TIPOLOGÍA Y CICLO DE VIDA DEL DATO

PRACTICA1: WEB SCRAPING

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS

Alumno: Eduardo Rivero Falcón.

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

El incendio en Gran canaria este verano puso de relieve los estragos que puede provocar estos fenómenos en la flora y los bosques. Los canarios en general y los de Gran Canaria en primera linea asistimos en esas fechas en los medios de comunicación a multitud de noticias que hablaban de la superficie y especies (flora) afectada, así como de las repercusiones y tiempo de recuperación.

La web "http://www.arbolappcanarias.es" ofrece un catálogo de las especies de árboles en Canarias con entre otros detalles descripción, localización e información de interés. La página es una iniciativa del Consejo Superior de Investigaciones Científicas (CSIC), el Real Jardín Botánico y el Jardín Botánico "Viera y Clavijo", unidad asociada al CSIC perteneciente al Cabildo de Gran Canaria. Ofrecen también una aplicación móvil que permite aparte de consultar datos identificar especies mediante una serie de preguntas que va contestando el usuario acerca de la corteza, hojas y otros detalles de los árboles.

En esta practica vamos a recopilar un dataset con datos de las especies de árboles de Canarias.

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

El título del dataset será arboles_canarias. Lo almacenaremos en el archivo 'species.csv'.

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).

Recopilaremos para cada especie nombres científico y común, localización en las islas y características acerca del orígen, si son autóctonas, naturalizadas e invasoras. Así mismo recopilaremos las imágenes disponibles de cada especie.

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

Los campos que forman la estructura del dataset son:

- Nombre científico: cadena (string)

- Nombre común: cadena (string)

- Autóctono: booleano

- Naturalizado: booleano

- Invasor: booleano

- Localización: cadena (string)

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

* *Nombre científico.*

Nombre con el que se identifica de forma unívoca la especie siguiendo los criterios aceptados por la comunidad científica.

* *Nombre común.*

Nombre o nombres con los que la población identifica a cada especie. Estos pueden no tener similitud con el nombre científico.

* *Autóctono.*

Del glosario de la propia página web para ser mas exactos en la definición, “Planta oriunda o nativa de un determinado país o región. No incluye a las plantas introducidas o naturalizadas”. El campo indica si es o no autóctona la especie.

* *Naturalizado.*

Del glosario “Planta que ha sido introducida voluntaria o involuntariamente por el ser humano fuera de su área de distribución natural y prospera en él como si fuese autóctona.”. El campo indica si la especie es naturalizada o no.

* *Invasor.*

Del glosario “Planta que ha sido introducida voluntaria o involuntariamente por el ser humano fuera de su área de distribución natural y que se ha naturalizado, resultando dañina para otras especies autóctonas o propias de esa región. ”. El campo indica si la especie es naturalizada o no.

* *Localización.*

Islas del archipiélago canario en las que se encuentra la especie en cuestión.

* *Aparte del dataset: Imágenes.*

Imágenes de la especie.

Estos atributos no presentan caducidad dada su naturaleza. Si acaso si hay cambio de criterio de la comunidad científica, de los botánicos y/ o investigadores al calificar una determinada especie como naturalizada o invasora o identificar una nueva especie en las islas. En cualquier caso estos posibles cambio se toman y aceptan en periodos largos, hablemos de varios años.

Los datos fueron tomados entre la primera semana del mes de Noviembre de 2019. Fueron recopilados mediante técnicas de “web scraping” en la página mencionada en el primer punto, por un agente (robot) codificado en el lenguaje Python.

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

Los propietarios son los promotores de la página web, el Jardín Botánico “Viera y Clavijo” (perteneciente al Cabildo de Gran Canaria), Real Jardín Botánico (estructura estatal) y el Consejo Superior de Investigaciones Científicas (CSIC). Mi agradecimiento a estas instituciones por la recopilación de la información y promoción al público de los árboles de Canarias.

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

Estos datos y las imágenes recopiladas podrían utilizarse en primer lugar, como punto de partida las imágenes como conjunto de entrenamiento para un algoritmo de reconocimiento, que permita identificar las especies tomando imágenes en vez de ir respondiendo a unas preguntas hasta llegar a la conclusión.

Una vez implementado el sistema de reconocimiento por imagen, se podría desde un dispositivo móvil con geolocalización y apoyándose de los campos del dataset delimitar zonas de especies de interés por ejemplo por ser autóctonas y raras, o por estar en peligro ante la presencia de especies invasoras. Gracias a la geolocalización y a los atributos podríamos delimitar esas zonas.

Se podría así mismo apoyándose en esos campos registrar la evolución temporal (extensión de las zonas y localización) de las especies autóctonas y naturalizadas, y las posibles invasoras, pudiendo a través de estos registros (series temporales) poder hacer predicciones de evolución que permitirían adoptar medidas por las autoridades de por ejemplo protección de las especies amenazadas.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

Los únicos propietarios y autores de estos datos son las organizaciones mencionadas promotoras de la página web (“http://www.arbolappcanarias.es “), es decir el Consejo Superior de Investigaciones Científicas (CSIC), el Real Jardín Botánico y el Jardín Botánico “Viera y Clavijo”, unidad asociada al CSIC perteneciente al Cabildo de Gran Canaria.

De acuerdo a los términos de uso disponibles en la página se debe hacer mención de esta autoría (nombrar dichas organizaciones) y no hacer un uso comercial.

En el caso de las imágenes habrá que hacer mención adicional de los autores:

FOTOGRAFÍAS: Magui Olangua, Felipe Castilla, Águedo Marrero, Carlos Aedo, Violeta Vicente, Josefa Navarro, Manuel Quevedo, Pilar Fernández, John Tann, Bidgee, Dinkum, Miwasatoshi, Ramin Nakisa, H.-U. Küenle, H. Zell, Peter O’Connor y Armin Kübelbeck.

Por todo lo anterior el tipo de licencia elegido es el “Released Under CC BY-NC-SA 4.0 License” por las condiciones que impone a la hora de compartir y adaptar el dataset:

Reconocimiento: debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.

No comercial: No puede utilizar el material para una finalidad comercial.

Compartir igual: Si remezcla, transforma o crea a partir del material, deberá difundir sus contribuciones bajo la misma licencia que el original.

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

Dado el origen y la autoría de estos datos quedan claros y justificados los fines de promoción, divulgación científica y los de bien de interés general, nunca intereses lucrativos de ninguna especie.

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

El código en Python se encuentra en la carpeta ‘src’, los archivos ‘main.py’ y ‘scraper.py’

10. Dataset. Presentar el dataset en formato CSV.

El archivo ‘species.csv’ se encuentra en la carpeta ‘data’.