2019

PEC 2 - Web Scraper

ALEJANDRO TORRES BRITO / CRISTINA RODRIGUEZ MARTINEZ - 10/04/2019

PEC 2 Web Scraper

WEB SCRAPER CON LIBRERÍA BEAUTIFUL SOUP

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

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

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

3. 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).

4. Representación gráfica. Presentar una imagen o esquema que identifique el

dataset visualmente

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de

los datos y cómo se ha recogido.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario

incluir citas de investigación o análisis anteriores (si los hay).

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

8. 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

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

10. Dataset. Presentar el dataset en formato CSV

1. **Contexto:**

En la práctica hemos recopilado una lista con los libros de referencia de Python mejor valorados en la web openlibra.com. Este sitio web es una librería especializada en material científico con un alto número de usuarios, que puede dar una idea de la popularidad de los títulos publicados por materia.

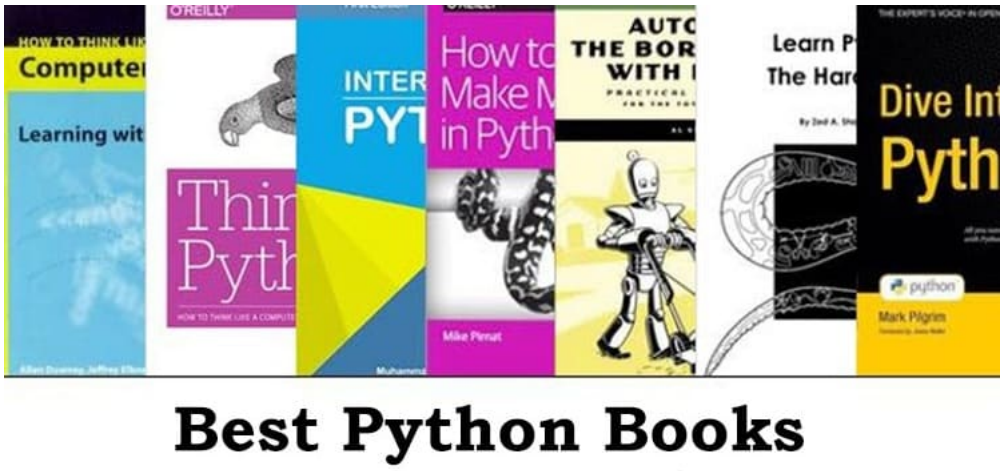
2. **Título:**

Los libros más valorados de Python

3. **Descripción del dataset:**

En el dataset se han recogido los libros científicos relacionados con el lenguaje de programación Python y la valoración que de ellos han realizado sus lectores.

4. **Representación gráfica (icono descriptivo):**



5. **Contenido:**

Los datos se han recogido haciendo un scraping del sitio:

https://openlibra.com/es/collection/search/category/programacion_python
Del listado de libros publicado en la página, se han obtenido campos descriptivos de la obra y las correspondientes reseñas de los lectores.

La página web lleva vendiendo libros de esta temática desde su inicio en 2011, por lo que los datos recogidos corresponderían con los años 2011 – 2019, periodo de tiempo en el que se han recogido las valoraciones de los libros que, a su vez, se actualizan diariamente. (¿?)

Los campos de los que consta el dataset son: título, autor, editorial, número de páginas, puntación media y número de votos. Estos campos se han recogido con un scraper implementado con la librería Beautiful Soup con la versión 3.7 de Python con un Jupiter Notebook ejecutado desde Anaconda. Como herramiento de colaboración para el desarrollo software se ha utilizado un repositorio en GitHub, con un Master y dos branches, que han servido para generar el código final.

6. Agradecimientos:

En este caso los datos están publicados en el sitio web Openlibra.com, perteneciente al proyecto y web etnasoft.com desarrollado por Carlos Benítez. Ambos sitios están publicados bajo licencia Creative Commons 3.0, cuyo texto se detalla debajo. Cabe agradecer el trabajo realizado por Carlos Benítez para publicar todo este contenido de forma desinteresada.

7. Inspiración:

El dataset es relevante para identificar fácilmente qué textos han sido más útiles a los lectores para avanzar en su formación y poder recomendarlos como material de estudio en Universidades o cursos.

8. Código:

El código con el que se ha implementado el scraper se adjunta en un archivo .txt y en un archivo .ypnb. El repositorio creado en GitHub está configurado como privado por lo que es necesario que el usuario interesado solicite acceso para acceder a él (https://github.com/crodriguezmartinez/Web-Scraper)

Contribución	Firma
Investigación previa	Alejandro Torres Brito
Redacción de las respuestas	Cristina Rodríguez Martínez
Desarrollo código	Especialización HTML: Alejandro Torres Brito Especialización Pandas y Data Frames: Cristina Rodríguez Martínez