Tipología y ciclo de vida de los datos

Alumno: Javier Cantero Lorenzo (Aula 3)

PRA1: Web Scraping

Pregunta 1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

La **micología** es una rama de la biología que se dedica al estudio de los hongos. El estudio y uso culinario de las setas ha fascinado a la humanidad desde tiempo prehistóricos, siendo consideradas como un alimento sagrado por muchas civilizaciones de todo el mundo, especialmente las orientales. [1]

Siendo típicas de platos otoñales, actualmente el consumo de setas se asocia a la cocina *gourmet*, constituyendo un alimento que aporta gran sabor a los platos siendo bajas en grasas y calorías. [2]

Sin embargo, en el contexto local de las Islas Baleares, el uso de setas en la gastronomía ha estado tradicionalmente asociado a la economía de subsistencia mallorquina en tiempos de pobreza previos al boom del turismo [3]. En la actualidad la recogida de las setas es una tradición cultural en muchas ocasiones transmitida generacionalmente.

No todas las setas son comestibles, incluso algunas pueden llegar a ser letales. Se estima que en España se producen anualmente más de 400 intoxicaciones graves [4] por el consumo de setas (micetismo), consecuencia entre otros factores de la falta de un catálogo actualizado y la falta de una formación específica para la recogida e identificación de las setas.

En el año 2007 una colaboración entre el Museu Balear de Ciències Naturals y el área de botánica de la Universitat de les Illes Balears, en un proyecto financiado por la Fundació Sa Nostra, digitalizó el catálogo micológico balear redactado un año antes por Carles Constantino y Josep L. Siquier. Los resultados de este proyecto pueden ser consultados en http://bolets.uib.es/cat/index.html.

En este proyecto se propone el diseño de un script de web scraping que permita la extracción de la versión más actualizada del catálogo micológico balear como punto de partida para el desarrollo de proyectos académicos especializados o productos de inteligencia artificial que ayuden a los recolectores baleares amateur en la identificación de la toxicidad de una cierta especie.

[1] https://www.math.uci.edu/~vbaranov/nicetexts/eng/mushrooms.html

[2] https://www.alimente.elconfidencial.com/gastronomia-y-cocina/2020-02-06/setas-hongos-gourmet-incluir-dieta_1836486/

[3] http://www.centpercent.cat/tomeu-caldentey-a-la-cuina-cada-bolet-te-un-tractament-especific/

[4] https://www.heraldo.es/noticias/salud/2018/10/08/cada-ano-producen-400-casos-graves-intoxicaciones-por-setas-1270707-2261131.html

Pregunta 2. Definir un título para el dataset. Elegir un título que sea descriptivo.

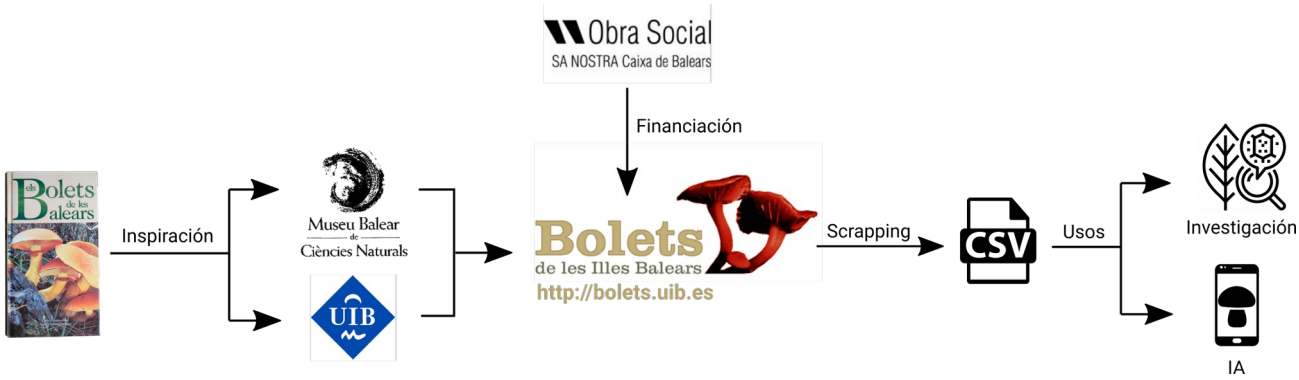
Catálogo micológico de las Islas Baleares.

Pregunta 3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

En el dataset asociado a esta práctica se recoge en detalle el conjunto del catálogo micológico característico de las Islas Baleares. Para cada seta se presenta su taxonomía identificativa, así como información acerca de su geografía, morfología y toxicidad alimentaria.

El conjunto de los datos se presenta en formato CSV. Ha pasado por un primer proceso de limpieza para eliminar la duplicidad de entradas y asegurar la integridad de los argumentos. Sería necesario llevar a cabo una segunda fase de procesado previo a un análisis específico de su contenido.

Pregunta 4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.



Pregunta 5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Cada una de las entradas presentes en el conjunto de datos corresponde a una seta diferente de la micología balear.

La información alojada en la página web se fundamenta en un catálogo publicado en papel a fecha de 2006. La última actualización de la web fue el 21 de Diciembre de 2007. Se espera que en un futuro la página pudiera ser actualizada con un catálogo más completo y actualizado.

Los atributos extraídos para cada una de las setas son los siguientes:

- **scientific_name:** Nombre científico oficial.
- **family:** Familia taxonómica.
- **genre:** Género taxonómico.
- **alternative_scientific_names:** Nombres científicos alternativos siguiendo otros sistemas de nomenclatura.

- **ca_common_name:** Nombres populares en catalán.
- **es_common_name:** Nombres populares en español.
- **description:** Descripción detallada.
- **additional_info:** Notas adicionales.
- **islands:** Islas del archipiélago balear en donde se encuentra la seta.
- **habitat:** Hábitats.
- **edibility:** Toxicidad alimentaria.

Notar que no todas las entradas poseen información acerca de todos los atributos.

Los datos han sido recogidos utilizando técnicas de web scraping mediante el lenguaje Python. En concreto se ha hecho uso de la librería *BeautifulSoup* para analizar la estructura HTML de la página.

La información de cada una de las setas se encuentra alojada en la página web http://bolets.uib.es/cat/bolet/[bolet_id].html donde **[bolet_id]** representa el código identificativo de cada seta.

En una primera fase se han obtenido los códigos de todas las setas disponibles a partir de la indexación escalonada de las setas por familia, género y nombre científico. Seguidamente para cada una de las entradas se ha extraído los detalles proporcionados por la página. Finalmente se han aplicado técnicas de preprocesado para eliminar registros duplicados (por ejemplo, misma setas con nombres científicos alternativos) o para presentar de forma sencilla el valor de atributos multicategóricos.

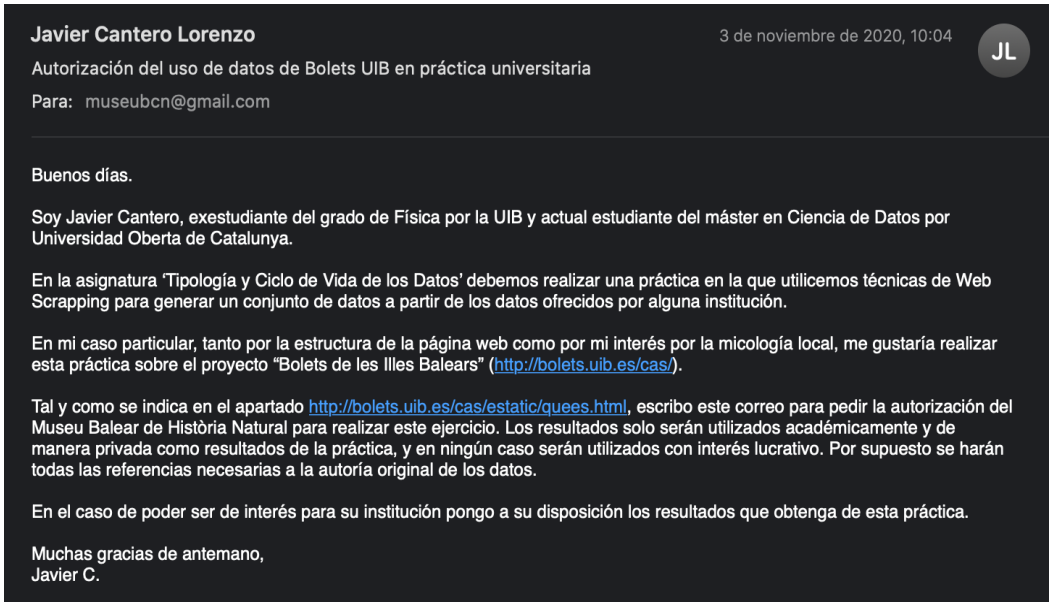
Pregunta 6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Esta práctica no podría haber sido realizada de no ser por los autores del libro '*Els bolets de les Balears*' (Micobalear, C.B.) Carles Constantino y Josep L. Siquier, quienes condujeron el peso del estudio y catálogo de la micología de las Islas Baleares.

Del mismo modo esta práctica pretender ser un paso extra y fundamentado en el trabajo de digitalización realizado por el Museu Balear de Ciències Naturals y la Universitat de les Illes Balears con el patrocinio de Obra Social Sa Nostra Caixa de Balears.

Antes de realizar el proceso de web scraping se consultó el archivo **robots.txt** asociado a la página web de la Universitat de les Illes Balears (https://www.uib.cat/robots.txt). En éste destaca el bloqueo de servicios de scraping especialmente a información relacionada con el personal de la universidad. Al ser una práctica con fines académicos se consideró oportuno seguir adelante.

En la página del proyecto se indica que tanto las imágenes como los textos están protegidos por derechos de autor, y no se puede hacer ningún uso comercial o electrónico sin el consentimiento explícito del Museo Balear de Ciències Naturals. El correo de contacto en la página se encuentra fuera de servicio, con lo que se contactó con el museo a partir del correo que aparece en su página oficial http://www.museucienciesnaturals.org/.



Javier Cantero Lorenzo 3 de noviembre de 2020, 10:04

Autorización del uso de datos de Bolets UIB en práctica universitaria

Para: museubcn@gmail.com

Buenos días.

Soy Javier Cantero, exestudiante del grado de Física por la UIB y actual estudiante del máster en Ciencia de Datos por Universidad Oberta de Catalunya.

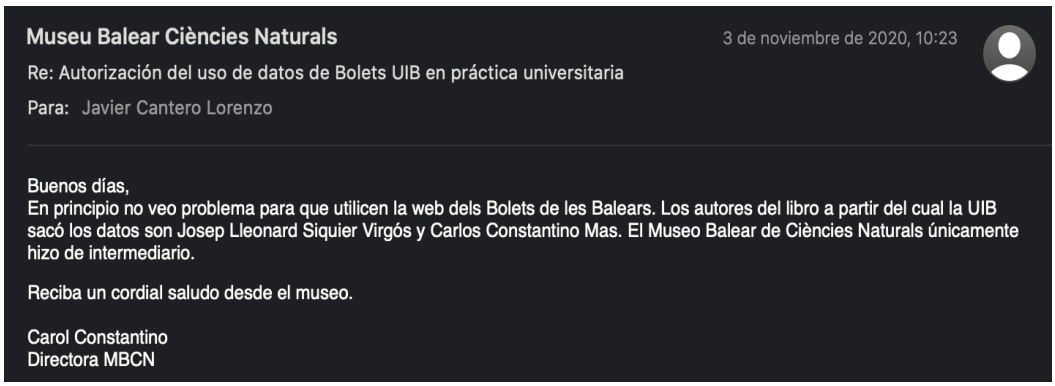
En la asignatura 'Tipología y Ciclo de Vida de los Datos' debemos realizar una práctica en la que utilicemos técnicas de Web Scrapping para generar un conjunto de datos a partir de los datos ofrecidos por alguna institución.

En mi caso particular, tanto por la estructura de la página web como por mi interés por la micología local, me gustaría realizar esta práctica sobre el proyecto "Bolets de les Illes Balears" (http://bolets.uib.es/cas/).

Tal y como se indica en el apartado http://bolets.uib.es/cas/estatic/quees.html, escribo este correo para pedir la autorización del Museu Balear de Història Natural para realizar este ejercicio. Los resultados solo serán utilizados académicamente y de manera privada como resultados de la práctica, y en ningún caso serán utilizados con interés lucrativo. Por supuesto se harán todas las referencias necesarias a la autoría original de los datos.

En el caso de poder ser de interés para su institución pongo a su disposición los resultados que obtenga de esta práctica.

Muchas gracias de antemano,
Javier C.



Museu Balear Ciències Naturals 3 de noviembre de 2020, 10:23

Re: Autorización del uso de datos de Bolets UIB en práctica universitaria

Para: Javier Cantero Lorenzo

Buenos días,
En principio no veo problema para que utilicen la web dels Bolets de les Balears. Los autores del libro a partir del cual la UIB sacó los datos son Josep Lleonard Siquier Virgós y Carlos Constantino Mas. El Museo Balear de Ciències Naturals únicamente hizo de intermediario.

Reciba un cordial saludo desde el museo.

Carol Constantino
Directora MBCN

Pregunta 7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Existen varias motivaciones detrás de la elección de este proyecto.

La **primera** de ellas es la de trabajar con un conjunto de datos local y cercano a mi residencia, tratando con instituciones que me son familiares.

La **segunda** de ellas es la de promover la creación de un catálogo micológico balear en constante actualización que sirva de referencia para el avance de la investigación botánica y como fuente de cultura balear.

La **tercera** de ellas es servir como punto de inicio al desarrollo de una aplicación de IA utilizando técnicas de *computer vision* que permita ayudar a los aficionados de la micología a identificar las especies que pretenden recolectar o con las que se topan al realizar rutas de excursión. Este tipo de aplicación no solo sería útil a nivel gastronómico sino también como parte de la experiencia del senderismo de exploración de la fauna y flora local.

Previo al desarrollo de este último proyecto más ambicioso sería necesario conducir estudios posteriores para determinar si la información visual que puede recoger la cámara es suficiente para categorizar correctamente la especie.

Algunas de las preguntas que se tratarían de resolver con este conjunto de datos serían:

¿Qué setas son características de las Islas Baleares?
¿Cuáles son las mejores zonas de recolección de setas con fines gastronómicos?
¿Qué tipo de hábitats aloja la mayor proporción de setas tóxicas?
¿Con qué seta me he cruzado durante mi ruta de senderismo?
¿Qué relación hay entre la morfología de la seta y su toxicidad?

Pregunta 8. Licencia.

Para satisfacer la licencia de los datos originales y asegurar el buen uso de los mismos, la licencia escogida para este conjunto de datos sería **Released Under CC BY-NC-SA 4.0 License**. De esta manera:

- El usuario puede modificar y redistribuir el material en cualquier medio o formato.
- El usuario puede modificar, transformar y construir productos basados en los datos.
- El usuario debe atribuir crédito al autor original de los datos.
- El usuario debe distribuir el material derivado de los datos bajo la misma licencia.
- El usuario no puede comercializar los productos derivados del uso de los datos.

REF: https://creativecommons.org/licenses/by-nc-sa/4.0/

Pregunta 9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El conjunto del proyecto se puede encontrar en el siguiente repositorio de GitHub:

https://github.com/jcanterol/bolets_uib_scrapper/

La estructura de las páginas individuales para cada seta consta de múltiples caracteres en blanco que entorpecen la obtención de los datos (\r, \t, \s, \n, …). Además, el atributo correspondiente a los nombres científicos alternativos no consta de identificador ni criterio estandarizado de separación entre nombres. Para solucionar todos estos problemas se ha tenido que hacer un uso especial de las expresiones regulares.

Pregunta 10. Dataset. Presentar el dataset en formato CSV.

El conjunto de datos ha sido subido a la plataforma Zenodo con el siguiente DOI:
http://doi.org/10.5281/zenodo.4263163

Responsabilidades

Esta práctica ha sido realizado por completo de manera individual por **Javier Cantero Lorenzo** con la autorización del profesor Jose Moreira Sanchez y la profesora responsable a día 09 de Octubre.

Contribuciones	Firma
Investigación previa	JCL
Redacción de las respuestas	JCL
Desarrollo código	JCL