

Eating, Drinking & Data Mining

Tabla de contenido

1. Contexto	2
2. Título	3
3. Descripción del Dataset.....	4
4. Representación gráfica.....	4
5. Contenido	5
6. Agradecimientos	7
6.1 Análisis del archivo robots.txt	7
6.2 Análisis del mapa del sitio web	8
6.3 Tamaño de la página web	8
6.4 Tecnología	8
6.5 Propietario.....	9
6.6 Webs similares	9
6.7 Calidad de los datos	10
6.7 Otros estudios	11
7. Inspiración	12
8. Licencia	12
9. Código.....	13
src/scrapper.properties.....	13
src/main.py.....	14
Función main	14
src/scrapper.py.....	14
Función scrape	15
Función get_links.....	15
Función get_page	16
Función get_page_data.....	16
10. Dataset	17
Contribución.....	17

1. Contexto

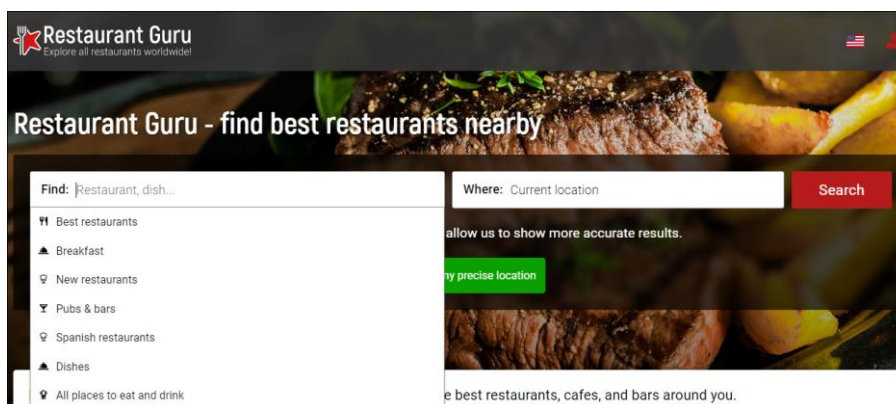
El negocio de la restauración es uno de los principales motores de la economía en España. Los restaurantes, bares, tabernas, cafeterías, mesones, pubs, etc. proliferan por cualquiera de nuestras ciudades y pueblos. Comer y beber en estos lugares públicos es parte de nuestra cultura.

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 hoy?, ¿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 sinnúmero 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, un dataset que contiene todos los negocios de restauración de dicho municipio. Este dataset guarda, para cada restaurante, información básica de contacto (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, o 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 este dataset es enorme.

Para la recolección de datos sobre restaurantes hemos utilizado la página 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 seleccionada la ciudad a analizar, restaurantguru.com nos lista todos los negocios de restauración en la ciudad. Y para cada uno de estos, nos proporciona:

- Nombre
- Clasificación según usuario final y tipo de comida
- Tipo de comida
- Horario de apertura
- Teléfono
- Link al menú 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 del dataset. Las reseñas, por ejemplo, quedan excluidas.

2. Título

El título del dataset es Eating, Drinking & Data Mining.

Este título aglutina:

- El propósito del dataset: Data Mining
- El objeto del dataset: Eating and Drinking.

La aplicación permite generar un dataset cada vez que se ejecuta. Con el objeto de identificar de forma unívoca los datasets generados, se ha definido una regla para nombrarlos. El nombre del dataset generado se compone de estos términos:

EDDM_<XXX>

Dónde:

- EDDM – es el acrónimo de Eat, Drink & Data Mining
- XXX – es el nombre del municipio.

A efectos de esta práctica, seleccionamos Alcalá de Henares como municipio de análisis, por lo que generamos el dataset:

EDDM_Alcala_de_Henares

3. Descripción del Dataset

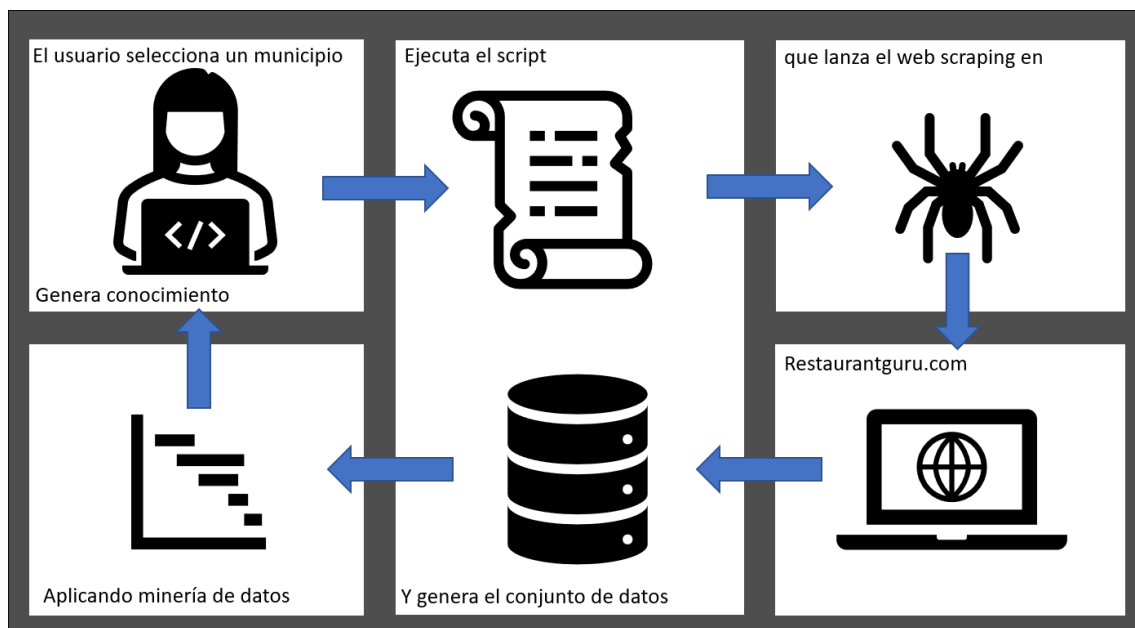
Este dataset contiene 1508 registros que se corresponden con todos los restaurantes en Alcalá de Henares. Para cada restaurante, el dataset ofrece sus datos de localización y contacto, datos del tipo de servicio ofrecido por el mismo, así como el impacto del restaurante en redes sociales.

Localización y contacto	Tipo de servicio	Impacto en Redes sociales
Nombre	Precio medio	Reseñas
Teléfono	Tipo de comida	Valoraciones
Dirección	Servicios generales	Comentarios
Localización	Servicios del restaurante	Restaurantes similares
Horario de apertura	Especialidades culinarias	
Página web		

El dataset generado con técnicas de web scraping requiere de limpieza y conversión de datos para su posterior tratamiento con técnicas de minería de datos.

4. Representación gráfica

El proceso de Eating, Drinking & Data Mining es sencillo. El usuario selecciona un municipio, corre el script, que llama al crawler que hará web scraping en restaurantguru.com. El resultado de este proceso es la dataset EDD_Alcala_de_Henares. La aplicación de técnicas de minería de datos sobre este dataset permitirá que el usuario adquiera el conocimiento deseado.



5. 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	Valoración media por tipo de comida	Json	{"rape":"excellent", "pescado": "excellent"}
ratings	Valoración media en redes sociales	Json	{"Foursquare":"Sin valorar aún", "Google":"4.4/5"}
comments	Número de reseñas en redes sociales	Json	{"Yelp":"1","Foursquare":"2"}
features	Características y facilidades del restaurante	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 recomendados del restaurante	String	Rape, pescado, hamburguesas, ...
restaurant_features	Otras características y 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 este 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 18 de Octubre de 2021.

A continuación se presenta una propuesta de conversión y limpieza:

Nombre campo	Tipo de datos	Conversión
name	String	No
avg_price	String	Si, categórico por rangos
cuisine_features	String -> Boolean	Si, este campo se parseará de manera que para cada feature se creará un nuevo campo que tomará valores booleanos
telephone	String	Si, formato específico para teléfono
address	String	No
location	String	No
opinions	Json -> Numeric	Si, este campo se parseará de manera que para cada tipo de comida se creará un nuevo campo que tomará valores numéricos de valoración
ratings	Json -> Numeric	Si, este campo se parseará de manera que para cada red social se creará un nuevo campo que tomará valores numéricos de valoración
comments	Json -> Numeric	Si, este campo se parseará de manera que para cada red social se creará un nuevo campo que tomará valores numéricos con el número de reseñas
features	String -> Boolean	Si, este campo se parseará de manera que para cada feature se creará un nuevo campo que tomará valores booleanos
schedule	Json -> String & Numeric	Si, este campo se parseará de manera que para cada día de la semana se creará un campo con la indicación de su horario de apertura, y se creará otro campo por cada día de la semana con las horas totales de apertura por día
apettizing_dishes	String -> Boolean	Si, este campo se parseará de manera que para cada plato recomendado se creará un nuevo campo que tomará valores booleanos
restaurant_features	String -> Boolean	Si, este campo se parseará de manera que para cada feature se creará un nuevo campo que tomará valores booleanos
similar_restaurants	String	No
web_site	String	No

6. Agradecimientos

Nuestro agradecimiento a la página web restaurantguru.com, a la que pertenecen todos los datos extraídos en esta práctica.

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

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.

A continuación, se detallan los pasos que se han seguido.

6.1 Análisis del archivo robots.txt

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

En el archivo [robots.txt](#) de restaurantguru.com, aparecen una serie de User-Agents para los que se aplican restricciones: 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 (Crawl-delay: 1), y se nos restringe el acceso a los directorios marcados como Disallow .

Nuestra aplicación cumple con las restricciones impuestas por la página web. El código tiene un retardo aleatorio de entre 1 y 2 segundos.

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

- The information provided here is being provided freely, and that no kind of agreement or contract is created between you and the owners, project administrators or users of this site, or anyone else who is in any way connected with this project or sister projects.

6.2 Análisis del mapa del sitio web

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.

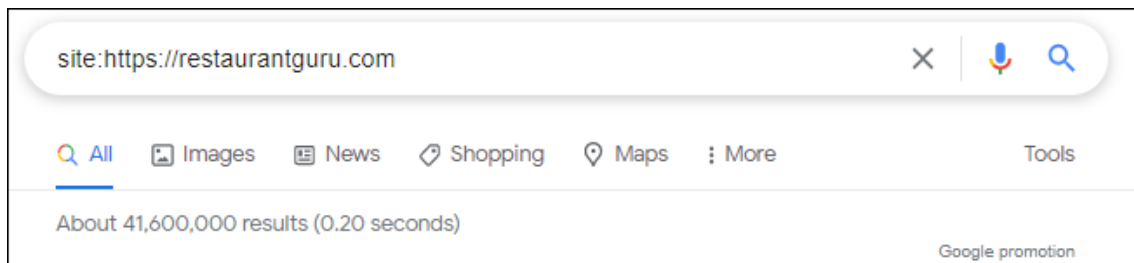
Para el dominio restaurantguru.com no hemos encontrado su sitemap al final del fichero robots.txt, ni tampoco en las siguientes direcciones:

- <https://restaurantguru.com/sitemap.xml>
- <https://restaurantguru.com/sitemap-index.html>
- https://restaurantguru.com/sitemap_index.html

6.3 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.



6.4 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.

```
In [7]: 1 builtwith.builtwith('https://restaurantguru.com/')
Out[7]: {'web-servers': ['Nginx'], 'advertising-networks': ['Google AdSense']}
```

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.

6.5 Propietario

Conocer el propietario de una página web es interesante para saber si es conocido por bloquear los procesos de web scraping. Buscamos el propietario de restaurantguru.com usando la librería whois de Python.











```
1 import whois
2 print(whois.whois('https://restaurantguru.com'))
```

```
{
  "domain_name": [
    "RESTAURANTGURU.COM",
    "restaurantguru.com"
  ],
  "registrar": "GoDaddy.com, LLC",
  "whois_server": "whois.godaddy.com",
  "referral_url": null,
  "updated_date": [
    "2021-09-08 12:29:11",
    "2019-05-15 19:46:48"
  ],
  "creation_date": [
    "2001-09-23 00:14:00",
    "2001-09-22 19:14:00"
  ],
  "expiration_date": [
    "2022-09-07 11:59:59",
    "2022-09-07 06:59:59"
  ],
  "name_servers": [
    "NS-1471.AWSDNS-55.ORG",
    "NS-1865.AWSDNS-41.CO.UK",
    "NS-283.AWSDNS-35.COM",
    "NS-696.AWSDNS-23.NET"
  ],
  "status": [
    "clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited",
    "clientRenewProhibited https://icann.org/epp#clientRenewProhibited",
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited"
  ],
  "emails": [
    "abuse@godaddy.com",
    "restaurantguru.com@domainsbyproxy.com"
  ],
  "dnssec": "unsigned",
  "name": "Registration Private",
  "org": "Domains By Proxy, LLC",
  "address": [
    "DomainsByProxy.com",
    "2155 E Warner Rd"
  ],
  "city": "Tempe",
  "state": "Arizona",
  "zipcode": "85284",
  "country": "US"
}
```

6.6 Webs 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.

	Site	Monthly visits
1	 zmenu.com	10.35M
2	 zomato.com	19.47M
3	 menupix.com	6.23M
4	 restaurantji.com	7.49M
5	 blackwoodbbq.com	<50K
6	 sirved.com	2.70M
7	 allmenus.com	3.87M
8	 parachuterestaurant.com	<50K
9	 gastroguide.de	457.90K
10	 eat-list.fr	720.68K

6.7 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	Chequeo	Resultado
Compleitud	Proporción de datos almacenados frente al potencial de datos completos	Número de celdas con valores no nulos / total de celdas en el dataset	76% de celdas atribuidas (5399 celdas nulas en un total de 22620 celdas)
Unicidad	No existencia de duplicados	El uso de diccionarios de Python durante la captura impide la presencia de registros duplicados	0% de duplicados
Puntualidad	Grado en que los datos representan la realidad desde el punto requerido en el tiempo	Se comprueba en una muestra el número de reseñas de Google en el dataset con el número de reseñas en Google en el momento del chequeo	100% de registros actualizados
Validez	Los datos se ajustan a la sintaxis de su definición	Se comprueba si el contenido de las celdas se ajusta al campo	Se observan un par de teléfonos con una sintaxis diferente. Se corregirá en la fase de limpieza de datos
Exactitud	Los datos se ajustan a la realidad que están describiendo	Requiere verificación de campo	No se comprueba
Consistencia	Ausencia de diferencias con respecto a su definición	Se comprueba en una muestra si el contenido coincide con Google	100% de registros consistentes

6.7 Otros estudios

No se han encontrado estudios o análisis relativo al uso de los datos ofrecidos por restaurantguru.com.

Sin embargo, en internet, 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. [Link](#)
- Allika, Krishnakanth, Insights into Toronto's Foodservice Market using Data Science Tools (August 08, 2019). Available at [SSRN](#)

7. Inspiración

Este conjunto de datos es interesante porque mezcla datos geográficos de localización, con datos de servicios proporcionados y con datos de sentimientos de 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 tienen 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?
- Etc...

8. Licencia

El software está publicado con licencia [MIT](#).

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.

El dataset se ha publicado en Zenodo con licencia [Creative Commons Attribution 4.0 International](#). Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad que esta presenta en relación con el trabajo realizado:

- 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.

9. Código

El código se encuentra disponible en la carpeta src de este repositorio [Git](#).

Durante el desarrollo del programa de web scraping se han encontrado una serie de dificultades que se enumeran a continuación, así como la solución que se les ha dado:

Dificultad	Solución
Evitar ser confundidos con un robot, para no ser bloqueados y que no se muestren captchas	Modificar la cabecera de las peticiones indicando entre otras opciones el user-agent: función <i>get_headers</i> Introducir retardos entre peticiones: función <i>sleep_random</i>
Emular el comportamiento humano de hacer scroll sobre una página para cargar todo su contenido	Usar webdriver de la librería selenium: función <i>scroll_down_to_bottom</i>
Gestionar reintentos de acceso a páginas en caso de error	Usar el objeto HTTPAdapter de la librería requests.adapters y el objeto Retry de la librería urllib3.util: función <i>get_request_session_with_retry_options</i>

La carpeta src contiene 3 ficheros que se explican a continuación.

src/scraper.properties

Fichero de propiedades donde se configura el modo de ejecución del script.

La ejecución puede ser:

- En modo interactivo: se solicitan los parámetros de ejecución del script por pantalla.
- En modo test: se leen los parámetros de ejecución del script de la sección Test del fichero de propiedades.

El fichero de propiedades se compone de 2 secciones:

- Run: dónde se indica el modo de ejecución del script (interactivo o test)
- Test: dónde se inician los parámetros a utilizar en modo test. Estos parámetros son:
 - debug_enabled: para habilitar/deshabilitar las trazas de ejecución del script.
 - scroll_down: para habilitar/deshabilitar scroll automático de una página hasta el final.
 - city: municipio de captura de los datos. También permite seleccionar áreas administrativas o países.

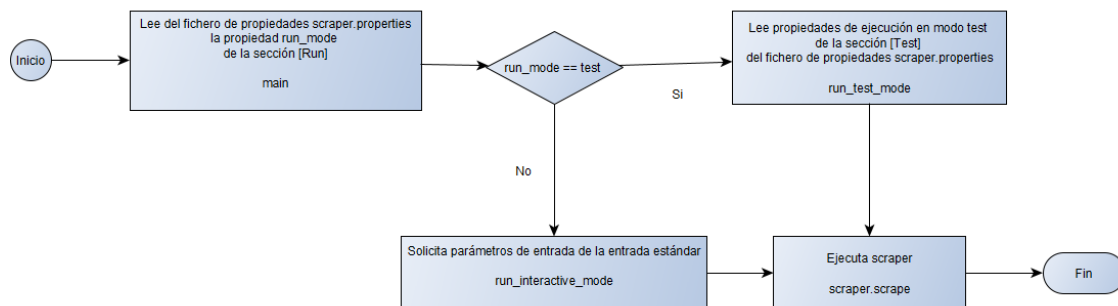
```
[Run]
# values: [test, interactive]
run_mode = interactive

[Test]
# Change this value for enable/disable debug trace
debug_enabled = True
# Change this value to see the difference between using Selenium (True) or not (False)
scroll_down = False
# Change this value to get data from another city
city = Alcala de Henares
```

src/main.py

Invoca al scraper según el modo de ejecución configurado en el fichero de propiedades scraper.properties.

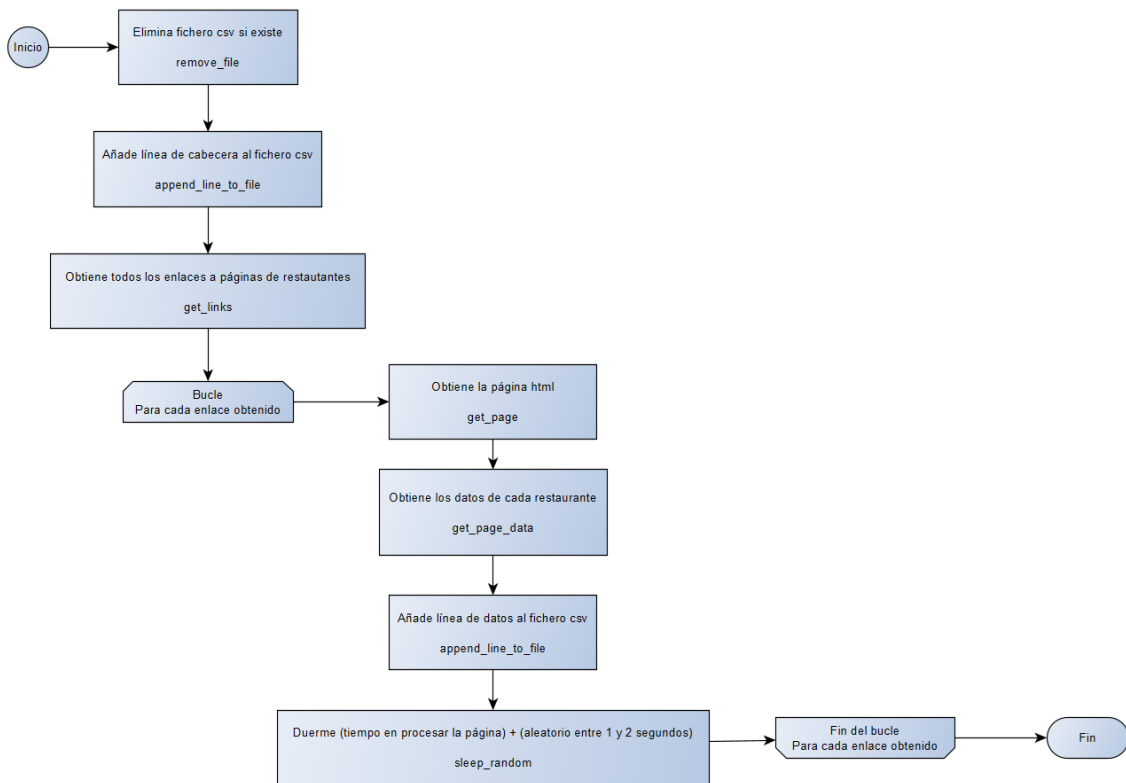
Función main



src/scraper.py

Realiza las funciones de scraping y guarda los resultados en un fichero en formato csv. A continuación se muestran diagramas de las funciones principales.

Función scrape



Función get_links

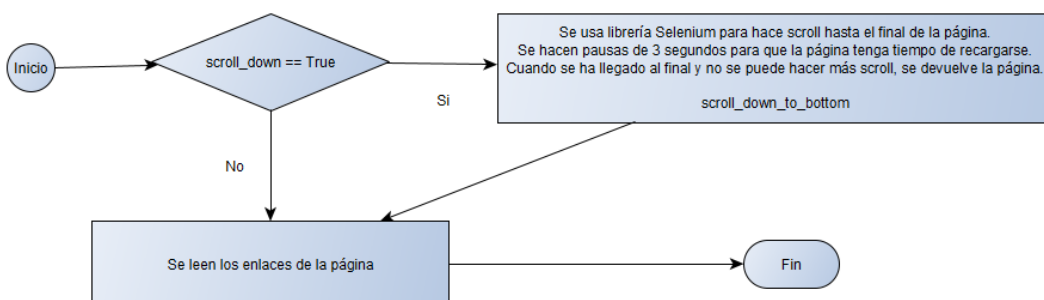
La página de inicio de la que se obtienen los enlaces a los restaurantes carga solo los 20 primeros. Para que la página cargue la totalidad de restaurantes existentes, es necesario ir haciendo scroll hacia abajo. Este comportamiento se emula en la función `scroll_down_to_bottom`, que usa la librería Selenium.

Si `scroll_mode=True`, se hace scroll automático hasta el final de la página.

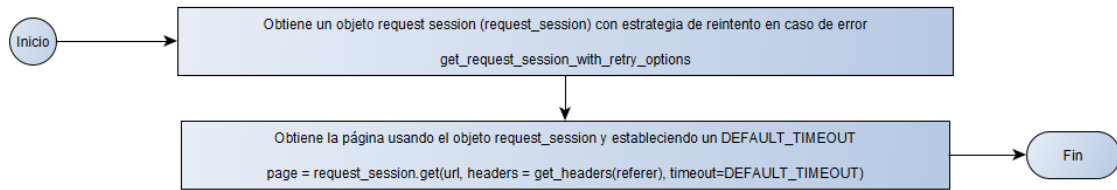
Si `scroll_mode=False`, no se hace scroll, y solo se obtienen enlaces a los 20 primeros restaurantes.

En el webdriver de Selenium se configuran las siguientes opciones:

- `user agent`: que indica el agente de usuario usado, en este caso Chrome
- `headless`: que indica que no muestre el browser



Función *get_page*



La función *get_page*, hace uso de la función *get_request_session_with_retry_options*, que devuelve un objeto request sesión configurado con estrategia de reintento en caso de error.

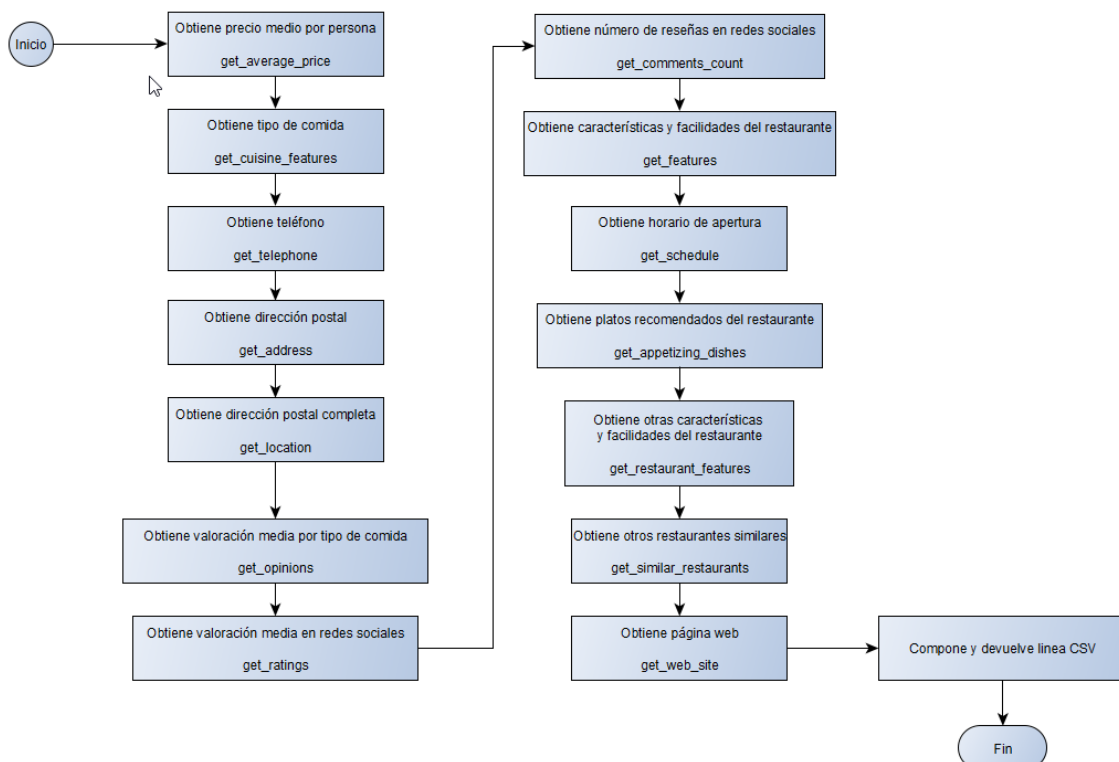
Las peticiones http sobre las que se hacen reintentos en caso de error son HEAD, GET y OPTIONS, y se controlan los errores de cliente (4xx) y servidor (5xx) más comunes:

- 429: too many requests
- 500: internal server error
- 502: bad gateway
- 503: service unavailable
- 505: gateway timeout

Se establece también el máximo de reintentos a 3 y un tiempo de espera entre fallos que responde al siguiente algoritmo:

$\{\text{backoff factor}\} * (2 ** (\{\text{number of total retries}\} - 1))$, donde backoff factor = 1

Función *get_page_data*



10. Dataset

Se encuentra publicado en [Zenodo](#).

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

El desarrollo del código se ha realizado usando la técnica “pair programming”, consistente en trabajar en el mismo equipo ambos programadores de forma conjunta, uno (el conductor) escribiendo el código, y otro (el observador) supervisándolo.

Los roles se han ido alternando en las sesiones de Teams en las que se ha compartido pantalla, cambiando de conductor a observador y viceversa.

Al haber trabajado ambos programadores en el mismo equipo, todos los commits de código al repositorio se han hecho por el mismo programador.