

Web Crawling y análisis de datos con Python

GABRIEL M. LESKE

Tutor: Dr. Isaac Lera Castro

Trabajo Final del Máster Universitario en
Tecnologías de la Información
Universidad de las Islas Baleares

Introducción

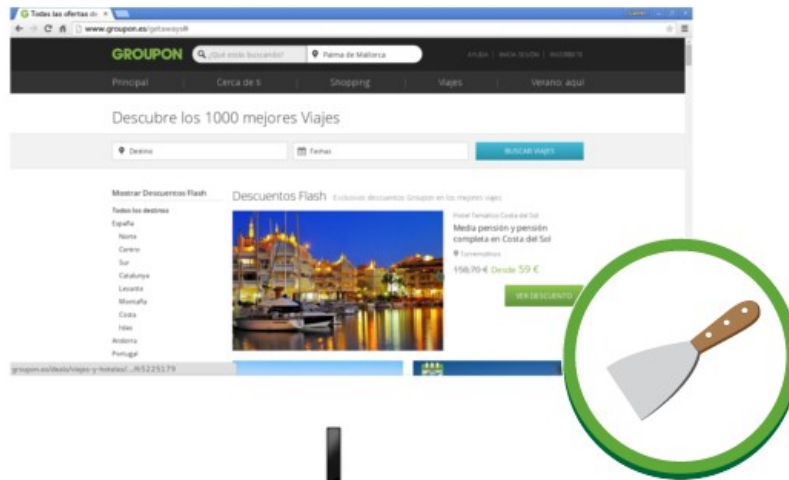
- Objetivos:
 - Obtención de datos de una web turística con ofertas para su explotación.
 - Encontrar indicadores de calidad de las ofertas.
- Procedimiento:
 - Extracción datos de forma automatizada.
 - Almacenamiento de datos.
 - Análisis de datos.

Técnicas utilizadas

- El **web crawling** (raspado web) consiste en la extracción de los datos significativos de la web para su posterior manipulación.
- El de **machine learning** (aprendizaje automático) es una rama de la inteligencia artificial que intenta crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de muestras.

Arquitectura

1. Web Crawling con Scrapy



2. Obtención de ficheros JSON



4. Análisis de la información



3. Importar JSON a MongoDB



Key	Value	Type
(1) ObjectId["55697a..."]	{ 12 fields }	Object
(2) ObjectId["55697a..."]	{ 12 fields }	Object
_id	ObjectId["55697a1bc6dda..."]	ObjectId
location	San Antonio	String
stars	3.000000	Double
url	http://www.groupon.es/deal...	String
price	89.000000	Double
discount	58.000000	Double
place	Love Motel	String
options	Lo que obtienes Incluido en t...	String
description	Motel LuVe Motel LuVe 3* s...	String
title	Valencia: 1 noche para 2 en ...	String
timestamp	15-06-06 10:35:42:1603...	String
address	Cigüeña, n8 bajo San Antoni...	String
(3) ObjectId["55697a..."]	{ 12 fields }	Object

Desarrollo: Extracción

¡Nuevo! Buscar hoteles en cualquier destino, cualquier día

Con más descuentos que nunca y miles de hoteles que ofrecen 5% de crédito Bucks Groupon, siempre encontrará el mejor valor en Groupon. Busca ahora.

Ahorra hasta un 70% en Madrid

Sí, quiero recibir emails de Groupon en relación con las últimas ofertas de bienes, servicios y viajes en mi ubicación, personalizadas en función de las cookies usadas en mis dispositivos, y de mi ubicación e historial de navegación y compras, u otra información que proporcione. [Declaración de Privacidad.*](#)

Explora Groupon GRATIS y tendrás acceso a ofertas para restaurantes, spas, vacaciones y mucho más ...

Dirección de email

Elige tu ciudad Madrid

¡Al Groupon de hoy!

*Puedes rechazar y eliminar las cookies directamente a través de tu navegador y gestionar tu suscripción a newsletters en cualquier momento pinchando el link para darse de baja incluido al pie de las newsletter de Groupon que recibas.

Gracias, ya estoy inscrito y acepto la Política de Privacidad de Groupon.


Hotel PlayaMaro
Nerja
62% Descuento
52 €

Hospedería Porta Coeli
Sigüenza
68% Descuento
25 €

49 de 188

MOSTRAR 48 OFERTAS MÁS

Sanxenxo: 5 o 7 noches para dos con desayuno buffet y late check-out en Hotel Farsund, a 50m de la playa desde 159€



DESD E
159 €

OPCIONES ▼

VALOR	DESCUENTO	TÚ AHORRAS
340 €	53%	181 €

(TIEMPO LIMITADO!)
7 días 10:40:23


55 comprados

COMPARTIR ESTA OFERTA

Me gusta 2

Hotel Farsund

Web de la empresa



Destacados

Hotel situado a 50 metros de la Playa de Ares, España

Lo que obtienes

Incluido en todas las opciones: alojamiento en habitación doble, desayuno buffet, late check-out hasta las 13h

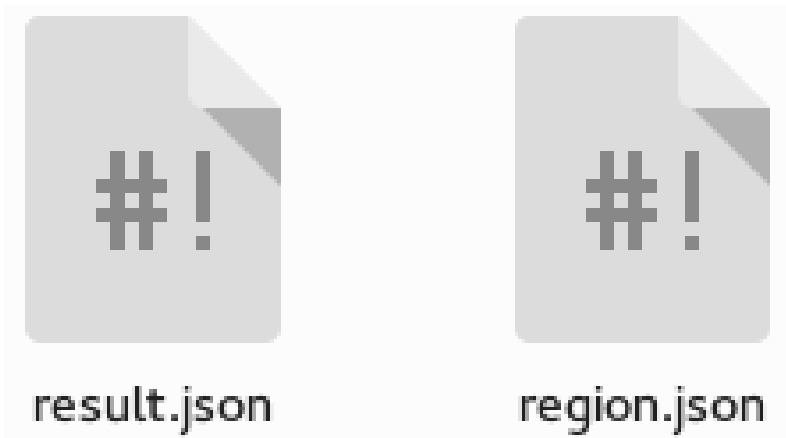
- Temporada media: 5 noches - 2 pers - 159€ (53% de descuento)
- Temporada media: 7 noches - 2 pers - 209€ (56% de descuento)
- Temporada alta: 5 noches - 2 pers - 199€ (49% de descuento)

Guía Groupon de Sanxenxo

Historia, naturaleza, gastronomía, playa, ría, Atlántico, Isla de La Toja.

Desarrollo: Almacenamiento

- Scrapy permite exportar los resultados en JSON, que pueden ser importados fácilmente a MongoDB.



db.result.find() x

localhost:27017 tfm

```
db.result.find()
```

result 0.418 sec. 0 50

Key	Value	Type
▶ (1) ObjectId("55697a... { 12 fields }		Object
▼ (2) ObjectId("55697a... { 12 fields }		Object
_id	ObjectId("55697a1bc6ddda...)	ObjectId
location	San Antonio	String
stars	3.000000	Double
url	http://www.groupon.es/deal...	String
price	89.000000	Double
discount	58.000000	Double
place	Luve Motel	String
options	Lo que obtienes Incluido en t...	String
description	Motel LuVe Motel LuVe 3* s...	String
title	Valencia: 1 noche para 2 en ...	String
timestamp	15-06-06 10:35:42:1603...	String
address	Cigueña, n8 bajo San Antoni...	String
▶ (3) ObjectId("55697a... { 12 fields }		Object

Desarrollo: Análisis

- Pymongo permite interactuar con MongoDB desde Python.
- Scikit-learn es un conjunto de herramientas simples y eficientes para la minería y el análisis de datos que incluye:
 - Pandas: permite crear una estructura para hacer el análisis de datos.
 - NumPy: permite trabajar con vectores y matrices.
 - Matplotlib: permite la representación gráfica en 2D/3D.

Desarrollo: Análisis

- Procedimiento:
 - Tokenizer: trocear los textos en palabras, eliminando símbolos y números (NLTK).
 - Stop of words: eliminar palabras que no aportan ningún significado (artículos, conjunciones y otras)
 - Obtención de las matrices y aplicación del algoritmo de clústering.
 - Repetir el proceso con diferentes agrupaciones (por categoría y por región).

Desarrollo: Análisis

- Se realizan los siguientes análisis:
 - Clustering con Kmeans y extracción de términos más relevantes por clúster.
 - Representación gráfica de los clústers.
 - Agregación:
 - Cálculo del precio medio y descuento medio por clúster y por categoría.
 - Términos más relevantes por región y por categoría.

Resultados

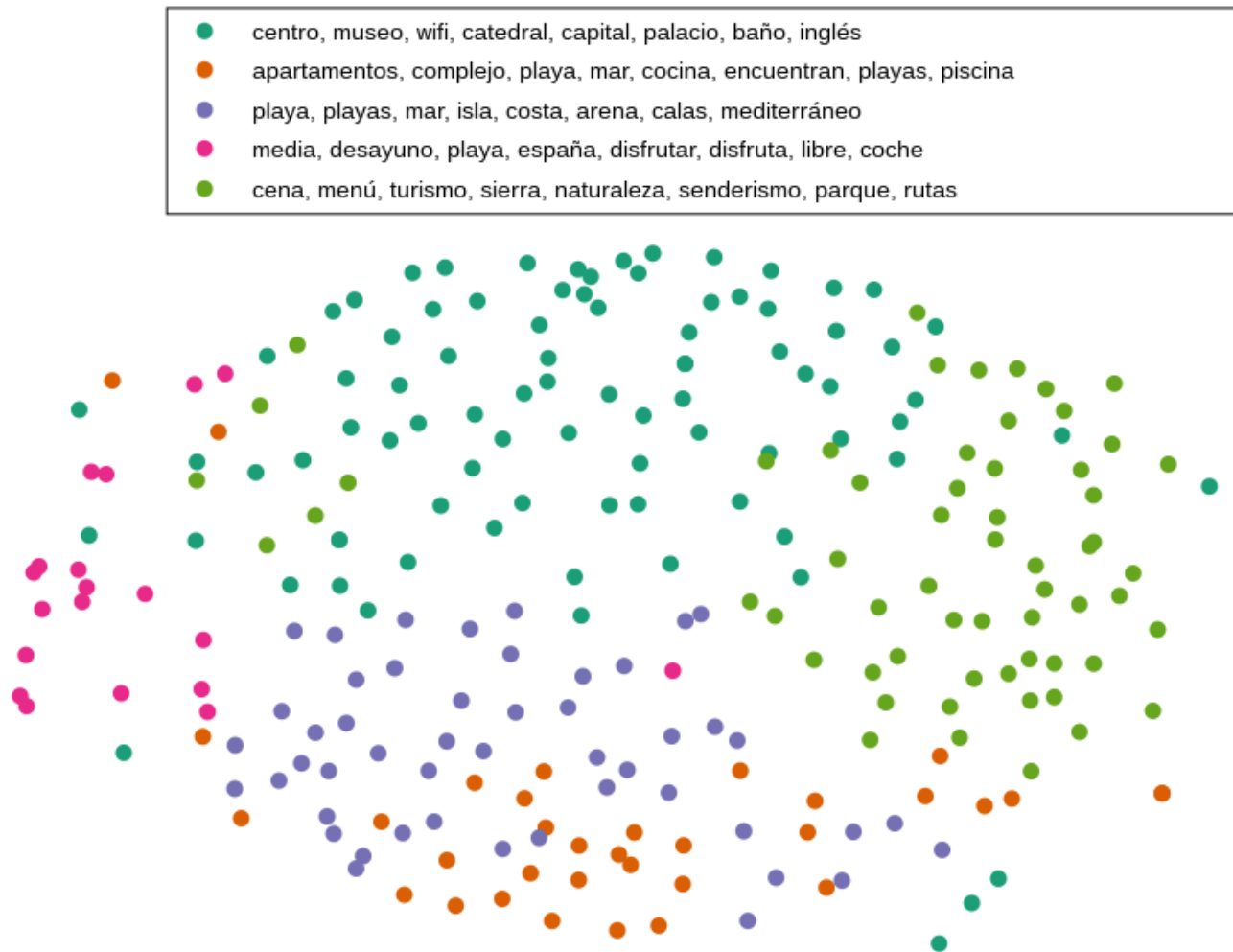
- Resultado del Clustering

Cluster	Cantidad	Palabras
0	60	naturaleza, turismo, sierra, senderismo, parque, región, rutas, activo
1	80	playa, mar, playas, apartamentos, costa, isla, arena, metros
2	20	cena, menú, bebidas, vino, casa, restaurante, cafetería, patrimonio
3	68	centro, palacio, capital, catedral, museo, estación, ofrece, wifi
4	20	desayuno, media, playa, disfrutar, España, libre, coche, pequeño

- Clúster 0 está relacionado con la naturaleza.
- Clúster 1 está relacionado con zonas de costa.
- Clúster 2 está relacionado con alimentación y bebidas.
- Clúster 3 está relacionado con zonas de ciudad.
- Clúster 4: observando sólo las palabras no queda claro exactamente su temática, pero luego de analizar las ofertas se deduce que está relacionadas con relax (spa, balnearios, masajes, etc).

Resultados

- Representación gráfica del clustering



Resultados

- Agregación

- Descuento y precio medio por cluster:

Cluster	Descuento Medio	Precio Medio
0	55.196429	63.206897
1	54.75	133.909091
2	55.263158	50.578947
3	53	97.651163
4	56.5	194.333333

- Descuento medio por clúster es muy similar, entre el 53-56%.
- El clúster 2 tiene el precio medio más bajo y el menor número de ofertas con términos como cena, menú y restaurante.
- El clúster 4 precio medio más caro 194.3€, por lo tanto los términos de este clúster como disfrutar, coche y libre le dan más valor a las ofertas.
- El clúster 1 precio medio de 133.9€, siendo el clúster con mayor cantidad de ofertas con términos como costa, mar e isla.
- El término playa es significativo en ambos clústers.

Resultados

- Agregación
 - Descuento y precio medio por cluster:

Palabras clave por categoría (estrellas):

1: apartamentos, canaria, playa, junio, mar, mayo, palmera, agosto
2: hostel, municipio, wifi, zona, patrimonio, aprovecha, comarca, playa
3: wifi, centro, baño, guía, tv, catedral, mascotas, partir
4: centro, aire, piscina, baño, spa, wifi, palacio, tv
5: desayuno, media, pensión, palacio, disfrutar, golf, tiempo, locales

- El término centro es relevante en ofertas de 3 y 4 estrellas.
- El término palacio lo es para las de 4 y 5,
- El término wifi lo es para las de 2 y 3. P
- La mayoría de las ofertas de 5 estrellas son de media pensión con desayuno,
- Las de 2 estrellas son hostales.
- Las de 3 estrellas donde más mascotas se admiten.

Resultados

- Agregación

- Palabras clave por región:

Islas: playa, wifi, mar, centro, zona, playas, baño, gratuito
Especial: precio, viaje, oportunidad, directo, mundo, pensión, españa, mínimo
Europa: museo, centro, baño, tv, inglés, guía, aire, encuentra
España: playa, mar, wifi, zona, baño, parque, guía, naturaleza
Portugal: playa, wifi, zona, mar, baño, centro, guía, gratuito
Costa: mar, playa, costa, castillo, municipio, mascotas, visitar, zona
Norte: mar, gastronomía, naturaleza, restaurante, wifi, baño, playa, guía
Centro: cena, madrid, palacio, centro, aire, rural, tv, wifi
Montaña: playa, wifi, mar, baño, zona, centro, gratuito, playas
Catalunya: barcelona, mar, zona, actividades, municipio, mediterráneo, playas, guía
Andorra: julio, andorra, jueves, apartamentos, caldea, sábado, montaña, rutas
Sur: granada, cena, wifi, playa, parque, partir, guía, zona
Fuera de Europa: ammán, barcelona, madrid, desayuno, salida, prevista, traslado, llegada
Levante: playa, apartamentos, murcia, mar, castillo, costa, histórico, castellón

- Los términos playa y wifi son de los que más abundan.
- En regiones de Montaña el término más destacado sea playa y que en Centro aparezcan términos como rural y aire.
- Las palabras hacen referencia a la temática de cada región como por ejemplo en Europa museo e inglés, España playa, en el Norte gastronomía y naturaleza, y en Levante castillo e histórico.

Conclusiones

- La extracción de datos de la web daba pocas muestras (aprox. 200 como máximo), por lo que ha sido necesario realizar más extracciones para anidar datos nuevos.
- El análisis ha permitido encontrar patrones comunes en las ofertas, haciendo posible:
 - Agruparlas en función de su contenido.
 - Deducir cuáles son las palabras que les proporcionan mayor o menor valor.
- Destacar la rapidez y simplicidad de las tecnologías: Scrapy, MongoDB y Scikit-learn.

Conclusiones

- La extracción de datos de la web daba pocas muestras (aprox. 200 como máximo), por lo que ha sido necesario realizar más extracciones para anidar datos nuevos.
- El análisis ha permitido encontrar patrones comunes en las ofertas, haciendo posible:
 - Agruparlas en función de su contenido.
 - Deducir cuáles son las palabras que les proporcionan mayor o menor valor.
- Destacar la rapidez y simplicidad de las tecnologías: Scrapy, MongoDB y Scikit-learn.