/src/read_recipes_new.py

```
from bs4 import BeautifulSoup
from urllib.request import urlopen
import re

from selenium import webdriver
from selenium.common.exceptions import NoSuchElementException
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
```

```
import time


import pandas


PAGE_URL = "https://www.saboresdehoy.com/recetas"


#Función para obtener todos los links de recetas

#Utilizamos selenium para poder simular el click en el botón de Más recetas

#De esta manera nos aseguramos de que se muestran todas las recetas
existentes

def getLinks(url):

	PATIENCE_TIME = 60

	driver = webdriver.Chrome('./chromedriver.exe')

	driver.get(PAGE_URL)

	driver.maximize_window()


	loadMoreButton =
driver.find_element_by_xpath("//div[@id='barracookies']/a[1]")

	time.sleep(2)

	loadMoreButton.click()


	while True:

		try:

			loadMoreButton = driver.find_element_by_id("loadMore")

			time.sleep(2)

			loadMoreButton.click()

			time.sleep(5)

		except Exception as e:

			print(e)

			break


	time.sleep(10)


	links = []
```

```
	links_elements = 
driver.find_elements_by_xpath("//a[contains(@href,'recetas/')]")


	for link in links_elements:
		links.append(link.get_attribute("href"))


	driver.quit()


	return list(set(links))


#Función para obtener los datos de la receta que vamos a incluir en nuestro 
dataset
#Utilizamos BeautifulSoup para extraer la información
def getRecipe(url):
	page = urlopen(url)
	soup = BeautifulSoup(page, "html.parser")


	nombre_tag = soup.find("span",attrs={"class": "titulo"})
	nombre = nombre_tag.text.replace(","," ")


	autor_tag = soup.find("span",attrs={"itemprop": "name"})
	autor = autor_tag['content'].replace(","," ")


	personas_tag = soup.find("span",attrs={"itemprop": "recipeYield"})
	personas = personas_tag.text


	calorias_tag = soup.find("div",attrs={"class": "textocalorias"})
	calorias = calorias_tag.text


	image_tag = soup.find("meta",attrs={"itemprop": "url"})
	image = image_tag['content']


	image_width_tag = soup.find("meta",attrs={"itemprop": "width"})
	image_width = image_width_tag['content']
```

```
	image_height_tag = soup.find("meta",attrs={"itemprop": "height"})
	image_height = image_height_tag['content']


	alergenos = ""
	i = 0


	div_alergenos = soup.find("div",attrs={"id":
"alergenos"}).find_all("img")


	for child in div_alergenos:
		if (i == 0):
			alergenos = child.get('title')
		else:
			alergenos = alergenos + "|" + child.get('title')


		i = i + 1


	listado_ingredientes = soup.find_all("span",attrs={"itemprop":
"ingredients"})


	filas = []
	for child in listado_ingredientes:
		txt_ingrediente = child.text
		campos = child.text.split(' ', maxsplit=2)
		cantidad = campos[0]
		unidad = campos[1]
		ingrediente = campos[2]



	filas.append([url,nombre,autor,personas,alergenos,calorias,ingrediente
,cantidad,unidad,image,image_width,image_height])


	return filas


#Obtenemos todos los links
```

```
url_recetas = getLinks(PAGE_URL)

#Guardamos en un dataset todos los registros de cada una de las recetas
dataset = []
for urls in url_recetas:
	dataset = dataset + getRecipe(urls)

#Utilizamos pandas para exportar a csv los registros del dataset
pd = pandas.DataFrame(dataset)
pd.to_csv('recipes.csv',index=False,encoding='iso-8859-1')
```

Pregunta 10

Dataset: Dataset en formato CSV

Disponible en:

https://github.com/lmesamo/web-scraping/blob/master/src/recipes.csv

Bibliografía

Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

Masip, D. (2010). El lenguaje Python. Editorial UOC.

Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

Tutorial de Github https://guides.github.com/activities/hello-world.

https://es.wikipedia.org/wiki/Licencias_Creative_Commons

https://stackoverflow.com/questions/39112138/use-selenium-to-click-a-load-more-button-until-it-doesnt-exist-youtube

https://selenium-python.readthedocs.io/locating-elements.html

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

https://www.saboresdehoy.com/robots.txt

https://www.saboresdehoy.com/sitemap.xml

https://www.saboresdehoy.com/aviso-legal