M2.851 - Tipología y ciclo de vida de los datos

PRA 1

Nombre: PRA1 - Web Scraping
Fecha: 10 de abril de 2019
Autores: Azucena González Muiño;Jesús Márquez Díaz

Contenido

M2.851 - Tipología y ciclo de vida de los datos

PRA 1

Contexto

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

Una de las principales enfermedades de la sociedad actual es el estrés que, según los últimos estudios, está muy relacionado con los hábitos alimentarios, repercutiendo negativamente en estos y produciendo un círculo vicioso que actúa en detrimento de una buena salud. En un entorno de concienciación creciente tanto por parte de los ciudadanos como de los gobiernos por mejorar los hábitos alimenticios y conocer los principios de una alimentación saludable, creemos fundamental facilitar información nutricional precisa y veraz de los alimentos que ingerimos.

La web elegida para nuestro estudio es la Base de Datos Española de Composición de Alimentos (BEDCA), que ha sido desarrollada y es mantenida por la red BEDCA, un conjunto de centros de investigación públicos, administración e instituciones privadas.

La red BEDCA se ha formado por la colaboración del Ministerio de Ciencia e Innovación y el Ministerio de Sanidad, Servicios Sociales e Igualdad (a través de la Agencia Española de Seguridad Alimentaria y Nutrición).

El objetivo de la web BEDCA es el de elaborar una base de datos española de composición de alimentos unificando los datos fragmentados y no documentados que existían. Los valores de composición de alimentos recogidos en esta base de datos han sido obtenidos de distintas fuentes que incluyen laboratorios, industria alimentaria y publicaciones científicas, o calculados. Está construida con los estándares europeos desarrollados por la Red de Excelencia Europea EuroFIR y se encuentra incluida en la lista de bases de datos de composición de alimentos de dicha asociación.

Título para el dataset

Elegir un título que sea descriptivo para el conjunto de datos.

El nombre elegido para nuestro dataset es Composición Nutricional de Alimentos, ya que consideramos que describe de manera adecuada el contenido de nuestros datos.

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).

El dataset generado recoge la composición nutricional de un conjunto de alimentos en base a la información recopilada de la web de la red BEDCA (Base de Datos Española de Composición de Alimentos).

Dentro del conjunto de datos se encuentra el nombre común y científico de los alimentos, así como el contenido detallado en grasas, hidratos de carbono, vitaminas minerales y proximales.

Las unidades de medida que aplican son las siguientes:

- gramos (g): para la mayoría de los componentes como pueden ser las grasas, el agua o los ácidos grasos.
- kilojulios (kJ): para medir la energía total.
- microgramos (ug): para medir colesterol, la mayoría de las vitaminas o el folato.
- miligramos (mg): para medir el calcio, el hierro o el potasio.

Los posibles valores que pueden tomar los componentes alimenticios son:

- Valor numérico: indica la cantidad del componente que tiene el alimento.
- Valor TR: existen trazas del componente en el alimento.
- Valor NA: no se dispone de información sobre la cantidad del componente en el alimento.
- Valor BL: el valor del componente está por debajo del límite de detección.
- Valor UD: el valor del componente es indecidible o imposible de detectar.

Representación gráfica

Presentar una imagen o esquema que identifique el dataset visualmente.

Contenido

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

En la web de la red BEDCA se indica que el contenido de la base de datos de alimentos se actualiza de forma periódica, aunque no se especifica un periodo concreto. Se ha solicitado vía email esta información y, a la fecha de realización de este documento, aún no se ha recibido respuesta. Los datos del csv entregado se han obtenido a fecha 10/04/2019.

A continuación, se enumeran todos los campos que componen nuestro dataset. A menos que se indique lo contrario, todas las cantidades hacen referencia a 100 gramos de porción comestible.

Nombre del campo	Tipo de dato	Descripción del campo
f_id	Numérico	Identificador del alimento.
f_ori_name	Texto	Nombre del alimento.
sci_name	Texto	Nombre científico.
edible_portion	Numérico	Porción comestible.
alcohol (etanol)	Numérico	Cantidad del alcohol medida en gramos.
energía total	Numérico	Energía medida en kilojulios.
grasa, total (lipidos totales)	Numérico	Cantidad de grasa medida en gramos.
proteína, total	Numérico	Cantidad de proteína medida en gramos.
agua (humedad)	Numérico	Cantidad de agua medida en gramos.
carbohidratos	Numérico	Cantidad de carbohidratos medida en gramos.
fibra, dietetica total	Numérico	Cantidad de fibra medida en gramos.
ácido graso 22:6 n-3 (ácido docosahexaenóico)	Numérico	Cantidad de ácido graso específico medida en gramos.
ácido graso 20:5 (ácido eicosapentaenóico)	Numérico	Cantidad de ácido graso específico medida en gramos.
ácido graso 12:0 (láurico)	Numérico	Cantidad de ácido graso específico medida en gramos.
ácido graso 14:0 (ácido mirístico)	Numérico	Cantidad de ácido graso específico medida en gramos.
ácido graso 16:0 (ácido palmítico)	Numérico	Cantidad de ácido graso específico medida en gramos.
ácido graso 18:0 (ácido esteárico)	Numérico	Cantidad de ácido graso específico medida en gramos.
ácido graso 18:1 n-9 cis (ácido	Numérico	Cantidad de ácido graso específico

oléico)		medida en gramos.
ácido graso 18:2	Numérico	Cantidad de ácido graso específico medida en gramos.
ácido graso 18:3	Numérico	Cantidad de ácido graso específico medida en gramos.
ácido graso 20:4 n-6 (ácido araquidónico)	Numérico	Cantidad de ácido graso específico medida en gramos.
ácidos grasos, monoinsaturados totales	Numérico	Cantidad de ácido graso específico medida en gramos.
ácidos grasos, poliinsaturados totales	Numérico	Cantidad de ácido graso específico medida en gramos.
ácidos grasos saturados totales	Numérico	Cantidad de ácido graso específico medida en gramos.
ácidos grasos, trans totales	Numérico	Cantidad de ácido graso específico medida en gramos.
colesterol	Numérico	Cantidad de colesterol medido en microgramos.
Vitamina A equivalentes de retinol de actividades de retinos y carotenoides	Numérico	Cantidad de vitamina medida microgramos.
Vitamina D	Numérico	Cantidad de vitamina medida microgramos.
Vitamina E equivalentes de alfa tocoferol de actividades de vitámeros E	Numérico	Cantidad de vitamina medida en miligramos.
folato, total	Numérico	Cantidad de ácido fólico medida en microgramos.
equivalentes de niacina, totales	Numérico	Cantidad de niacina (o equivalente) medida en microgramos.
riboflavina	Numérico	Cantidad de riboflavina medida en miligramos.
tiamina	Numérico	Cantidad de tiamina medida en miligramos.
Vitamina B-12	Numérico	Cantidad de vitamina medida en microgramos.
Vitamina B-6, Total	Numérico	Cantidad de vitamina medida en miligramos.
Vitamina C (ácido ascórbico)	Numérico	Cantidad de vitamina medida en miligramos.
calcio	Numérico	Cantidad de calcio medida miligramos.
hierro, total	Numérico	Cantidad de hierro medida en miligramos.
potasio	Numérico	Cantidad de potasio medida en miligramos.
magnesio	Numérico	Cantidad de magnesio medida en

		miligramos.
sodio	Numérico	Cantidad de sodio medida en miligramos.
fósforo	Numérico	Cantidad de fósforo medida en miligramos.
ioduro	Numérico	Cantidad de ioduro medida en microgramos.
selenio, total	Numérico	Cantidad de selenio medido en microgramos.
zinc (cinc)	Numérico	Cantidad de zinc medida en miligramos.

Agradecimientos

Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Nos gustaría reconocer el trabajo de los profesionales que han colaborado con la red BEDCA en el desarrollo de una base de datos fidedigna y completa, así como el de las organizaciones participantes (universidades, agencias, federaciones y fundaciones), en especial:

- Agencia Española de Seguridad Alimentaria y Nutrición (AESAN).
- Universidad de Granada.
- Centre d'Ensenyament Superior de Nutrició i Dietética (CESNID).
- Universidad de Murcia.
- Universidad de Córdoba.
- Consejo Superior de investigaciones Científicas (CSIC).
    - Instituto de la Grasa de Sevilla.
    - Instituto del Frío de Madrid.
- Universidad Complutense de Madrid.
- Universidad Autónoma de Madrid.
    - Hospital Universitario Puerta de Hierro.
- Universidad de Barcelona.
- Federación Española de Industrias y Alimentación y Bebidas (FIAB).
- Fundación Triptolemos.

Inspiración

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

La composición nutricional de los alimentos (nutritions facts en inglés) proporciona información importante a los consumidores, de manera que pueden elegir una mejor nutrición que afecte de manera positiva a su salud.

Este tipo de datos, en conjunto con estudios médicos, puede ayudar a relacionar tipos de alimentos con enfermedades gracias a la minería de datos, relacionando componentes básicos a enfermedades concretas.

De manera específica estos datos sirven para:

- El comercio, la exportación y la legislación de los alimentos.
- El desarrollo de nuevos alimentos y platos.
- La producción, sostenibilidad y seguridad alimentaria.
- La investigación básica, epidemiológica y clínica.
- Clínicos y profesionales de la salud.
- Educadores y planificadores de políticas alimentarias.
- Ciudadanos que desean conocer en profundidad las características de los alimentos que ingieren.

Un ejemplo de estudio que hace uso de una BBDD similar, es el Proyecto Bacchus (http://www.eurofir.org/bacchus/), basado en la BBDD europea alimenticia EuroFir, este proyecto tiene como objetivo dar apoyo a los estudios que relacionan los beneficios de ingerir cierto tipo de componentes alimenticios con enfermedades cardiovasculares.

La recopilación de información nutricional realizada, así como el formato escogido, facilitan la explotación de estos datos por terceros (instituciones gubernamentales, empresas, ONGs, etc.) para la elaboración de sus propios estudios.

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

La licencia seleccionada para el dataset es la CC BY-NC_SA 4.0, que permite:

- Compartir: copiar y redistribuir el material licenciado en cualquier medio o formato.
- Adaptar: combinar, transformar y construir tomando como base el material licenciado.

Siempre y cuando se cumplan las siguientes condiciones:

- Se reconozca la autoría del material, proporcionando un enlace a la licencia e indicando si se realizaron cambios.
- Las nuevas obras derivadas de la original estén bajo una licencia con los mismos términos.
- No se utilice con fines comerciales.
- No se apliquen restricciones adicionales o medidas tecnológicas que limiten legalmente a otros realizar aquello que la licencia permite.

Se ha elegido esta licencia para el dataset porque se ajusta a las indicaciones facilitadas por el autor de la web que es fuente origen de la información, AESAN/BEDCA, que se resumen a continuación:

- La información recopilada se pone a disposición del público, pero no puede ser reproducida sin una clara indicación de la fuente original (AESAN/BEDCA Base de Datos Española de Composición de Alimentos v1.0).
- La reproducción, traducción o cualquier uso que no sea personal, educacional o no comercial de los datos en formato electrónico estará sujeto a la autorización expresa de AESAN/BEDCA.

En cuanto al código fuente, se ha seleccionado la licencia GNU AGPLv3, que es íntegramente una GNU GPL con una nueva cláusula que añade la obligación de distribuir el software si éste se ejecuta para ofrecer servicios a través de una red de ordenadores.

Código y dataset

Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.
Presentar el dataset en formato CSV.

El código fuente escrito así como el dataset obtenido pueden encontrarse en el siguiente repositorio de GitHub: <https://github.com/jmarquezd/NutriScraper>

Contribuciones

Apartados	Firma
Investigación previa	AGM, JMD
Contexto	AGM, JMD
Título del dataset	AGM, JMD
Descripción del dataset	AGM, JMD
Representación gráfica	AGM, JMD
Contenido del dataset	AGM, JMD
Agradecimientos	AGM, JMD
Inspiración	AGM, JMD
Licencia	AGM, JMD
Desarrollo de código	AGM, JMD
Generación del dataset	AGM, JMD
Bibliografía	AGM, JMD

Bibliografía

Brody, Hartley (2017). *The Ultimate Guide to Web Scraping* (1ª ed.). Victoria: Lean Publishing.

Lawson, Richard (2015). *Web Scraping with Python* (1ª ed.). Birmingham: Packt Publishing Ltd.

Mitchell, Ryan (2018). *Web Scraping with Python* (2ª ed.). Sebastopol: O'Reilly Media, Inc.

"Sobre las licencias CC" (noviembre de 2017). *Creative Commons*. [artículo en línea]. [Fecha de consulta: 5 de abril de 2019]
<https://creativecommons.org/licenses/>

"Red BEDCA" (julio de 2009). *AESAN/BEDCA Base de Datos Española de Composición de Alimentos*. [artículo en línea]. [Fecha de consulta: 25 de marzo de 2019]
<http://www.bedca.net/>