

Práctica 1: Web Scraping

Autoras: Sandra Campos Suárez y M^a de los Ángeles García Carrión

Contexto

El conjunto de datos que se desea estudiar recoge los diferentes animales que se encuentran en el bioparc de Fuengirola. Se ha escogido este bioparc debido a que es uno de los 8 mejores zoológicos de España, conocido como un lugar bien condicionado y altamente ético donde aprender sobre los animales. Algunos de los últimos animales criados en el Bioparc son unas tortugas boba descubiertas en la propia Fuengirola y unos leopardos de Sri Lanka.



Figure 1: *Figura 1: Leopardo de Sri Lanka*

Para la elección de este dataset hemos evaluado que se cumplan los siguientes aspectos: el archivo robots.txt, el mapa del sitio web, su tamaño, la tecnología usada y el propietario del mismo. Encontraremos en el apartado 6 un análisis detallado de estos puntos.

Título

Características de los Animales del bioparc de Fuengirola.

Descripción del dataset

En el dataset se encontrará información de los animales que se encuentran en el Bioparc de Fuengirola. En este conjunto de datos encontraremos la familia perteneciente de cada animal junto con su especie, orden, hábitat, clasificación, zona, dieta, gestación y grado de amenaza.

Representación gráfica

En el siguiente diagrama podemos observar la tabla Animales que será la que se va a analizar en esta práctica. Se ha añadido la tabla Bioparc ya que la tabla Animales recoge el detalle de la variable Animales, relacionadas ambas por el id que será añadido en la extracción generada.

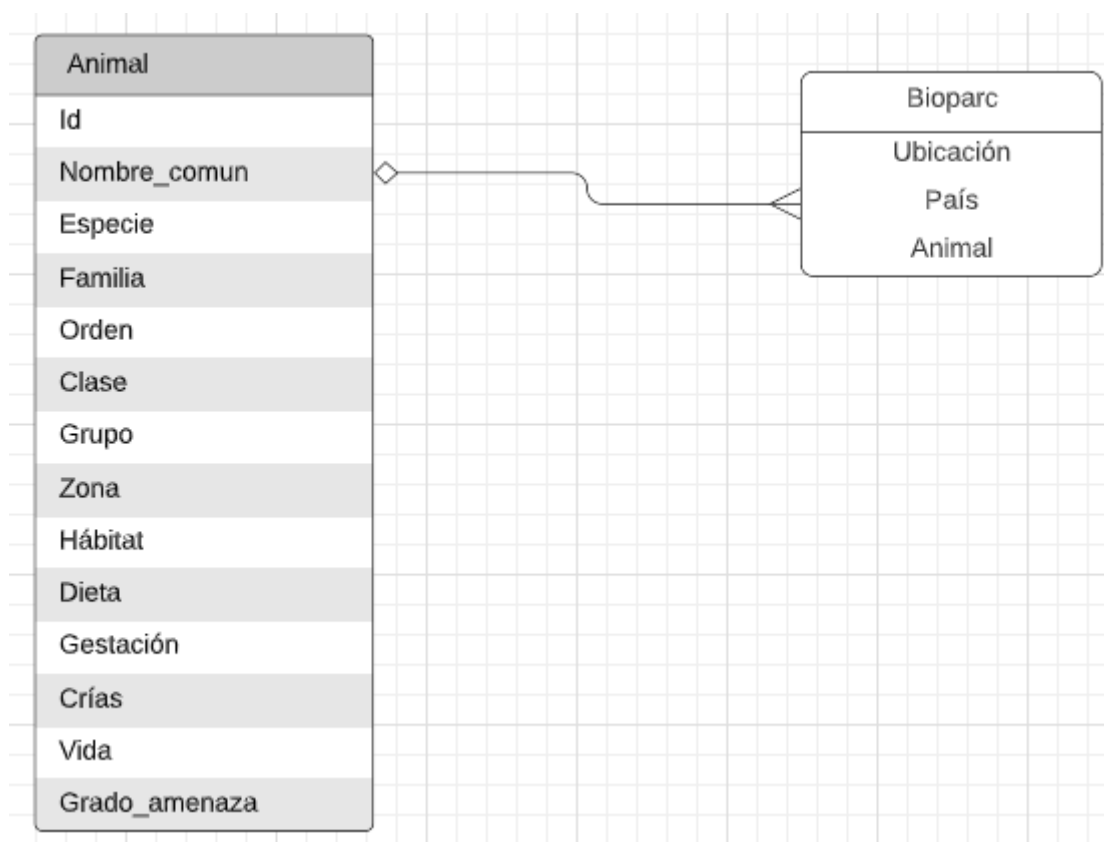


Figure 2: *Figura 2: Diagrama dataset*

Contenido

El dataset recoge todos los animales existentes en el Bioparc de Fuengirola durante la primera semana de abril de 2022. Extraídos del siguiente enlace: <https://www.bioparcfuengirola.es/animales/clasificacion-animal/>

Los campos que encontraremos en el conjunto de datos son los siguientes:

- **Nombre_comun:** nombre del animal.
- **Especie:** categoría de clasificación del animal.
- **Familia:** familia del animal.
- **Orden:** categoría taxonómica entre la clase y la familia.
- **Clase:** categoría en la taxonomía, situada entre el filo o la división y el orden. Conocido a su vez como subgrupo de los animales.
- **Grupo:** clasificación de los animales que se presentan la misma especie (peces, anfibios, reptiles, aves y mamíferos entre otros).
- **Zona:** zona geográfica donde habita el animal.
- **Habitat:** lugar de condiciones apropiadas para que viva un organismo, especie o comunidad animal o vegetal.
- **Dieta:** alimentación del animal, es decir, carnívoros, herbívoros y omnívoros.
- **Gestacion:** tiempo de gestación.
- **Crias:** número de crías.
- **Vida:** esperanza de vida del animal.
- **Grado_amenaza:** Grado de extinción.

Agradecimientos

En cuanto a los pasos llevados a cabo para la elección de este dataset se ha realizado una evaluación inicial de los siguientes aspectos:

- **archivo robots.txt:** Las restricciones a tener en cuenta cuando se pretende rastrearlas. En este caso a /wp-admin/

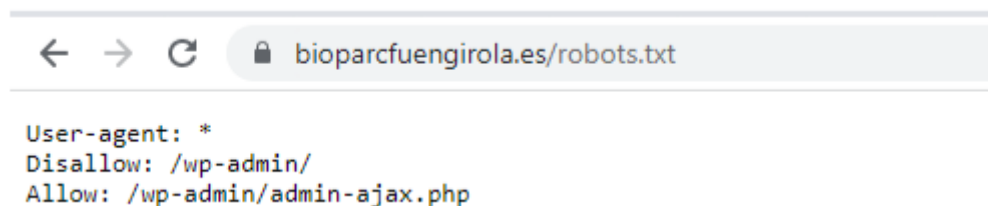


Figure 3: Image 1: archivo robots.txt


- **mapa del sitio web:** En este caso cumple el formato Simplemaps XML. Para el caso de la web de Fuengirola accedemos al siguiente enlace (https://www.bioparcfuengirola.es/sitemap_index.xml)
- **Tamaño:** Al realizar la búsqueda en Google para Bioparc de Fuengirola obtenemos 389.000 resultados.
- **Tecnología usada:** Para la extracción de el dataset se ha utilizado la herramienta Python y se ha hecho uso de las librerías Python Requests y BeautifulSoup, diseñadas para la extracción de contenido web.
- **Propietario del mismo:** Hemos utilizamos dos formas para conocer el propietario de la página web del bioparc de Fuengirola. En primer lugar, hemos accedido a la página web <https://whois.domaintools.com/bioparcfuengirola.es>

Además, hemos utilizado la sentencia *whois* en Python para conocer el propietario de la web

— Domain Profile

Registrar Status	taken	
Name Servers	NS1.INETSERVER.COM.ES (has 23 domains) NS2.INETSERVER.COM.ES (has 23 domains)	↻
Tech Contact	—	
IP Address	51.178.156.6 - 19 other sites hosted on this server	↻
IP Location	 - Madrid - Madrid - People SI Internet	
ASN	 AS16276 OVH, FR (registered Feb 15, 2001)	
Hosting History	1 change on 2 unique name servers over 7 years	↻

— Website

Website Title	 500 SSL negotiation failed:	↻
Response Code	500	

Whois Record (last updated on 2022-04-09)


```
% NOTE: The registry for this domain name does not publish ownership
%       records (whois records) in the standard format.  This data
%       represents the most likely status of the domain based on
%       information provided by the Internet's domain name servers (DNS).

domain: bioparcfuengirola.es
status: taken
nameserver: ns1.inetserver.com.es
nameserver: ns2.inetserver.com.es


% For more information, please visit http://www.nic.es/
```

Figure 4: *Image 2: Mapa del sitio web*

— Domain Profile

Registrar Status	taken	
Name Servers	NS1.INETSERVER.COM.ES (has 23 domains) NS2.INETSERVER.COM.ES (has 23 domains)	↻
Tech Contact	—	
IP Address	51.178.156.6 - 19 other sites hosted on this server	↻
IP Location	 - Madrid - Madrid - People SI Internet	
ASN	 AS16276 OVH, FR (registered Feb 15, 2001)	
Hosting History	1 change on 2 unique name servers over 7 years	↻

— Website

Website Title	 500 SSL negotiation failed:	↻
Response Code	500	

Whois Record (last updated on 2022-04-09)

```
% NOTE: The registry for this domain name does not publish ownership
%       records (whois records) in the standard format.  This data
%       represents the most likely status of the domain based on
%       information provided by the Internet's domain name servers (DNS).

domain: bioparcfuengirola.es
status: taken
nameserver: ns1.inetserver.com.es
nameserver: ns2.inetserver.com.es

% For more information, please visit http://www.nic.es/
```

Figure 5: *Image 3: Propietario*

```
In [1]: import whois
        print(whois.whois('https://www.bioparcfuengirola.es'))

{
  "domain_name": null,
  "registrar": null,
  "whois_server": null,
  "referral_url": null,
  "updated_date": null,
  "creation_date": null,
  "expiration_date": null,
  "name_servers": null,
  "status": null,
  "emails": null,
  "dnssec": null,
  "name": null,
  "org": null,
  "address": null,
  "city": null,
  "state": null,
  "zipcode": null,
  "country": null
}
```

Figure 6: *Image 4: Propietario*

**** Inspiración ****

El objetivo principal de este estudio es analizar los diferentes animales existentes en el Bioparc de Fuengirola. Para ello se ha estructurado en los siguientes puntos donde se centrará especialmente en un análisis cualitativo de las variables:

- Número de animales especies por familia, orden clase, grupo junto con porcentaje que representa respecto al total.
- Intervalo de gestación, vida y crías por especie.
- Qué tipo de especie presenta mayor gestación o vida.
- Conocer el número de especies por zona junto con su hábitat.
- Relación entre los diferentes tipos de animales junto con la zona geográfica
- Analizar el grado de amenaza de las especies para saber así cuales son las especies que están o no en estado de extinción.

Licencia

Hemos escogido la licencia *Released Under CC0* ya que no afecta en ninguna forma los derechos de patentes o de marcas sobre la obra, ni derechos que otras personas puedan tener en la obra o en cómo la obra es usada, como derechos de publicidad o privacidad.

Con esta licencia se permite que otros usuarios puedan usar esta información libremente para cualquier propósito sin restricción.