PRÁCTICA 1 - Web Scraping

Tipología y ciclo de vida de los datos

Miguel Angel Sosvilla Luis

08-11-2020

Índice general

Contexto

La creación de este conjunto de datos se desarrolla como parte de la asignatura *Tipología y ciclo de vida de los datos* del Máster Universitario de Ciencia de Datos de la UOC. Los datos recogen los movimientos de pasajeros en los aeropuertos de españoles y se obtienen de la página web de estadísticas de tráfico aéreo de AENA, empresa pública española que gestiona los aeropuertos de interés general en España.

Título del dataset

Movimiento Aéreo de Pasajeros

Descripción del dataset

El conjunto de datos contiene el número mensual de viajeros de llegada y salida en los aeropuertos españoles desde el año 2005. Los datos están desglosados por compañía aérea.

Los datos se presentan en dos dataset uno con formato de tabla ancha, respetando el formato que ofrece AENA y otro con formato de tabla larga para facilitar el análisis de los mismos.

Representación gráfica

Figura 1: Representación gráfica

Contenido

Los datos quedan recogidos en dos conjuntos de datos en formatos de tabla ancha y tabla larga, con los siguientes campos:

Formato de tabla ancha - movimiento_pasajeros_ancha:

- **airline**: nombre de la compañía aérea que realiza el vuelo.
- **movimiento**: tipo de movimiento de pasajeros, llegada o salida.
- **aeropuerto**: aeropuerto de referencia de los datos.
- **year**: año.
- **total**: total de pasajeros embarcados o desembarcados en el año.
- **ene, feb,... dic**: número de pasajeros embarcados o desembarcados en el mes.

Formato de tabla larga - movimiento_pasajeros_laga:

- **airline**: nombre de la compañía aérea que realiza el vuelo.
- **movimiento**: tipo de movimiento de pasajeros, llegada o salida.
- **aeropuerto**: aeropuerto de referencia de los datos.
- **fecha**: mes en formato `aaaa-mm`.
- **num_pasajeros**: número de pasajeros embarcados o desembarcados en el mes.

Agradecimientos

El conjunto de datos ha sido creado gracias a la publicación de datos realizada por AENA en su página web de Estadísticas de Tráfico Aéreo.

Inspiración

La idea inicial de para el conjunto de datos fue analizar el impacto de la pandemia de la COVID19 en el movimiento de personas y, al disponer de datos desde el 2005, comparar y poner en contexto la crisis sanitaria actual con la crisis económica del 2007.

El conjunto de datos permite hacer comparativas entre aeropuertos y compañías aéreas e identificar de las más afectadas o más resistentes a las crisis.

Licencia

La licencia seleccionada es la Open Data Commons Open Database License (ODbL) ya que se trata de una licencia *Attribution Share-Alike for data/databases*, es decir, es específica para datos o bases de datos y permite:

- Compartir y adaptar el conjunto de datos mientras se reconozca al autor.
- La publicación de bases de datos derivadas mientras se mantenga también abierta usando la misma licencia.

Código

El código usado para crear los conjuntos de datos está disponible en el repositorio de Github https://github.com/msosvi/scraping_aena

Se trata de un script de Python que usa la librerías Selenium y BeatifulSoup para automatizar las consultas y recolectar los datos obtenidos en la página web de Estadísticas de Tráfico Aéreo de Aena.

La página presenta algunas dificultades para realizar el *scraping*:

- El diseño de la página se ha quedado anticuado (y lento) con uso intensivo de tablas.
- El uso de páginas diferentes para obtener los datos del mes actual y los anteriores.
- La estructura del resultado que no es constante y depende de la existencias de valores.

En la solución propuesta se combina el uso de Selenium y BeatifulSoup para solventar la lentitud de Selenium, herramienta pensada inicialmente para hacer test de aplicaciones web y que muestra alguna carencia al tener que recorrer la tabla de resultados, tarea que BeatifulSoup hace con mucha más soltura.

Dataset

Conjuntos de datos de muestra están publicados en Zenodo con el DOI 10.5281/zenodo.4242905.