Eating, Drinking & Data Mining

# Contexto

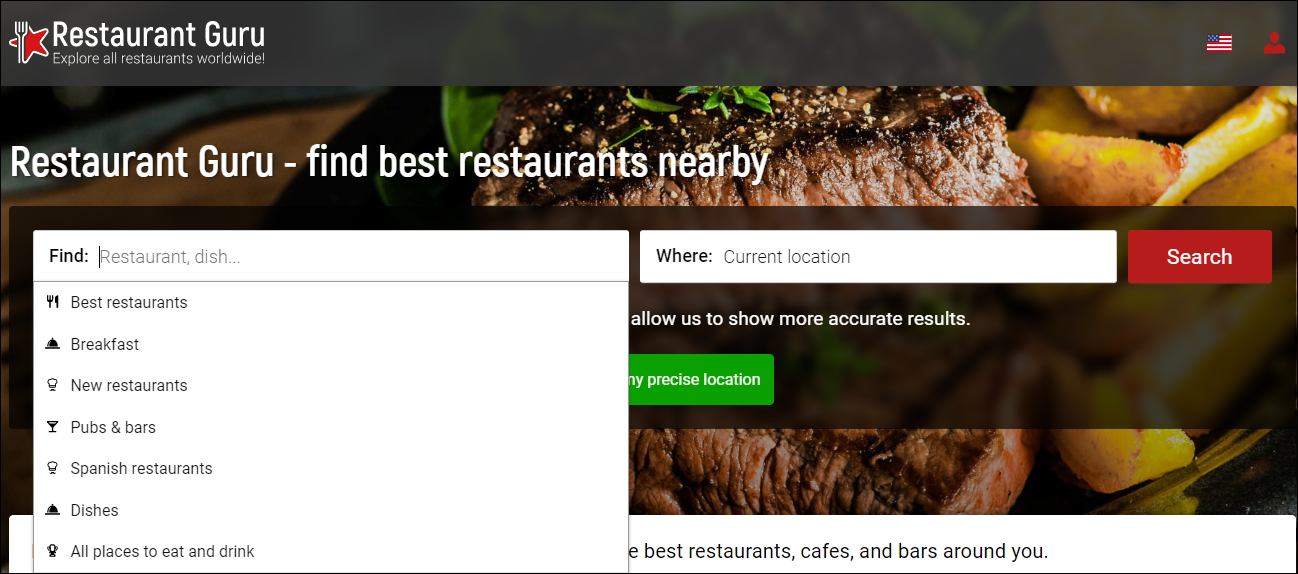
El negocio de la restauración es uno de los principales motores de la economía en España. Los restaurantes, bares, tabernas, cafeterías, mesones, pubs, etc. proliferan por cualquiera de nuestras ciudades y pueblos. Comer y beber en estos lugares públicos es parte de nuestra cultura.

Existen muchos agregadores en la web dónde podemos buscar este tipo de negocios, en estos agregadores se muestran datos que permiten al usuario tomar una decisión del tipo: ¿Dónde voy a comer? ¿Qué restaurante está más cerca de esta calle? ¿Dónde puedo comer comida japonesa? ¿Cuánto me va a costar? ¿Puedo reservar? ¿Qué opinan otros usuarios de este restaurante? ¿Es accesible para minusválidos? ¿Les gustará a los niños?, etc. Hay un sinfín de preguntas que se pueden responder con estos datos.

La aplicación eating, drinking & data mining genera, a partir de un municipio seleccionado por el usuario, una dataset conteniendo todos los negocios de restauración dicho municipio. Esta dataset contiene, para cada restaurante, información básica de contacto del restaurante: {nombre, dirección, teléfono, website}, así como información avanzada para la toma de decisiones: {tipo de comida, horario de apertura, rango de precios, número de opiniones en redes sociales, ratings, así como un inventario de las palabras más mencionadas en las reseñas de sus usuarios}

El potencial de estos datos es enorme, de ahí el término Data Mining en el nombre de la aplicación. Con técnicas de minería de datos podemos agrupar restaurantes por precio, por tipo de comida, y clasificarlos en función de sus reseñas. Podemos predecir cuál será el éxito del negocio en función del tipo de comida. Podemos ver cuál es la correlación entre las horas de apertura y el número de reseñas, etc. El abanico de preguntas que se puede responder con esta dataset es enorme.

Para la recolección de datos sobre restaurantes hemos utilizado la página [restaurantguru](https://restaurantguru.com/). En esta página podemos buscar por distintos tipos de negocio de restauración en cualquier parte del mundo. Se trata de un agregador de contenidos cuya principal ventaja es que aglutina información sobre reseñas y ratings de otros agregadores populares como Facebook, Google, Yelp, etc.



Una vez seleccionados estos dos parámetros (¿Qué? y ¿Dónde?), restaurantguru nos lista todos los negocios que cumplen las condiciones indicadas. Y para cada uno de estos, nos proporciona:

* Nombre
* Clasificación según usuario final y tipo de comida en el municipio seleccionado
* Tipo de comida
* Horario de apertura
* Teléfono
* Link a la carta del restaurante
* Rango de precio
* Dirección
* Fotografías
* Descripción del restaurante
* Términos más mencionados en las reseñas
* Ratings en redes sociales: Yelp, Foursquare, Google, Tripadvisor y Facebook
* Número de reseñas en redes sociales
* Horarios de apertura
* Website
* Servicios del restaurante
* Reseñas

Por motivos legales y de cumplimiento de las restricciones de esta página, no todos los atributos formarán parte de la dataset.

## Título

El nombre de la dataset generada se compone de estos términos:

<EDDM>\_<ISO-3>\_<XXX>

Dónde:

* EDDM – es el acrónimo de Eat, Drink & Data Mining
* ISO -3 es el código [ISO 3166](https://unstats.un.org/unsd/tradekb/Knowledgebase/50347/Country-Code) que regula los códigos de país
* XXX – es el nombre del municipio.

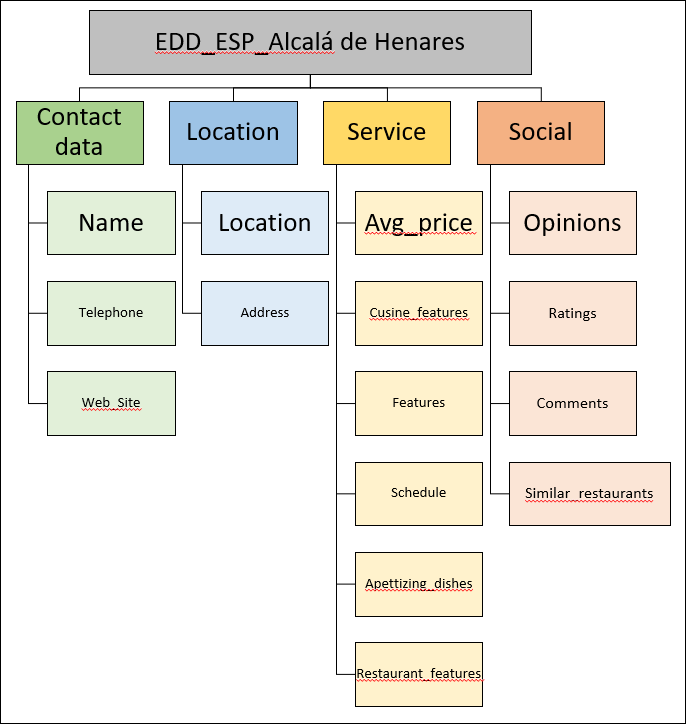
A efectos de esta práctica, generaremos la dataset:

**EDDM\_ESP\_Alcala\_de\_Henares**

# Descripción del Dataset

Este dataset contiene 979 registros que se corresponden con todos los restaurantes en Alcalá de Henares. Para cada restaurante, la dataset ofrece sus datos de localización y contacto, datos del tipo de servicio ofrecido por el mismo, así como del impacto del restaurante en redes sociales.

# Representación gráfica



# Contenido

Este dataset se compone de los siguientes campos:

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre campo | Descripción | Tipo de datos | Ejemplo |
| Name | Nombre del negocio de restauración | String | Texas Burger Alcalá |
| avg\_price | Precio medio por persona | String | Rango de precios por persona 11€-20€ |
| cuisine\_features | Tipo de comida | String | Comida rápida, americana, opciones vegetarianas |
| telephone | Teléfono | String | <tel:+34912801333> |
| address | Dirección postal | String | Rda. De la Pescadería 5, Alcalá de Henares |
| location | Dirección postal completa | String | Rda. De la Pescadería 5, Alcalá de Henares, Comunidad de Madrid, España |
| opinions | Opinión media por tipo de comida | Json | {“rape”:”excellent”, “pescado”: “excellent”} |
| ratings | Valoración media en cada red social | Json | {“Foursquare”:”Sin valorar aún”, “Google”:”4.4/5”} |
| comments | Número de reseñas en cada red social | Json | {“Yelp”:”1”,”Foursquare”:”2”} |
| features | Facilidades del rsturante | String | Tarjetas de crédito aceptadas, Entrega, Asientos al aire libre, … |
| schedule | Horario de apertura | Json | {“Lunes”:”cerrado”, “Martes”:”12.00-16.00, 19:30-00:00”…} |
| apettizing\_dishes | Platos del restaurante | String | Rape, pescado, hamburguesas, … |
| restaurant\_features | Otras facilidades del restaurante | String | Comida para llevar, estupenda ubicación, pedir comida a domicilio,… |
| similar\_restaurants | Otros restaurantes similares | String | Buddy Hollys, Dulcinea, Restaurante Oh la La … |
| web\_site | Página web | String | http://texasburgueralcala.com/ |

Los datos de esta dataset se han obtenido en bruto con la técnica de web scraping. Estos datos requieren de limpieza para su adecuación a posteriores procesos de análisis.

La fecha de adquisición de los datos fue el 16 de Octubre de 2021.

# Agradecimientos

Nuestro agradecimiento a la página web [restaurantguru.com](https://restaurantguru.com/) de la que pertenecen todos los datos extraídos en esta práctica.

Restaurantguru es un metabuscador de restaurantes que ayuda al usuario a elegir el restaurante adecuado en cualquier ciudad.

Hemos realizado una evaluación inicial de este website, con el objetivo de adecuarnos a las restricciones impuestas por esta página web y reducir así las posibilidades de ser bloqueados. Para ello hemos seguido estos pasos:

Análisis del archivo robots.txt

El archivo robots.txt contiene las restricciones a tener en cuenta al rastrear las páginas web.

En este archivo aparecen una serie de User-Agents con unas restricciones específicas para: Googlebot, Googlebot-Mobile, Mediapartners-Google, Yandex, y PetalBot. Para el resto de user-agents se indican las siguientes restricciones:

User-agent: \*

Crawl-delay: 1

Disallow: /image/\*

Disallow: /web/\*

Disallow: /link/\*

Disallow: /slink/\*

Disallow: /ajax/\*

Disallow: /user/\*

Disallow: /login/\*

Disallow: /compare/\*

Disallow: /search\*

Disallow: /\*/reviews\*

Disallow: /\*/emotions\*

Disallow: /\*/load-menu

Disallow: /get-widget\*

Disallow: /geohelp\*

Disallow: /emotions?h=\*

Disallow: /amp/\*

Disallow: /malls/\*

Disallow: /islands/\*

Disallow: /desktop\_init\*

Disallow: /widget-landing?id=\*

Disallow: /widget-awards?id=\*

Disallow: /comment-widget-preview?\*

Disallow: /qr-menu-landing\*

Disallow: /widget-demo\*

Se nos pide un retardo de 1 segundo entre llamadas, y se nos restringe el acceso a todas estas direcciones.

Nuestra aplicación cumple con las restricciones impuestas por la página web. (LO PODEMOS HACER MEDIANTE CODIGO CON ESTO… o dejarlo como está dicho)

Import robotparser

rp = robotparser.RobotFileParser()

rp.set\_url(‘https://restaurantguru.com/robots.txt

rp.read()

while crawling …

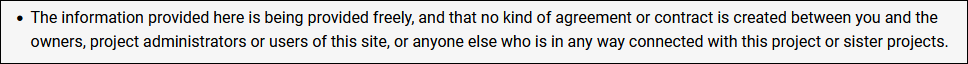
if rp.can\_fetch(user\_agent, url):

....

else :

print (‘Blocked by robots.txt’,url)

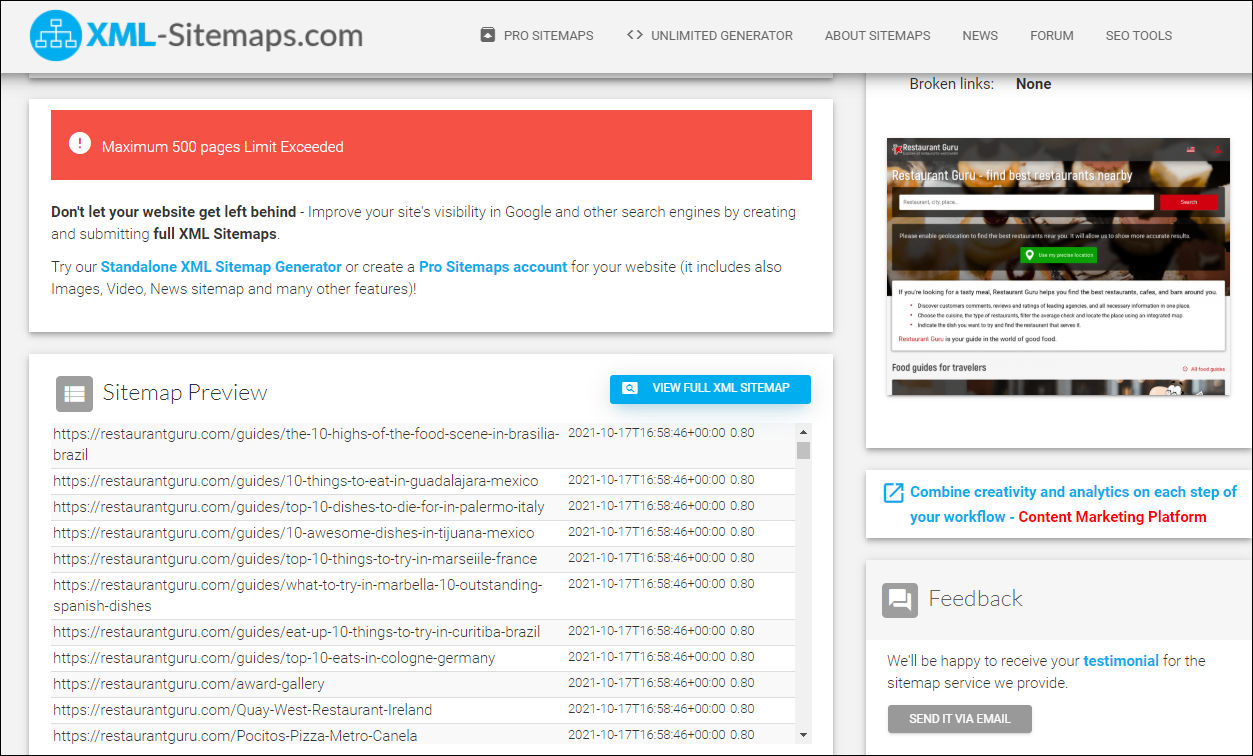
Es importante destacar que entre los [términos y condiciones](https://restaurantguru.com/disclaimer) de esta página web se explicita que la información proporcionada es libre, y no está sujeta a ningún tipo de acuerdo o contrato.



Análisis del mapa del sitio web.

Se ha analizado la estructura de la web utilizando [xml-sitemaps](https://www.xml-sitemaps.com/). El análisis del mapa de la página nos permite localizar el contenido que buscamos sin tener que rastrear todas las páginas que componen el sitio.

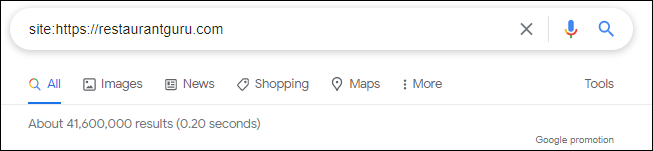
En el caso de esta página web cuenta con más de 500 páginas por lo que se necesita una licencia de pago para obtener el mapa completo del sitio. El tamaño del sitio es el esperado, ya que existe al menos una página por cada restaurante almacenado.



## Tamaño de la página web.

La manera más eficaz de estimar el tamaño de la web es buscándola en Google con la palabra clave site.

Se trata de un sitio bastante grande con más de 41 millones de páginas indexadas.



## Tecnología.

La tecnología utilizada por el sitio web condicionará el tipo de web scraping utilizado. Se utiliza la herramienta builtwith de Python para saber más acerca de la tecnología de esta web.



Sin embargo, no obtenemos muchos detalles sobre la tecnología que usa esta web.

* Nginx es un servidor web/proxy de alto rendimiento, que se usa como proxy inverso, cache de HTTP y balanceador de carga.
* Google AdSense permite monetizar una página web mediante la colocación de anuncios en sitios web.

## Propietario.

Conocer el propietario de una página web es interesante para saber si es conocido por bloquear los procesos de web scraping. Para ello usamos la librería whois de Python.

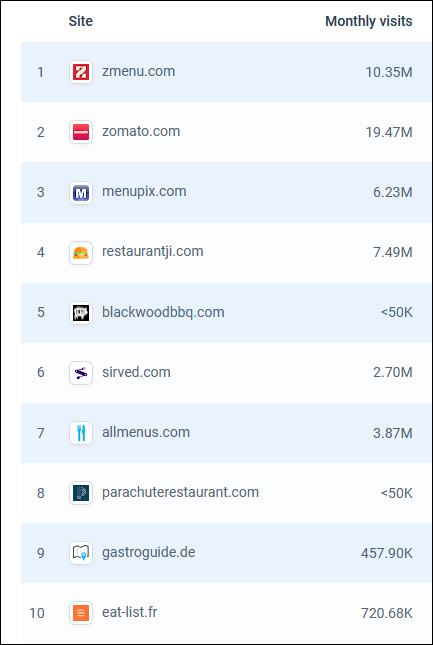




## Web similares.

Se han identificado otras páginas web similares a restaurantguru.com

La principal diferencia entre ellas es el número de visitas, pero consideramos que restaurantguru.com cumple con las necesidades de calidad y consistencia de los datos que venimos buscando.



## Calidad de los datos.

Los datos que hemos examinado de restaurantguru.com ofrecen garantías en las 6 dimensiones especificadas por la International Data Management Association del Reino Unido, estas son:

|  |  |  |
| --- | --- | --- |
| Dimensión | Descripción | Resultado |
| Completitud | Proporción de datos almacenados frente al potencial de datos completos | 80% de celdas con valores nulos en el total del dataset |
| Unicidad | No existencia de duplicados | 100% de registros no duplicados |
| Puntualidad | Grado en que los datos representan la realidad desde el punto requerido en el tiempo | 100% de registros actualizados. El chequeo se hace comprobando el número de reseñas de Google en una muestra reducida |
| Validez | Los datos se ajustan a la sintaxis de su definición | 100% validos |
| Exactitud | Los datos se ajustan a la realidad que están describiendo | No se ha comprobado porque requiere trabajo de campo |
| Consistencia | Ausencia de diferencias con respecto a su definición | Errores despreciables, solo se han detectado un par de teléfonos con una sintaxis diferente |

## Otros estudios

Existen numerosos estudios y artículos en la red sobre el uso de análisis de datos en restaurantes.

Podemos destacar entre otros:

* Rahman, Md. Muminur. (2017). IoT & Big Data Based Applications for Restaurant Questions of Opportunities, Challenges, Benefits & Operations. 10.13140/RG.2.2.35231.05287.
* Allika, Krishnakanth, Insights into Toronto’s Foodservice Market using Data Science Tools (August 08, 2019). Available at SSRN: <https://ssrn.com/abstract=3652350> or [http://dx.doi.org/10.2139/ssrn.3652350](https://dx.doi.org/10.2139/ssrn.3652350)

# Inspiración

Este conjunto de datos es interesante porque mezcla datos geográficos de localización, con datos de servicios proporcionados y con datos de sentimientos de sus usuarios. Esto nos permite mediante técnicas de minería de datos responder a múltiples preguntas, entre ellas:

* ¿Qué restaurante es más valorado por tipo de comida?
* ¿Qué tipo de comida tiene restaurantes más valorados?
* ¿Cuáles son los servicios más comunes en los restaurantes más valorados?
* ¿Cuáles son los servicios más comunes en los restaurantes más valorados?
* ¿Cuál es el tiempo medio de apertura en los restaurantes más valorados?
* ¿Existe correlación entre el horario de apertura y la valoración de un restaurante?
* ¿Cuál es la correlación entre el precio y la valoración de un restaurante?
* ¿Existe correlación entre el precio de un restaurante y su localización?
* ¿Cuáles son los términos más mencionados en las reseñas?

# Licencia

El software está publicado con licencia [MIT](https://mit-license.org/).

Se ha seleccionado esta licencia porque es sencilla y permite al usuario hacer cualquier cosa con el software licenciado mientras mantengan la copia de la licencia, incluyendo el aviso de copyright.

Esta licencia impone muy pocas limitaciones en la reutilización del software, permite:

* Uso comercial
* Distribución
* Modificación
* Uso privado del contenido licenciado.

La dataset se ha publicado en Zenodo con licencia [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/legalcode). Esta licencia tiene estas características:

* Es una licencia de carácter global, gratuita, no transferible a terceros, no exclusiva para reproducir y compartir el material licenciado, en su totalidad o en parte; y para producir, reproducir y compartir material adaptado.
* No es aplicable donde se apliquen excepciones y limitaciones al uso del material licenciado.

# Código

El código se encuentra disponible en este repositorio [Git](https://github.com/gromerof1974/Eating-Drinking-Data-Mining).

# Dataset

Se encuentra publicada en [Zenodo](https://zenodo.org/record/5574493#.YWyA7t_iuUl).

# Contribución

La presente tabla confirma la participación de los dos integrantes del grupo en todas las actividades realizadas en esta práctica.

|  |  |
| --- | --- |
| Contribuciones | Firma |
| Investigación previa | BLB, GRF |
| Redacción de las respuestas | BLB, GRF |
| Desarrollo del código | BLB, GRF |