Práctica 1: Web Scraping

Componentes del grupo de trabajo:

- Carlos Muñiz Solaz
- Gregorio de Miguel Vadillo

Repositorio:

https://github.com/gregoriodmv/gmv_cms_web_scraping

Tabla de contenido

Descripción de la práctica

En este apartado se contesta a las preguntas formuladas en el enunciado de la práctica:

1. Título del dataset

Dataset de Composición de Alimentos

2. Subtítulo del dataset

Repositorio generado durante la Práctica 1 de la asignatura *Tipología y ciclo de vida de los datos* del *Máster de Ciencia de Datos* de la *UOC*. En esta práctica hemos realizado *Web Scraping* sobre la **Base de Datos Española de Composición de Alimentos (BEDCA)** para generar un dataset sobre la composición de los alimentos. En el repositorio se puede encontrar el código desarrollado, así como el dataset generado.

Página oficial de BEDCA: http://www.bedca.net/

3. Imagen

4. Contexto

Las **Bases de Datos de Composición de Alimentos (BDCA)** son un conjunto de informaciones que detallan los componentes más importantes de los alimentos, así como sus valores de energía y nutrientes. Entre estos componentes, encontramos las proteínas, los carbohidratos, las grasas, las vitaminas, los minerales y otros componentes nutricionales importantes como la fibra. Para más información sobre las Bases de Datos de Composición de Alimentos ver:

http://www.fao.org/3/a-y4705s.pdf
https://en.wikipedia.org/wiki/Food_composition_data

5. Contenido

El dataset inicial que hemos generado lo hemos guardado en el fichero: *alimentos_raw.csv*. Posteriormente, hemos generado otro dataset más procesado al que hemos llamado *alimentos_procesado.csv*.

El número de registros que hemos conseguido extraer son 949 y el número de campos son 41 en el caso de *alimentos_raw.csv* y de 42 en el caso de *alimentos_procesado.csv*, ya que incluye una columna con el índice.

Los campos que encontramos en los datasets generados son los siguientes:

- Nombre del alimento
- Campos sobre carbohidratos: fibra o carbohidratos
- Campos sobre grasas: datos sobre los distintos ácidos grasos
- Campos sobre vitaminas: A, D, E, ...
- Campos sobre minerales: hierro, calcio, ...
- Campos proximales: alcohol, energía total, grasa total, proteína total y agua

La mayoría de los campos son valores continuos que representan valores de masa (en g, mg o ug).

Existen campos de energía expresados en kJ o kCal.

Algunos micronutrientes cuyo porcentaje en el alimento es ínfimo se representan mediante la etiqueta "traza" en lugar de un valor numérico.

A la hora de analizar los datos hay que tener en cuenta que:

- Existe una gran variabilidad de la composición de los alimentos entre los distintos países
- Existen muchos datos incompletos tanto de alimentos como de nutrientes al desconocerse su valor

Por último, es importante tener en cuenta el *periodo de tiempo de los datos*. Debido a limitaciones en los recursos, mucho de los valores no están actualizados o son muy

difíciles de obtener con exactitud. También hay que tener en cuenta que los métodos para obtener los componentes van variando a lo largo de los años, con lo que es posible que el dataset que hemos generado se quede obsoleto a lo largo de los años. Por ello, y para asegurar la validez de los datos, habría que repetir el proceso de web scraping de forma periódica o incluso modificar el proceso si la forma de disponer la información variara.

6. Agradecimientos

Nos gustaría agradecer a la Red BEDCA, fuente original de los datos obtenidos:

AESAN/BEDCA Base de Datos Española de Composición de Alimentos v1.0 (2010)

Por la elaboración de la base de datos con la composición de alimentos y la página web que nos ha permitido generar nuestro dataset:

http://www.bedca.net/

Agradecer también a las distintas fuentes que han permitido conocer la composición de los alimentos y que han posibilitado su incorporación en la base de datos. Se puede encontrar más información sobre las distintas fuentes en el siguiente enlace:

http://www.bedca.net/bdpub/index.php

7. Inspiración

Los alimentos que ingerimos afectan de manera asombrosa sobre la salud de las personas. Pero no solo a nivel físico, sino también a nivel psicológico. Se dice que hay alimentos "que curan". Se sabe que hay alimentos que mejoran el estado de ánimo. Otros mejoran la concentración. Otros mejoran el rendimiento deportivo, así como otros consiguen todo lo contrario. Se habla también de *superalimentos*. Existen testimonios de personas cuya vida ha cambiado simplemente realizando un cambio en su alimentación. Por lo tanto, para que un individuo pueda desempeñar sus funciones correctamente, es fundamental que su alimentación sea la correcta, y eso pasa por tener conocimiento de la composición de los alimentos que forman esa alimentación. Estos y otros motivos hacen que el conjunto de datos propuesto sea de interés máximo.

Las aplicaciones de la *Base de Datos de Composición de Alimentos* son muy amplias y permiten responder por sí mismas a distintas dudas y problemas presentados en la sociedad. Algunos ejemplos para los que se usan estas bases de datos son la:

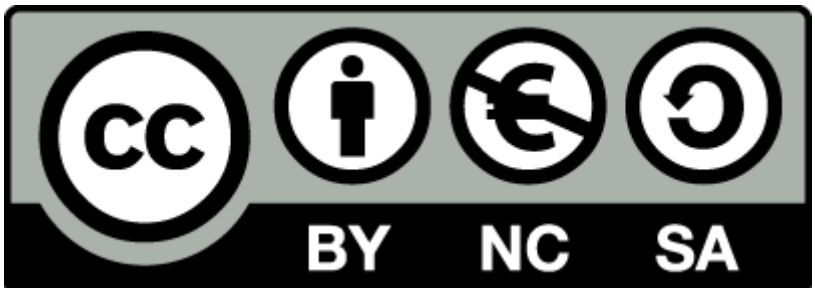
- Elaboración de *dietas terapéuticas*: para tratar la obesidad, diabetes, alergias e intolerancias a alimentos.

- Elaboración de *dietas institucionales*: colegios, hospitales, prisiones, centros de día.
- Elaboración de *dietas epidemiológicas*: estudiar el efecto de las dietas sobre la población. Por ejemplo, para perder peso, ganar musculo, ...
- Realización de estudios cuantitativos sobre la nutrición humana.
- Para educar a la población sobre los alimentos y sus nutrientes.
- Para elaborar las etiquetas de los productos procesados.

8. Licencia

El dataset de Composición de Alimentos de Gregorio de Miguel Vadillo y Carlos Muñiz Solaz se encuentra bajo la licencia **CC-BY-NC-SA 4.0**.
https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

como se explica en:
https://creativecommons.org/licenses/



Se ha elegido está licencia ya que permite a otros distribuir, remezclar, retocar, y crear a partir de tu obra de modo no comercial, siempre y cuando te den crédito y licencien sus nuevas creaciones bajo las mismas condiciones. Esto se debe a la licencia de la página de BEDCA que no permite utilizar los datos para fines comerciales como se detalla en el siguiente documento:
Uso de la Base de Datos

9. Código

El código empleado se encuentra en la siguiente ruta del repositorio:

https://github.com/gregoriodmv/gmv_cms_web_scraping/blob/2017_11_09_web_scraping/src/

Se incluyen tanto los ficheros utilizados para la extracción de información y generación del dataset:

- *alimento.py*
- *alimentos_scraper.py*
- *main.py*

Como el script encargado del procesado en R para generar un dataset limpio:

- *procesado.py*

En la Wiki del repositorio hay información adicional sobre los archivos.

10. Datasets

La ruta que contiene los datasets es la siguiente:

https://github.com/gregoriodmv/gmv_cms_web_scraping/blob/2017_11_09_web_scraping/dataset

Contiene tanto el dataset en crudo obtenido durante el proceso de web scraping (**alimentos_raw.csv**) como el dataset obtenido tras procesarlo con R (**alimentos_procesado.csv**).