1. Extracción de un dataset de Ratings de Estandard de Higiene en los restaurantes de Belfast, Irlanda del Norte

2. Introducción

He creado un conjunto de datos de los Ratings de Food Hygiene Standards utilizando los datos disponible en el web del Foods Standards Agency (FSA) UK. El GUI del web permite busquedas con parametros de Ciudad, Codigo Postal, tipo del comercio, nombre del comercio. Se puede hacer un injección de parametros de busqueda en el URL. En este caso he seleccionado datos de restaurantes en la ciudad de Belfast, Irlanda del Norte.

3. Imagen identificativa

Logo del los ratings del FSA

4.Contexto

Los datos se tratan de un rating de inspecciones de la higiene de los restaurantes. Las inspecciones suelen de repetirse al mínimo cada 2 años. Los inscpectores hacen un informe y el informe tiene una clasificación (rating)

5. Contenido

He extraido los siguientes campos.
Restaurant Name – Nombre del restaurante
Post Code – código postal del restaurante
Rating Text – texto del rating del nivel de higiene del restaurante 1 peor- 5 mejor
Inspection Date – fecha de la inspección en formato “dd Mon yyyy”
Rating Number – numer del rating del nivel de higiene del restaurante 1-5

Los datos se refrescan en cuanto se haga una nueva inspección del restaurante. El periodo estaría del orden de meses.

He recogido los datos de todos los restaurantes de Belfast y lo más viejo es de 2013, aunque 98% de los datos son de 2017 y 2018.

6. Agradecimientos

El propietario del la base de datos es el Food Standards Agency del gobierno británico (https://www.food.gov.uk/). Tiene una base de datos on-line con Web interface y permiten acceso via la licencia (Open Government Licence) http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/. Esta licencia es similar a CC-BY-SA-4.0 y permite reproducción de los satos.

7. Inspiración

Estoy muy interesado en analisar las opiniones de restaurantes pero me asustó leer los Terms and Condicions de Tripadvisor que explícitamente dice que web-scraping esta prohibido. Así que encontré algo similar pero un poco diferente – datos sobre la calidad de higiene en los restaurantes. Si obtengo permisión escrita de Tripadvisor me gustaría hacer una comparación de los "bubbles" de TripAdvisor y los ratings de higiene de los restaurantes.

8. Licencia

He seleccionado "Released Under CC BY-SA 4.0 License" porque la licencia original de los datos (Open Government Licence) (OGL) estipula:

You are free to:

- copy, publish, distribute and transmit the Information;
- adapt the Information;
- exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your own product or application.

You must (where you do any of the above):

- acknowledge the source of the Information in your product or application by including or linking to any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence;

Y así es muy similar a lo de CC-BY-SA-4.0. Además el OGL dice que es compatible con el Creative Commons Attribution License 4.0.

9. Código

El código de Python del extrato de los datos esta incluida en este mismo repositorio de GitHub dentro del folder "code".

10. Dataset

El dataset está incluido en este mismo repositorio de Github en formato CSV.

Recursos

Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC
Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd.