Práctica 1 Agustín Torres Nieto

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes por un proyecto analítico y usar las herramientas de extracción de datos. Para hacer esta práctica tendréis que trabajar en grupos de 3 o 2 personas, o si preferís, también podéis hacerlo de manera individual. Tendréis que entregar un solo fichero con el enlace Github (https://github.com) donde haya las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos de vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Podéis mirar ejemplos de Kaggle como referencia de vuestra práctica, como por ejemplo este https://www.kaggle.com/natt77/rare-diseases-on-facebook-groups.

Descripción de la Práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos al web. Tenéis que indicar las siguientes características del dataset general:

1. Título del dataset. Poned un título que sea descriptivo.
2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.
3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente
4. Contexto. ¿Cuál es la materia del conjunto de datos?
5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?
6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.
7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

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
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
10. Dataset: Dataset en formato CSV

Recursos

Los siguientes recursos son de utilidad por la realización de la PEC:

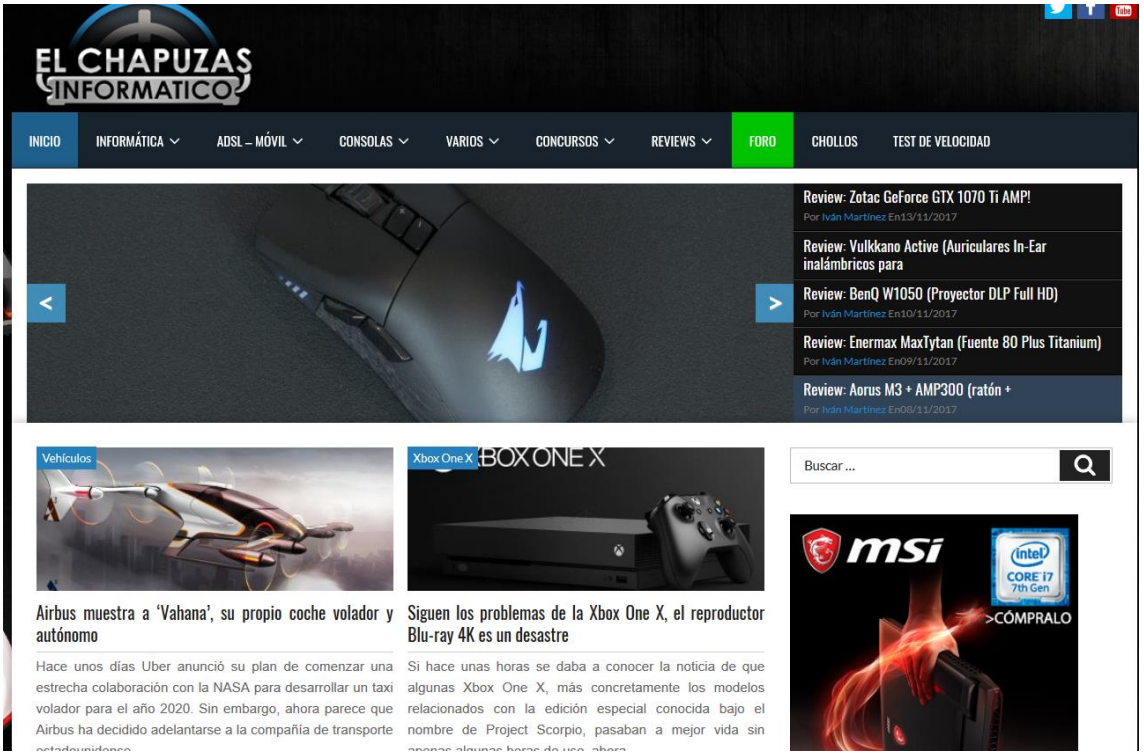
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- Los apartados 1, 2, 3 y 4 valen 0,25 puntos cada uno.
- Los apartados 5, 6, 7, 8 valen 1 punto cada uno.
- Los apartados 9 y 10 valen 2,5 puntos cada uno.

Respuestas

1. Título: “Entradas del blog El chapuza informático”
2. Subtitulo: Recolección de las entradas título, autor y fecha de las 20 páginas más recientes del blog.
3. 

4. La materia son las entradas que se van realizando en el blog.
5. Los campos que incluyen son el título, el autor y la fecha de publicación de las entradas. La recolección de datos se realizó el día 13 de noviembre de 2017 a las 21:02 de las entradas que había en ese momento en las 20 páginas más recientes.
6. Agradecimientos al blog https://elchapuzasinformatico.com/, ya que es donde he sacado los datos.
7. Este conjunto de datos es interesante porque habla de temas como la informática, ciencia, tecnología…

8. La licencia utilizada es la misma que en la página web: Reconocimiento-NoComercial-SinObraDerivada 3.0 España (CC BY-NC-ND 3.0 ES)

Usted es libre de:

Compartir — copiar y redistribuir el material en cualquier medio o formato

El licenciador no puede revocar estas libertades mientras cumpla con los términos de la licencia.

Bajo las condiciones siguientes:

Reconocimiento — Debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios<. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.

NoComercial — No puede utilizar el material para una finalidad comercial.

SinObraDerivada — Si remezcla, transforma o crea a partir del material, no puede difundir el material modificado.

No hay restricciones adicionales — No puede aplicar términos legales o medidas tecnológicas que legalmente restrinjan realizar aquello que la licencia permite.

9.

```
#Agustin Torres Nieto
#2017/11/13
#Cabecera
import csv
import requests
from lxml import html
from bs4 import BeautifulSoup

# Definimos la página web
page_inicial = "https://elchapuzasinformatico.com/"

# Hayamos cual es el numero de paginas totales del blog por si quisieramos
# descargar todas las entradas (en esta practica solo se hace hasta la pagina 20)
firstpage = requests.get(page_inicial)
tree = html.fromstring(firstpage.text)
numpage = int(float(tree.xpath('//*[@id="main"]/div/div/nav/div/a[4]/text()')[0])*1000)

#Creamos nuestro archivo csv
with open('Dataset.csv', 'w', encoding = "utf-8") as csvfile:
    writer = csv.writer(csvfile)

    #Bucle para recorre las 20 paginas mas recientes
    for i in range(1, 20):

        #Hacemos evolucionar las pagina a cargar
        if i > 1:
            url = "%spage/%d/" % (page_inicial, i)
        else:
            url = page_inicial

        #Petición a la web
        page = requests.get(url)

        #Pasamos el continido de la web a un objeto BeautifulSoup()
        soup = BeautifulSoup(page.text, "html.parser")

        #Buscamos todas las class donde estan las entradas
        entradas = soup.find_all(class_='post-content')

        #En cada entrada buscamos el titulo, autor y la fecha
        for entrada in entradas:
            titulo = str(entrada.find(class_='entry-title').getText())
            metadatos = entrada.find(class_='entry-meta')
            autor = str(metadatos.find(class_='post-author-name').getText())
            fecha = str(metadatos.find(class_='post-date').getText())

            #Almacenamos los datos en una lista
            datos = [titulo, autor, fecha]

            #Imprimimos los datos
            print (datos)

            #Guardamo s los datos en un archivo CSV
            writer.writerows([datos])


print("Fin")
```

10. Entradas-blog-ElChapuzaInformatico