UNIVERSITAT OBERTA DE CATALUNYA

Web scraping

Tipología y ciclo de vida de los datos

Isabel González Valle;Cristina Merino García de la Reina

Abril 2020

Contenido

1. Contexto

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

Se ha recolectado información de los productos sin gluten que oferta el supermercado Eroski aprovechando que en su web ofrece la posibilidad de visualizar todos los productos que no contienen esta proteína.

Eroski en su supermercado online identifica ciertas marcas para agrupar productos con unas características específicas:

Elegimos extraer la información de todos los alimentos sin gluten que este supermercado online ofrece a sus clientes para poder generar un conjunto de datos que permita conocer la lista de alimentos que se podrían comprar en esta web y de cara a que puede servirnos para realizar comparativas con otros supermercados o crear listas de la compra para personas celiacas o con cierta intolerancia o sensibilidad al gluten.Nos gustaría hacer hincapié en que la web elegida proporciona la lista de alimentos sin gluten de todos los ámbitos de la alimentación (carnes, lácteos, bollería…) y no se centra solamente en productos que tradicionalmente se elaboran con gluten como puede ser la pasta o el pan, como ocurre en la mayoría de secciones sin gluten de otros supermercados, donde la etiqueta “sin gluten” solo aparece en productos específicos.

2. Definir un título para el dataset

Elegir un título que sea descriptivo.

El título que hemos elegido para este conjunto de datos es: “AlimentosSinGluten”. En principio no se valora incluir el nombre del supermercado del que se extrae la información, puesto que este dataset podría ser ampliado realizando web scraping de otras páginas con información de este tipo de productos.

3. 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 conjunto de datos extraído contiene información de detalle de los productos sin gluten que podemos encontrar en este supermercado, tales como datos relativos a cómo se presenta el producto al público, el precio, si el producto se encuentra o no en oferta, valor nutricional y valoración de los clientes para el producto en concreto.

En principio, no se identificar en el título del dataset el supermercado online accedido, puesto que esta información podría llegar a ampliarse con la información de otros supermercados online.

Del conjunto de datos podemos extraer la siguiente información:

- Product description: Nos ofrece la descripción del producto (nombre + cantidad)
- Product name: Nombre del producto
- Product quantity: Cantidad del producto
- Kilo evaluation: Indica el precio del kilo
  Offer description: Muestra el precio actual del producto y en caso de estar en oferta, también muestra el precio anterior.

```
1|Filetes de lomo adobado de cerdo extrafino EROSKI, bandeja 300 g | Filetes de lomo adobado de cerdo extrafino EROSKI|bandeja 300 g|None|1 KILO A|12,83 €|4,5||Ahora 3,85 €

2|Carne picada de ternera EUSKO LABEL SELEQTIA, bandeja 500 g | Carne picada de ternera EUSKO LABEL SELEQTIA|bandeja 500 g|None|1 KILO A|12,80 €|4,6||Ahora 6,40 €

3|Picada mixta cerdo-vacuno burger meat EROSKI basic, bandeja 1 kg | Picada mixta cerdo-vacuno burger meat EROSKI basic|bandeja 1 kg|None|NaN|NaN|4,0|5,30 € Antes|Ahora 4,99 €
```

4. Representación gráfica

Presentar una imagen o esquema que identifique el dataset visualmente.

Ilustración 1 - Lineal productos sin gluten

(https://cronicaglobal.elespanol.com/, 2019)

5. Contenido

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

Del conjunto de datos podemos extraer la siguiente información (los nombres de las variables no son definitivos):

- **Artículo**: Nos ofrece la descripción del producto (nombre + cantidad)
- **Nombre**: Identifica el producto. Es la descripción del producto.
- **Presentación**: Indica el formato en el que se presenta el producto. Puede ser peso o unidades.
- **Nutri-Score**: Etiquetado internacional que indica el valor nutricional de los alimentos, se representa con 5 colores, desde el verde al rojo, asociados a 5 letras, de la A a la E.
- **Cantidad Base**: Peso base de referencia (Kilo, Litro, unidad, …)
- **Precio Cantidad Base**: Precio de la cantidad base para poder comparar en el mismo producto o tipo de producto los precios en diferentes presentaciones.
- **Valoración**: Puntuación dada por los consumidores (del 1 al 5)
- **Precio Anterior**: Si el producto se encuentra en oferta, se muestra el precio anterior.
- **Precio Actual**: Precio de venta actual.
- **Fecha Extracción**: Fecha en la que se realiza el volcado de datos.

Como este dataset tiene dependencia temporal se incluye un campo fecha que va a identificar la fecha de la extracción de los datos.

6. Agradecimientos

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

El propietario del conjunto de datos es EROSKI. S. COOP.

El Grupo Eroski es una empresa cooperativa de distribución de las más importantes de España perteneciente a la Corporación Mondragón. Tiene su sede en la localidad vizcaína de Elorrio en el País Vasco, España.

Se fundó en el año 1969 y es una de las empresas de distribución más importante de España contando con una plantilla de más de 35 000 trabajadores repartidos por toda España. La empresa cuenta con alrededor de 2000 establecimientos de diferentes marcas, entre las que se incluyen los hipermercados "Eroski", supermercados "Eroski City" y "Eroski Center", supermercados "Eroski Merca", supermercados "Cash Record", supermercados "Caprabo", supermercados "Familia", autoservicios "Aliprox", "Eroski Viajes", "Viajes Caprabo", "Eroski Óptica", "Estaciones de Servicio Eroski" y "Tiendas de Deporte FORUM". (Wikipedia, s.f.)

7. Inspiración

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

La Enfermedad Celíaca (EC) es una enfermedad multisistémica con base autoinmune provocada por el gluten. Se estima que el 1% de la población la padece aunque diversos estudios indican que sobre el 75% de los casos están sin diagnosticar.

A día de hoy no existe cura para esta enfermedad, por lo que el único tratamiento para este colectivo es seguir una dieta sin gluten de por vida.

Aunque cada vez es más fácil encontrar alimentos sin gluten, no todos los supermercados disponen de un etiquetado reconocible para los productos que no son elaborados específicamente para los celiacos, por lo hacer la compra no resulta una tarea sencilla.

Además de la falta del etiquetado sin gluten, debemos sumar otro problema más: el precio.

La Federación de Asociaciones de Celiacos de España, más conocida por sus siglas FACE, elabora cada año un informe de precios en el que se muestra con carácter semanal, mensual y anual el gasto extra que supone para una persona celíaca seguir la dieta sin gluten.

Las conclusiones del informe de precios 2020 son:

> *"Teniendo como base los resultados obtenidos se puede concluir que una familia con una persona celiaca entre sus miembros, que tenga un patrón alimentario con aporte calórico de 2000 a 2200 kcal, tendrá un* ***incremento estimado en la adquisición de la cesta de la compra de 18,97€ a la semana, 75,89 € al mes, y de 910,73 € al año****, en relación con otra familia que adquiera productos con gluten."* (FACE, 2020)

Por lo tanto, nuestro conjunto de datos pretende facilitar la compra al colectivo celiaco identificando qué productos son los que pueden consumir con total tranquilidad e informando de su precio posibilitando así la realización de comparativas con otros supermercados.

8. 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

El propietario de los datos es EROSKI.

https://www.eroski.es/terminos-y-condiciones-de-uso/

La licencia que se ha elegido para la creación y publicación del conjunto de datos es la CC BY-NC-SA 4.0 por ser la que, bajo nuestro punto de vista, más se ajusta al trabajo que se ha realizado.

- CC BY-NC-SA 4.0
  Esta licencia 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 condiciones idénticas.

9. Código

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

10. Dataset

Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

11. Entrega

Presentar el trabajo con el DOI del dataset en Github.

12. Contribuciones al trabajo

Contribuciones	Firma
Investigación previa	IGV, CMGR
Redacción de las respuestas	IGV, CMGR
Desarrollo código	IGZ, CMGR

Bibliografía

Eroski. (s.f.). *https://supermercado.eroski.es/*. Obtenido de https://supermercado.eroski.es/es/supermercado/SinGluten/

FACE. (2020). *Informe de precios sobre productos sin gluten 2020.* Madrid: Federación de Asociaciones de Celiacos de España (FACE).

https://cronicaglobal.elespanol.com/. (27 de 05 de 2019). *https://cronicaglobal.elespanol.com/*. Obtenido de https://cronicaglobal.elespanol.com/vida/gluten-free-pueblos-ciudades-pequenas_247779_102.html

Wikipedia. (s.f.). Obtenido de https://es.wikipedia.org/wiki/Eroski