PRÁCTICA 1: WEB SCRAPING

Byron Lima Rojas
bvlima@uoc.edu

28 de octubre de 2017

Índice

1. Título 3

2. Subtítulo 3

3. Imagen - Identificativo del Dataset 3

4. Contexto. ¿Cuál es la materia del conjunto de datos? 3

5. Contenido. ¿Qué campos incluye?¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido? 3

6. Agradecimientos. ¿Quién es el propietario del conjunto de datos? 4

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder a la comunidad? 4

8. Licencia 5

9. Código fuente 5

10. Dataset 6

11. Bibliografía 7

1. Título

Alojamientos en Loja - Airbnb en conjunto con el Festival de Artes Vivas 2017

2. Subtítulo

Información de alquiler de habitaciones, casas y departamentos en la ciudad de Loja, Ecuador. Airbnb como otra alternativa para asistir al Festival de Artes Vivas a desarrollarse en el mes de Noviembre en la capital musical.

3. Imagen - Identificativo del Dataset

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

El conjunto de datos es analizado en su mayor parte por la estadística, en donde cada columna representa el valor de una variable y cada fila es considerado un registro o información de un elemento del conjunto, también conocido como dato. En la actualidad esta información puede ser extraída desde una base de datos o mediante técnicas de extracción de información, la misma tendrá que pasar un ciclo de vida hasta llegar a la publicación de datos y conclusiones de la misma (Wiki,2017).

En la extracción de datos de la página AirBNB es importante señalar que, en este proceso se ejecutaran las fases de captura y almacenamiento de datos.

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

Mediante la libreria scrapy de Python 2.7 se obtienen los siguientes datos de Airbnb:

- Baños
- Dormitorios
- Huéspedes

- Tipo

De la librería de scrapy, se utilizan las siguientes clases para obtener los datos:

- from scrapy.item import Item, Field: permite crear la clase tipo Ítem y definir la estructura de los datos que vamos a obtener de un sitio web.
- from scrapy.spiders import CrawlSpider, Rule: define las reglas que debe seguir el spider para revisar datos y se crea la clase que hereda CrawlSpider que nos permite ir de forma horizontal y vertical.
- from scrapy.loader import ItemLoader: establece el inicio de la recolección de datos por las diferentes etiquetas del sitio web, los datos se obtienen mediante código Xpath del navegador Chrome.
- from scrapy.linkextractors import LinkExtractor: define el recorrido que debe realizar el spider para obtener los datos, adicional define hacía que página o subsitios puede acceder.
- from scrapy.loader.processors import MapCompose: permite editar los datos extraídos mediante funciones lambda.

El periodo de tiempo de los datos en la parte de recolección tardó un minuto, en la parte de vigencia de los mismos puede ser de 15 días por el incremento de lugares que puedan ofertarse.

6. Agradecimientos. ¿Quién es el propietario del conjunto de datos?

El propietario de estos datos es el administrador de Airbnb, que ha permitido la posibilidad de capturar datos desde su sitio web.

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

El conjunto de datos obtenido desde sitios web de alojamientos es importante entre grandes empresas, familias numerosas y viajeros, es una forma de encontrar sitios a buen precio cuando se trata de viajar a otra ciudad, de esta forma se optimizan gastos de estadía.

De acuerdo a los datos obtenidos, nos permitirá indicar al usuario el número de habitaciones que cuenta el lugar, número de baños, cantidad de camas y el máximo de personas admitidas por lugar, de esta forma el usuario/empresa determinará el sitio que se acople a sus necesidades.

8. Licencia

9. Código fuente

Para la obtención de datos con Scrapy de Python se utilizaron las librerias de Scrapy, Buittwith, Python-Whois y Urllib2.

El código fuente es el siguiente:

```
from scrapy.item import Item, Field
from scrapy.spiders import CrawlSpider, Rule
from scrapy.loader import ItemLoader
from scrapy.linkextractors import LinkExtractor
from scrapy.loader.processors import MapCompose
__autor__ = 'ByronLimaRojas'

class HospedajeItem(Item):
    camas = Field()
    banios = Field()
    dormitorios = Field()
    huespedes = Field()
    tipo = Field()

class Airbnb(CrawlSpider):
    name = "Hospedaje"
    start_urls = ['https://www.airbnb.com/s/Loja']
    allowed_domains = ['airbnb.com']

    rules = (
        Rule(LinkExtractor(allow=r'page=')),
        Rule(LinkExtractor(allow=r'/rooms'), callback='parse_items')
    )

    def parse_items(self, response):
        item = ItemLoader(HospedajeItem(), response)
        item.add_xpath('camas',
                       '//*[@id="summary"]/div[2]/div[1]/div[2]/div[1]
                       /div[2]/div/div[3]/div/div[2]/span/text()',
                       MapCompose(lambda i: i[0])),
        item.add_xpath('banios',
                       '//*[@id="summary"]/div[2]/div[1]/div[2]/div[1]
                       /div[2]/div/div[4]/div/div[2]/span/text()',
                       MapCompose(lambda j: j[0])),
```

```
item.add_xpath('dormitorios',
                '//*[@id="summary"]/div[2]/div[1]/div[2]/div[1]
                /div[2]/div/div[2]/div/div[2]/span/text()',
                MapCompose(lambda k: k[0])),
item.add_xpath('huespedes',
                '//*[@id="summary"]/div[2]/div[1]/div[2]/div[1]
                /div[2]/div/div[1]/div/div[2]/span/text()',
                MapCompose(lambda l: l[0])),
item.add_xpath('tipo',
                '//*[@id="summary"]/div[2]/div[1]/div[2]/div[1]
                /div[1]/div/div/div[1]/div[2]/div/div/span/a[1]
                /text()')
yield item.load_item()
```

10. Dataset

Los datos se encuentran en la ruta del repositorio `github.com/byrvin17/WebScrapingPY` bajo el nombre de **bnb.csv**. Los datos fueron exportados por consola bajo el siguiente comando: scrapy runspider airbnb.py -o bnb.csv -t csv --set CLOSESPIDER_ITEMCOUNT=30, esto con la finalidad de obtener los 30 primeros registros y evitar el baneo desde AirBNB.

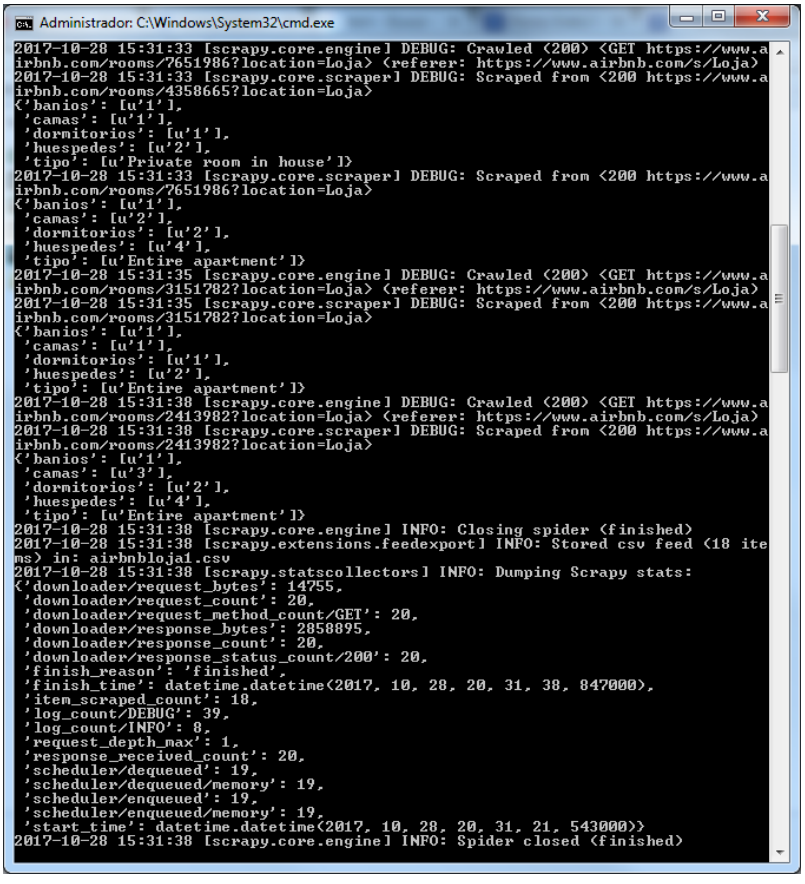

Figura 1: Ejecución de Scrapy en consola

11. Bibliografía

Referencias

[1] Conjunto de Datos (2017) [En línea].Wiki - UMAIC [Consulta: 28/10/2017]. https://wiki.umaic.org/wiki/Conjunto_de_datos.