PAC 1: Web Scrapping

Sergi Alonso Badia

Descripción

Quotes.toscrape, una web pública de aprendizaje.

El dataset se ha generado a partir de los datos obtenidos en la web http://quotes.toscrape.com, una página web diseñada expresamente para la práctica del scrapping. Se ha utilizado esta página ya que me ha parecido adecuada para la práctica, ya que, sin centrarse en el contenido de los datos, se pueden practicar las distintas habilidades que supone a nivel de desarrollar código Python para obtener el dataset.

El contenido de la practica y su código se pueden encontrar en el siguiente enlace:

https://www.github.com/sergialonsaco/uoc-web-scrapper

Imagen identificativa

Figura 1: imagen identificativa del dataset

La imagen identificativa seleccionada para el dataset no podría ser otra que la de unas marcas de **"cita"**, ya que representan adecuadamente el contenido del dataset.

Contexto

Como ya se ha comentado en la descripción, este dataset se ha obtenido simplemente para el desarrollo de la práctica, para desarrollar un código Python que generara correctamente el dataset, utilizando Selenium para navegar por el sitio web y BeautifulSoup para obtener los datos. También se ha utilizado Pandas para la captura de los datos en variable y facilitar la posterior generación del fichero csv.

Contenido

El dataset contiene los siguientes campos:

- **Author:** String con el nombre y apellido del autor de la cita.
- **Quote:** String con la cita completa.
- **Tags:** Array de strings con cada uno de los tags con los que se ha clasificado la cita.
- **Author_about:** String con la url a la descripción del autor de la cita.
- **Top_Ten_tags:** Boolean indicativo de si la cita contiene algún tag de la lista de Top Ten tags de la web. (en csv se ha representado con 1.0=True, 0.0=False)

Licencia

La licencia seleccionada para este dataset es **Released Under CC0: Public Domain License**, ya que el dataset se ha generado de una web pensada para prácticar el web scrapping y su contenido no tiene ninguna otra finalidad que la educativa.

Agradecimientos

Me gustaría dedicar unos agradecimientos a los autores de la web http://toscrape.com ya que gracias a esta web he podido realizar esta práctica de una manera muy didáctica, y que de seguro ha ayudado a muchos otros estudiantes.

Código fuente

El código fuente al completo se puede consultar en el repositorio de github:

http://www.github.com/sergialonsaco/uoc-web-scrapper

El dataset se puede obtener a partir de este enlace:

https://github.com/sergialonsaco/uoc-web-scrapper/releases/download/0.0.1/dataset_0.0.1.csv