TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Práctica 1: Web Scraping y Elaboración de un Dataset

Jose Luis Santos Durango
josant05@uoc.edu

María Isabel González Sánchez
mgonzalezsanchez19@uoc.edu

ÍNDICE

1. Contexto

Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

La página web que hemos elegido es Expatistan. Esta página web ofrece información sobre el **coste de vida** en distintos países del mundo y distintas ciudades, así como la opción de realizar una comparativa entre dos opciones elegidas. La información viene detallada por distintas categorías: comida, vivienda, ropa, transporte, cuidado personal y entretenimiento, y a su vez se categoriza por un nivel más detallado dentro de cada categoría. El objetivo del proyecto de creación de esta página era ofrecer una visión general del costo de vida en otros países/ciudades para que un expatriado pudiera hacerse una ligera idea del coste de vida en su futuro destino.

La forma en la que se añaden los datos es por **crowdsourcing**, de forma que los usuarios de la página web añaden los precios de sus propias ciudades. Posteriormente, mediante un modelo estadístico, se limpian dichos datos y se modelan para eliminar entradas erróneas y poder mostrar el promedio de los precios. Además, los datos introducidos se combinan con fuentes oficiales de consumo como: OCDE, INE, Consumer Expenditure Survey, Bureau of Labor Statistics, US DOL, entre otras.

Partiendo de un precio fiable de los productos o servicios ("ítems"), se realizará un cambio de divisa de la moneda local a una moneda común (euro), y se ponderan las distintas subcategorías, permitiendo así asignar una valoración al país o ciudad en cuestión, para poder generar un *ranking*.

El nivel de coste de vida en el *ranking* de cada país y ciudad tiene como base Praga, República Checa, de modo que se le asigna el nivel 100, siendo este el punto de referencia para el resto de comparaciones. El motivo de elegir Praga como punto de referencia es que no es una ciudad muy cara ni muy barata, con una gran comunidad de expatriados, y el propietario de la página es expatriado en Praga.

2. Título

Definir un título conciso y que sea descriptivo para el Dataset.

En este proyecto hemos extraído dos conjuntos de datos:

- **Cost of Living by Country**
  Este título corresponde al conjunto de datos almacenado en el fichero "cost_of_living_countries.csv" y contiene la información del coste de vida categorizado por país. El *scraping* ha sido realizado a la página que muestra el *ranking* por país e individualmente a cada una de las páginas de cada país del que se tiene información.

- **Cost of Living by City**

Este título responde al conjunto de datos almacenado en el fichero "cost_of_living_cities.csv" y contiene la información del coste de vida categorizado por ciudad. El *scraping* ha sido realizado a la página que muestra el *ranking* por ciudad e individualmente a cada una de las páginas de cada ciudad de la que se tiene información.

3. Descripción del Dataset

Desarrollar una breve descripción del conjunto de datos que se ha extraído. Es necesario que esta descripción sea coherente con el título elegido.

Los dos Datasets extraídos contienen información de precios de diferentes gastos que podemos tener en un **país** o **ciudad** determinados. Tal como se detalla en la página, los gastos de un "expatriado típico" (si existe tal cosa) se dividen en 6 grupos o **categorías**:

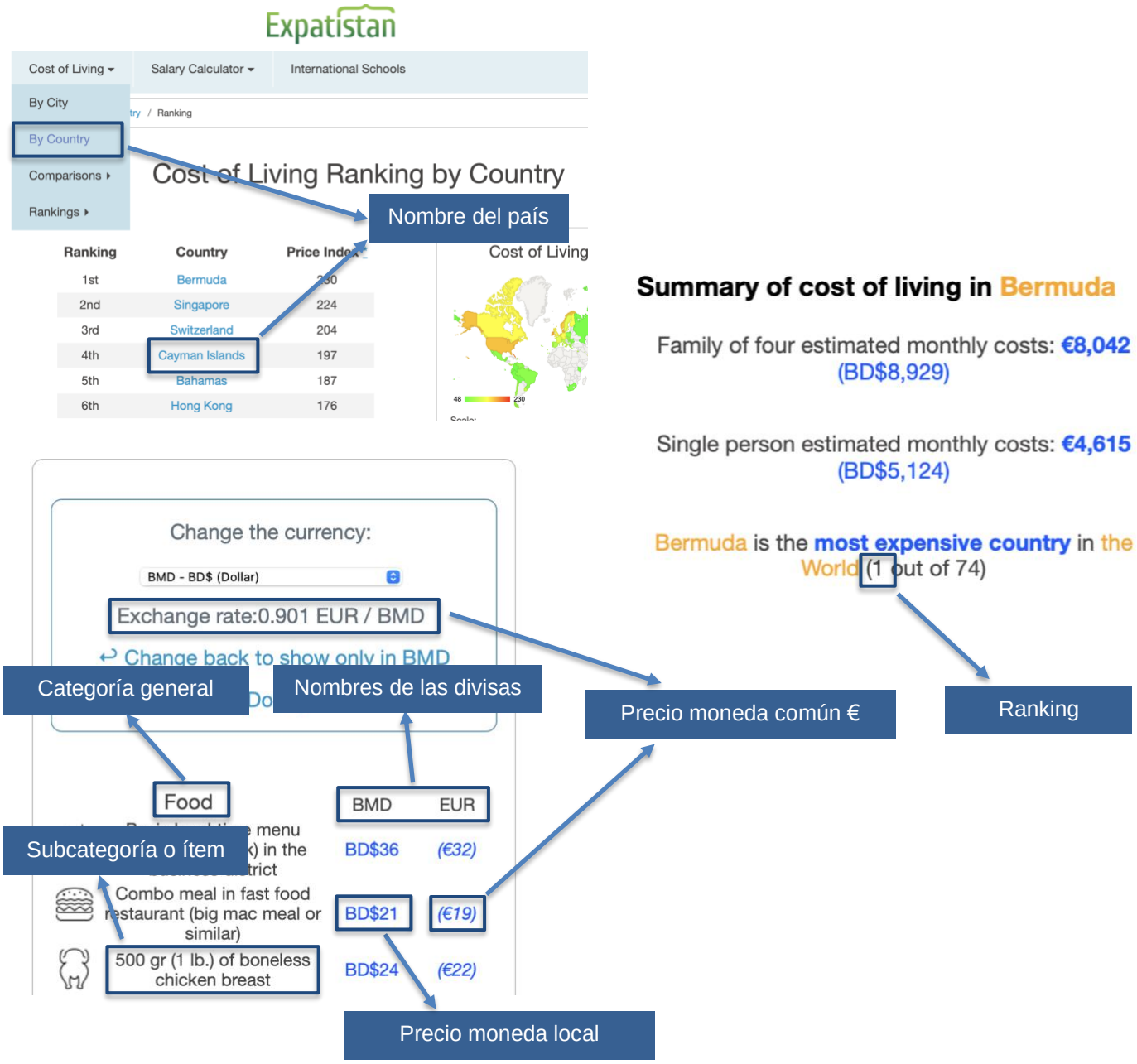
- Comida (*Food*)
- Alojamiento (*Housing*)
- Ropa (*Clothes*)
- Transporte (*Transportation*)
- Cuidado Personal (*Personal Care*)
- Entretenimiento (*Entertainment*)

Esta agrupación se basa en informes internacionales sobre variaciones en el coste de vida. A su vez, cada categoría tendrá sus subcategorías de productos o servicios (***items***), asociadas a su coste promedio en la moneda local (***original currency***). Dado que cada país o ciudad puede usar diferentes divisas, se ha escogido el Euro como moneda común para futuros estudios y comparativas (***exchanged currency***)

La diferencia entre ambos Datasets es la división geográfica. En el Dataset de Países, únicamente tenemos la información del **país**, mientras que en al Dataset de Ciudades, tenemos la categorización **país**, **ciudad** y **estado** (en caso de que la página tenga esa información).

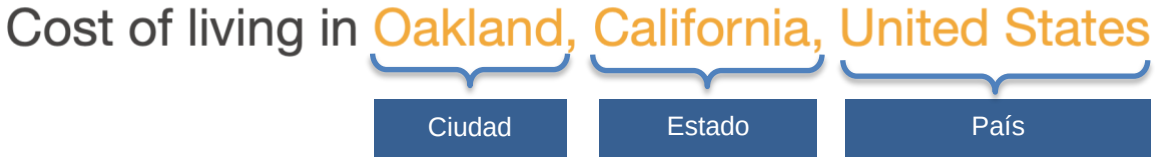
En la siguiente captura podemos apreciar de dónde se extrae cada columna de información explicada en los párrafos previos, siendo dichas columnas las resaltadas en **negrita**.

DATASET DE PAÍSES



DATASET DE CIUDADES

La principal diferencia con respecto al *scraping* de los datos de países, es la extracción del nombre del trío “ciudad, país, estado” en el caso de que existan para dicha ciudad:



4. Representación gráfica

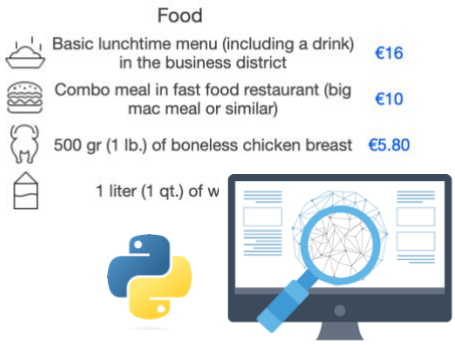
Dibujar un esquema o diagrama que refleje visualmente el Dataset y el proyecto elegido.

Fuentes de datos:

Diferentes canales oficiales.

Expatriados residentes en varios países actualizan los datos de precios en la propia página.

Procesado interno de los datos:
Calidad de datos, verificación, algoritmo de clasificación de precios, software interno para visualización de los datos



Web Scraping

Datasets

Análisis de los datos – PRA2

5. Contenido

Explicar los campos que se incluyen en el Dataset y el período de tiempo al que pertenecen los datos.

Los campos comunes a ambos Datasets son los siguientes:

- **Ranking position**
  Posición que ocupa el país o ciudad dentro del *ranking* de coste de vida de los países o ciudades incluidos en el estudio. Para los países puede ser del 1º al 74º y, en ciudades, del 1º al 230º.
- **Country**
  País sobre el que se realiza la valoración de los costes en el caso del *Web Scraping* de países o país al que pertenece la ciudad a la que se está aplicando el *Web Scraping* de ciudad.
- **Category**
  Categoría general de los costes e ítems. Este valor varía entre comida, alojamiento, ropa, transporte, cuidado personal y entretenimiento.
- **Items**
  Subcategoría de los costes. Consiste en productos o servicios ofrecidos en el país o ciudad, a los cuales se les asociará un valor monetario en la divisa local y en euros.
- **Original Currency**
  Acrónimo de la divisa local del país.
- **Original Currency Value**
  Coste del ítem en la divisa local.
- **Exchanged Currency**
  Acrónimo de la divisa común a la que se ha cambiado la divisa local. Por defecto en euros.
- **Exchanged Currency Value**

  Coste del ítem en la divisa común a la que se ha cambiado la divisa local.

Además, en el Dataset de ciudades se añaden los siguientes campos adicionales:

- **City**
  Ciudad sobre la que se realiza la valoración de los costos en el *Web Scraping* de ciudades.
- **State**
  Estado al que pertenece la ciudad a la que se está aplicando el *Web Scraping*, si existe.

El **periodo de tiempo** al que pertenece los datos podemos extraerlo de la siguiente cita del apartado "¿Cómo funciona?" de la web:

> "*La forma en que recopilamos datos es por "**crowdsourcing**". Nuestros usuarios añaden los precios para su propia ciudad de manera colaborativa* […]
>
> *Una vez que tenemos precios fiables para cada producto y ciudad, para comparar dos ciudades primero convertimos los precios de cada producto o servicio a una moneda común utilizando el tipo de cambio actual (**actualizado cada 3 días**)."*

Luego, podemos considerar que los datos se actualizan a medida que se van incorporando nuevos valores de las categorías en una ciudad en concreto. Entonces, cuando estos datos han sido validados por el algoritmo interno de la plataforma, se añadirán al cálculo de la media ponderada para sacar los *rankings*. Por lo tanto, podemos decir que la frecuencia de actualización es alta, y con respecto a los cambios de divisas podemos fijarlo en 3 días.

6. Propietario

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en su defecto, justificar esta búsqueda con análisis similares. Indicar qué pasos se han seguido para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto elegido.

La página web Expatistan nos presenta a su propietario de forma muy concisa y simple: **Gerardo Robledillo,** Ingeniero de Software especializado en aplicaciones web, en 2009 creó la web para poder dar respuesta a la pregunta típica de los expatriados "¿Cuánto dinero necesitaré en mi nueva ciudad?".

Siendo él mismo el "Provider", su página de Términos y Condiciones establece que él rige sobre la prestación de Servicios a través del sitio web www.expatistan.com a un tercero: el Usuario, que seremos nosotros. Legalmente se establecen una serie derechos y obligaciones para el Usuario además de una Licencia para casos concretos:

Citemos la Licencia (solo los puntos relevantes a nuestro trabajo):

"6. LICENSE

- *6.1. The Provider, within the Services, grants a license to the User to use the Services in a limited scope (hereinafter as the **"License"**).*
- *6.2. The License is granted:*
    - *6.2.1. as non-exclusive,*
    - *6.2.2. free of charge in case of the Cost of Living,*
    - *6.2.3. for the Price in case of the Salary Calculation,*

- *6.2.4. solely for personal and not commercial purposes only in accordance with these Terms and Conditions, and only in the appropriate manner,*
- *6.2.5. for a period of maximum one year,*
- *6.2.6. worldwide.* […]*"*

Gerardo Robledillo nos hace entrega de una licencia temporal para los valores que hemos extraído, que se corresponden con el punto 6.2.2. Por ende, el paso a seguir es ceñirnos solo a las URLs referentes al coste de vida.

Citemos los primeros Derechos y Obligaciones del Usuario:

"7. RIGHTS AND OBLIGATIONS OF THE USER

- *7.1. The User undertakes to use the Services solely in accordance with the legal regulations and these Terms and Conditions. The User is not entitled to use the Services for any other purpose or in any manner other than those set forth in these Terms and Conditions.*
- *7.2. Most importantly, the User hereby undertakes that he will not:*
  - *7.2.1. within the use of the Services or as a result of the use of the Services, interfere with the rights of third parties or the Provider,*
  - *7.2.2. interfere with the Services unjustifiably,*
  - *7.2.3. use the Services in a way that could damage it (including interference with the Website),*
  - *7.2.4. attempt to decompile or reverse engineer the Website or the Services,*
  - *7.2.5. conduct any systematic or automated data collection activities (including without limitation scraping, crawling, data mining, data extraction and data harvesting) on or in relation to the Website.* […]*"*

El punto que nos concierne es el 7.2.5, que nos prohíbe hacer un web scraping sin control ni limitaciones. Para ceñirnos a nuestras obligaciones, nuestro scraping se ha centrado en trozos concretos y acotados de los ficheros HTML del coste de vida. Por ende, estamos amparados en el marco legal de la página web.

Respecto a los análisis previos realizados con los datos de este dominio, podemos observar en el blog de la página, el propietario ha realizado un análisis del uso de la página tras ciertos acontecimientos como el Brexit o los resultados de las elecciones de Estados Unidos, donde el número de usuarios que comparan el precio de vivir en Reino Unido o EEUU con respecto a otras ciudades incrementa notablemente.

7. Inspiración

Explicar por qué puede ser interesante este conjunto de datos y qué preguntas se pretenden responder con ellos. Es necesario comparar con los análisis anteriores o análisis similares presentados en el apartado 6.

Ambos integrantes del grupo somos personas que nos gusta viajar y descubrir países y culturas nuevas. Por ende, esta página web nos ayudaba de manera muy intuitiva a preparar los gastos de cualquier viaje. De ahí la inspiración inicial, pero luego descubrimos su potencial estadístico.

Dado que con el *web scraping* del coste de vida hemos obtenido una gran cantidad de información sobre diversos países y ciudades, pretendemos responder a las siguientes preguntas:

1. De los Estados de US, ¿cuál de ellos presenta el acceso a medicamentos y doctores más caro y cuál el más barato? ¿cuáles son estas ciudades?
2. ¿Qué países puedes ajustarse a un perfil como el nuestro si en un futuro queremos ir a trabajar fuera?
3. ¿Qué presupuesto debemos planificar para viajar a los distintos países recopilados?
4. ¿Qué sueldo necesitaríamos por país y/o ciudad para poder alquilar un apartamento pequeño? ¿y uno grande?

Además, dado que los datos están siendo continuamente actualizados, y dada la situación de inflación global que actualmente estamos viviendo, será interesante realizar una comparativa entro los datos extraídos en esta práctica, con fecha 15 de abril de 2023, y los datos que se extraerán en un futuro para la práctica 2.

Con respecto a los análisis previos de los datos, tal como se puede ver en el blog, el único análisis visual realizado sobre los datos es la correlación del número de usuarios cuando sucede un acontecimiento importante en un país; no tenemos acceso al dato de usuarios de la página. Desde el punto de vista estadístico, podemos realizar comparativas de ciudades tal como se hace en la propia página.

8. Licencia

Seleccionar una licencia adecuada para el Dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse:

- **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.**
- **Otra (especificar cuál).**

Dados los claros comentarios de los Términos y Condiciones sobre el uso y distribución de los datos de la web, la licencia escogida para los Datasets será CC BY-NC-SA 4.0. Las causas de esta elección son las siguientes:

1. Se debe dar el correspondiendo crédito al autor de los Datasets, indicando la licencia y que cambios se han llevado a cabo.

2. Como sus siglas NC indican, no se permite el uso comercial del material. Esto lo tenemos que cumplir por el subapartado 6.2.4 del apartado "6. Licence" de los Términos y Condiciones que establece lo siguiente:
   " 6.2. The License is granted:
   […]
   - *6.2.4. solely for personal and **not commercial purposes** only in accordance with these Terms and Conditions, and only in the appropriate manner,"*

3. Cualquier cambio, transformación o mezcla de los datos de los Datasets debería distribuirse bajo la misma licencia que el original. Por lo tanto, esto permite mantener los términos establecidos con la licencia y el crédito al autor o autores de los conjuntos de datos.

9. Código

Código implementado para la obtención del Dataset, preferiblemente en Python o, alternativamente, en R.

- **Se deben indicar las librerías y versiones utilizadas.**
- **En la memoria en PDF, se deben comentar los aspectos más relevantes sobre cómo el código realiza el proceso de recolección de datos, qué dificultades presenta el sitio web elegido, y cómo se han resuelto.**

El código de esta primera práctica ha sido realizado en **Python 3.8.12**. Dado que nuestro IDE es Jupyter Notebook y extraer un archivo "requirements.txt" con *pip freeze* traerá todos los módulos de dependencias del *environment* general de Conda, se ha optado por complementar este fichero con el uso de la librería ***session-info,*** que es más específica.

```
In [26]: # Let's see this sessions dependencies and library versions
         session_info.show()

Out[26]: Click to view session information
         -----
         bs4                 4.10.0
         builtwith           NA
         pandas              1.3.3
         requests            2.26.0
         session_info        1.0.0
         utils               NA
         whois               NA
         -----
         Click to view modules imported as dependencies
         -----
         IPython             7.27.0
         jupyter_client      7.0.1
         jupyter_core        4.10.0
         jupyterlab          3.1.7
         notebook            6.4.3
         -----
         Python 3.8.12 (default, Oct 12 2021, 03:01:40) [MSC v.1916 64 bit (AMD64)]
         Windows-10-10.0.22621-SP0
         -----
         Session information updated at 2023-04-15 23:31
```

Esta librería de Python nos informa de:

1. Versión de módulos y librerías de la actual sesión (la del notebook).
2. Versión de Python
3. Sistema Operativo y CPU

Sin embargo, no es perfecta y faltan dos versiones: la de "*builtwith*" (1.3.4) y la de "*Python-whois*" (0.8.0). Estas versiones las podemos ver en el *output* de la celda de instalaciones:

```
In [27]: !pip install requests
         !pip install builtwith
         !pip install beautifulsoup4
         !pip install python-whois
         !pip install session-info

         Requirement already satisfied: requests in c:\users\isa31\anaconda3\lib\site-packages (2.26.0)
         Requirement already satisfied: charset-normalizer~=2.0.0 in c:\users\isa31\anaconda3\lib\site-packages (from requests) (2.0.4)
         Requirement already satisfied: certifi>=2017.4.17 in c:\users\isa31\anaconda3\lib\site-packages (from requests) (2021.10.8)
         Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\isa31\anaconda3\lib\site-packages (from requests) (1.26.7)
         Requirement already satisfied: idna<4,>=2.5 in c:\users\isa31\anaconda3\lib\site-packages (from requests) (3.2)
         Requirement already satisfied: builtwith in c:\users\isa31\anaconda3\lib\site-packages (1.3.4)
         Requirement already satisfied: six in c:\users\isa31\anaconda3\lib\site-packages (from builtwith) (1.16.0)
         Requirement already satisfied: beautifulsoup4 in c:\users\isa31\anaconda3\lib\site-packages (4.10.0)
         Requirement already satisfied: soupsieve>1.2 in c:\users\isa31\anaconda3\lib\site-packages (from beautifulsoup4) (2.2.1)
         Requirement already satisfied: python-whois in c:\users\isa31\anaconda3\lib\site-packages (0.8.0)
         Requirement already satisfied: future in c:\users\isa31\anaconda3\lib\site-packages (from python-whois) (0.18.2)
         Requirement already satisfied: session-info in c:\users\isa31\anaconda3\lib\site-packages (1.0.0)
         Requirement already satisfied: stdlib-list in c:\users\isa31\anaconda3\lib\site-packages (from session-info) (0.8.0)
```

Si nos centramos en cómo el código realiza el Web Scraping, es muy intuitivo. Para cada tipo de *scraping*, tenemos una clase representativa: **ExpatistanCountryScraper()** para la

extracción de países y **ExpatistanCityScraper(),** para ciudades. Estas clases serán las encargadas de llevar a cabo el *scraping* a base de llamar a su función principal "***scraping***".

Dado que ambas clases son muy similares, sus funciones comunes se han almacenado en el archivo "**utils.py**" en el mismo directorio que el código. Este fichero contendrá las funciones:

- **get_HTML(url)**
  Consigue y devuelve el archivo HTML de la URL introducida usando BeautifulSoup.

- **get_links(html, scrape_type)**
  Devuelve una lista con todos los *links* de ciudades (si el *scrape_type* es True) o países (si el *scrape_type* es False) que se encuentren el HTML pasado por parámetro

- **get_ranking_pos(html, num_digits)**
  Devuelve la posición del país o ciudad en su *ranking* respectivo. Dado que el *ranking* de países alcanza las 3 cifras (de 1º a 230º) y el de países solo llega hasta el puesto 74º, el parámetro "num_digits" determinará si la búsqueda es para países (valor igual a 2) o a ciudades (valor igual a 3).

- **saving(file_name, dataset)**
  Crea un Padas DataFrame a partir del Dataset introducido como parámetro y lo guarda en un fichero CSV con el nombre que contenga *file_name*.

Como ambas clases son similares, analizaremos el código de **ExpatistanCityScraper**() al ser la más completa. Primero, llamaremos a la única función pública: scraping(). Esta recuperará el HTML de la web inicial definida en el constructor de la clase y extraerá la lista de URLs de ciudades de dicho HTML. Luego, por cada ciudad, realizaremos su scraping individual con "**__scraping_single_city**".

Por cada ciudad, conseguimos su HTML y su puesto en el *ranking*. Luego, escogemos el punto de partida en el archivo para iniciar la extracción de información, siendo este el **primer problema** que nos encontramos. De primeras, como la información se encuentra almacenada en *tags* de tipo <tr>, extrajimos todos las etiquetas de este tipo del HTML. Por ende, a parte de la información relevante, también recibíamos etiquetas extras como, por ejemplo, *tags* sobre el formato de las columnas de la web. Así que la solución fue coger la tabla que contenía las etiquetas <tr> que buscábamos y luego extraer dichos *tags*.

Dado que estamos explicando el Web Scraping de ciudades, este tipo de extracción presenta el trío "ciudad + estado + país". He aquí el **segundo problema**: no todas las ciudades presentan "estado". Como dejar el valor en blanco era un claro error y no extraer esta información relevante también, se ha optado por usar el valor "No state" en caso de no existir el estado. De esta forma podemos hacer un estudio por estados en un futuro.

De seguido pasamos a extraer el resto de los valores comunes a ambos *scrapings*. Por cada etiqueta <tr> en la tabla con clase "comparison single-city", haremos las siguientes comparaciones:

- **La etiqueta <tr> presenta la clase "categoryHeader"**
  Esto nos informa que es la etiqueta que contiene el tag <th> con el nombre de la categoría a extraer. Por ende, encontramos su primera etiqueta <th> y guardamos su texto.

- **La etiqueta <tr> no tiene clase**
  En este caso, hay 3 posibilidades a estudiar según el número de tags <td> que contenga:

  - ***Contiene 2 tags <td>***
    Esto nos indica que estamos ante una ciudad con las columnas de moneda local propia y la del cambio de divisa a euro. En otras palabras, el texto de cada etiqueta <td> presenta el acrónimo de las respectivas divisas. Como hemos fijado que el euro sería siempre la segunda divisa, solo extraeremos y guardaremos el nombre de la moneda local.

  - ***Contiene 3 tags <td>***
    Nos encontramos en la URL de una ciudad con el euro como divisa local y este fue nuestro **tercer problema**: ¿qué ocurre con los países que usan euros? Al intentar añadir la misma divisa como moneda secundaria, la web no reconocía el cambio. Por ende, la mejor solución para mantener la integridad de los Datasets fue que en estos casos las columnas de divisa local y comunitaria serían iguales.

    Por otro lado, en estas 3 etiquetas encontraremos respectivamente un icono, el item y el coste del item en euros (divisa local). Puesto que el icono no nos aporta nada, saltaremos este *tag* y procederemos guardar los otros dos.

    En resumen, aquí guardamos una fila entera de datos extraídos en el Dataset definido en el constructor de la clase. Sin embargo, una vez realizado todo el scraping, nos fijamos en el **cuarto problema** con la web: no todas las ciudades tienen el coste de cada uno de los items. Estos aparecen con un guion en vez de un valor y no es explicativo. Por lo tanto, se estableció que si se daba el caso de que el coste es "-", se guardaría el identificador "Not defined" para su posterior tratamiento en la práctica 2.

  - ***Contiene 4 tags <td>***
    Nos encontramos en la URL de una ciudad con divisa local propia y divisa comunitaria establecida a euros. Su funcionamiento es igual al anterior caso, pero añadiendo un cuarto tag <td>, que contiene el coste del item en euros. Por lo tanto, también guardaremos una fila entera del Dataset, pero con dos divisas diferentes.

Una vez completado el Dataset, solo queda llamar a la función "*saving*" del archivo "*utils.py*" para que genere un Pandas DataFrame con el archivo JSON que es el Dataset y lo guarde como fichero CSV en el directorio /dataset.

10. Dataset

Publicar el Dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción de la misma. Obtener y adjuntar el enlace del DOI del Dataset (https://doi.org/…).

Los enlaces DOI de los Datasets son los siguientes:

- Cost of Living by Country: https://doi.org/10.5281/zenodo.7833244
- Cost of Living by City: https://doi.org/10.5281/zenodo.7833285

11. Repositorio GitHub

https://github.com/Tipologia-y-Ciclo-de-Vida-de-los-Datos/Practica1-Web_Scraping.git

12. Vídeo

Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), que deberá contar con la participación de los dos integrantes del grupo.

En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC.

El vídeo de la PRA1 se ha subido al apartado “Vídeo Práctica” del tablón de la asignatura.

13. Contribuciones

En este apartado, desglosaremos la contribución de cada integrante del grupo en las distintas fases y partes de la práctica. Siendo Jose Luís Santos Durango las siglas “JLSD” y María Isabel González Sánchez, “MIGS”, se concluye la siguiente tabla:

CONTRUBUCIONES	FIRMA
Investigación previa	JLSD, MIGS
Redacción de respuestas	JLSD, MIGS
Desarrollo del código	JLSD, MIGS
Participación en el vídeo	JLSD, MIGS

14. Referencias

- **Recurso web: Expatistan.** https://www.expatistan.com
- **Laia Subirats Maté, Mireia Calvo González.** *Web Scraping.* Universitat Oberta de Catalunya.
- **Richard Lawson (2015).** *Web Scraping with Python – Chapter 1 – Introduction to Web Scraping.*
- **Richard Lawson (2015).** *Web Scraping with Python – Chapter 2 – Scraping the data.*