M2.851 - Tipología y ciclo de vida de los datos

Práctica 1: WebScraping coffee_varieties

Equipo para la práctica

El equipo está formado por:

- Ivan García Jiménez (igarciajimenez0@uoc.edu)
- Itziar Ricondo Iriondo (iricondoi@uoc.edu)

Repositorio Github:

https://github.com/iricondoi/WebScraping_coffee_varieties

Contenido

1 Contexto

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

El café es uno de los productos más comercializados y una de las tres bebidas más consumidas del mundo (junto con el agua y el té) [1]. Se usan principalmente dos especies para la preparación su preparación: *Coffea arabica* y *Coffea canephora*. La primera especie abarca casi tres cuartas partes de la producción mundial y se cultiva principalmente desde México hasta Perú.

El cultivo del café está culturalmente ligado a la historia y al progreso de muchos países que lo han producido. En contra de lo que pueda creerse, alrededor de un 70% de la producción mundial de café proviene de explotaciones principalmente familiares de superficie inferior a diez hectáreas, incluso por debajo de cinco hectáreas.

Los factores de riesgo en el cultivo del café son las plagas como la roya [2], el cambio climático [3]–[5] y los precios que recibe el productor [6]. Estos factores tienen un impacto directo en la rentabilidad y sostenibilidad de los pequeños productores. La selección de las variantes de café a cultivar en este contexto es una decisión muy imponte [7], [8].

Todavía hoy en día siguen descubriéndose nuevas variedades silvestres de café en África. Además, se están desarrollando nuevas variedades en diferentes regiones de todo el mundo. En Centroamérica, Colombia, Brasil, Costa de Marfil o Kenia, los programas de mejora genética no solo se centran en el incremento de productividad, sino en la **tolerancia a plagas y enfermedades** (principalmente para Arábica) [9], **mejor adaptación fisiológica** a nuevas regiones cafeteras y cambio climático, y también a la diferenciación de atributos sensoriales.

El World Coffee Research (WCR) [10] es un programa colaborativo de investigación y desarrollo sin ánimo de lucro de la industria mundial del café para cultivar, proteger y mejorar el suministro de café de calidad, así como la mejora de calidad de vida de las familias que lo producen.

Este organismo sin ánimo de lucro ha publicado un catálogo de las variedades del café arábiga en Centro América y Caribe [11]. El objetivo del catálogo es **ofrecer información a los productores de café, sobre qué variedad de café se ajusta mejor a sus condiciones, con objeto de tomar la mejor decisión en cuanto a la variante a cultivar.** El trabajo se ha centrado en adquirir o rastrear la información de este catálogo disponible de forma abierta y generar un conjunto de datos.

2 Definir un título para el dataset

Elegir un título que sea descriptivo.

El título elegido para el dataset es **coffee_varieties**. El conjunto de datos muestra variedades de café arabica, por lo cual título describe concisamente la naturaleza del conjunto de datos.

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).

El conjunto de datos proporciona información sobre la apariencia, información agronómica, genética y disponibilidad de cada variedad de café arábica. La información más extensa es la agronómica, datos necesarios para identificar la idoneidad de determinada localización a las condiciones de crecimiento de la variedad de café.

4 Representación gráfica

Presentar una imagen o esquema que identifique el dataset visualmente.

Figura 1: Imagen representativa del conjunto de datos

Se ha optado por una composición de imagen que representa dos aspectos en el cultivo del café; por un lado, el trabajo del cultivo y recolección del café y, por otro, el trabajo de análisis y selección de variantes. En el cultivo y recolección del café el trabajo manual es imprescindible, representado por las manos del recolector. Por otro lado, el trabajo de análisis y decisión también se representa por las manos de una persona sobre un ordenador. La decisión de qué variedad de cultivo elegir viene soportada por la Ciencia de Datos, empezando por la obtención del conjunto de datos, que en el caso de este trabajo se ha realizado mediante Web Scraping.

5 Contenido

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

El conjunto de datos *coffee_varieties* extrae la información de catálogo Arabica Coffee Varieties [12] y la transforma a formato csv para que pueda ser analizada por cualquier usuario. El objetivo del conjunto, facilitado con la obtención del conjunto de datos, es analizar qué variantes del café se ajustan mejor a ciertas variables agronómicas o factores exteriores como son las plagas, para la toma de decisión. La selección de la variedad del café está directamente relacionada con la calidad final del café, la productividad de la explotación y, en consecuencia, la calidad de vida de los productores.

Los datos son estáticos, ya que representan propiedades sobre cada variedad de café que no se espera varíen a lo largo del tiempo. Por otro lado, podría darse la circunstancia que la organización propietaria de la fuente de los datos añadiera nuevas variedades. Al volver a ejecutar el archivo Python desarrollado para la obtención del conjunto de datos (archivo csv), todas las variantes expuestas en la página web serían rastreadas, adquiridas y guardadas en un archivo csv. A Noviembre de 2020, la actividad de Web Scraping sobre el catálogo da un resultado de 55 variantes de café.

Las variables del conjunto de datos se dividen en varias categorías:

- Apariencia: Engloba las variables *stature, leaf tip color, bean size*.
- Agronómica: Engloba las variables *optimal altitude*, *quality* potential at high altitude, *potential at high altitude, yield potential, coffee leaf, rust, coffee berry disease (CBD)*, *nematodes, year first production, nutrition requirement, ripening of fruit, cherry-to-green-bean outturn (CTGB), planting density, additional information.*
- Genética: Engloba las variables *lineage, genetic description, history.*
- Availability: Engloba las variables *breeder, intelectual property rights.*

El conjunto de datos incluye los siguientes campos:

Variable	Descripción	Tipo	Valores
Name	Nombre de la variedad de café.	String	
Description	Una breve descripción de la variedad de café.	String	
Stature	Indica el hábito de crecimiento de la planta	Discreto	Tall Dwarf/Compact
Leaf_tip_color	Indica cual es el color de la punta de las nuevas hojas de la planta.	Discreto	Green Light Bronze Bronze Dark Bronze Green or Bronze
Bean_size	Indica el tamaño del grano de café que produce la panta. Referecia, Caturra = Average, SL28 = Large, and Maragogipe = Very Large	Discreto	Below Average Average Large Very Large
Optimal_Altitude_1	Altitud para máximo rendimiento de calidad y agronómico. Rango latitud: 5°N to 5°S	String	>1600 m
Optimal_Altitude_2	Altitud para máximo rendimiento de calidad y agronómico. Rango latitud: 5–15°N and 5–15°S:	String	1000-1600m
Optimal_Altitude_3	Altitud para máximo rendimiento de calidad y agronómico. Rango latitud: >15°N and >15°S:	String	7000 m
Quality_potential	Indica cual es la calidad potencial del café de esta variedad si ha crecido a mayores altitudes.	Discreto	Very Low Low Good Very Good Exceptional
Yield_potential	Indica la cantidad potencial de frutos que se pueden llegar a recolectar de la variedad de café. Referencia: For reference, Caturra = Good and SL28 = Good	Discreto	Low Medium Good High Very high
Coffee_leaf_rust	Resistencia a la enfermedad de la roya.	Discreto	Resistant Tolerant Susceptible
CBD	Resistencia a la enfermedad de la baya del café.	Discreto	Resistant Tolerant Susceptible
Nematodes	Resistencia a los nematodos.	Discreto	Resistant Tolerant Susceptible

Variable	Descripción	Tipo	Valores
Year_First_Production	Indica el año en el que la planta producirá su primera fruta.	Discreto	Year 2 Year 3 Year 4
Nutrition_Requirement	Indica cuanta nutrición (por ejemplo, abono, fertilizante) requiere esta planta.	Discreto	Low Medium High Very High
Ripening_of_Fruit	Indica en qué momento de la temporada de cosecha madurará la fruta del árbol. Referencia: Caturra = Average	Discreto	Early Average Late Very late
CTGB	EL Cherry-to-Green-Bean Outturn, indica cuál es el tamaño del grano de café en relación con la fruta. Para referencia, Caturra = Average, SL28 = High	Discreto	Low Average High Very High
Planting_Density	Distancia que se debe utilizar para la siembra de esta variedad[1]	Discreto	1000-2000 a/ha (multiple-stem pruning) 2000-3000 a/ha (multiple-stem pruning) Less than 2000 a/ha (multiple-stem pruning) 3000-4000 (single stem pruning) 4000-5000 a/ha (single-stem pruning) 5000-6000 a/ha (single-stem pruning)
Additional	Incluye un breve texto con información agronómica adicional que no está incluida en las variables anteriores.	String	

[1] En Centro América, los árboles generalmente se podan para definirles un solo tallo vertical principal. En Africa, es típico podar los árboles dejándoles múltiples tallos principales verticales por árbol (2-3). Por lo tanto, aunque las densidades de la plantación de árboles son mucho más bajas en África, cada árbol está produciendo relativamente más porque tienen múltiples tallos principales.

Variable	Descripción	Tipo	Valores
Lineage	Indica de que variedad proviene esta variedad (cuando se conoce) o cuál es su linaje genético.	String	
Genetic Description	Las variedades de café están divididas en distintos grupos genéticos, este campo indica a cuál de ellos pertenece esta variedad.	Discreto	Ethiopian/landrace Bourbon-Typica group/Typica Bourbon-Typica group/Bourbon Bourbon-Typica group/Typica+Bourbon Introgressed/Catimor Introgressed/Sarchimor F1 hybrid/introgressed F1 hybrid/not introgressed
History	Este campo contiene un texto en el que se explica la historia de esta variedad de café.	String	
Breeder	Indica el nombre del cultivador de esta panta.	String	
Intellectual_Property_Rights	Propiedad intelectual	String	
Image	Indica la ruta en la que se encuentra la imagen de esta variedad.	String	

A continuación, se proporciona una descripción más informativa de algunos de los campos:

- **Optimal_Altitude**: Indica la altitud a la que se maximiza el potencial de calidad y rendimiento agronómico. La altitud óptima depende de la latitud de la granja. Las granjas ubicadas cerca del ecuador tendrán altitudes óptimas más altas que las que se encuentran más al norte o al sur del ecuador. Por ese motivo esta información está dividida en 3 campos:
  · **Optimal_Altitude_1**: Indica la Altitud óptima en el rango de latitud 5°N to 5°S (rango más cercano al Ecuador).
  · **Optimal_Altitude_2**: Indica la Altitud óptima en el rango de latitud 5–15°N and 5–15°S.
  · **Optimal_Altitude_3**: Indica la Altitud óptima en el rango de latitud >15°N and >15°S (rango más lejano al Ecuador).

- **Coffee_leaf_rust:** Indica la resistencia actual de la variedad de café a la enfermedad de la roya. Una variedad que es resistente a la roya hoy, puede que no lo sea mañana.

- **CBD:** El Coffee Berry Disease (CBD) o enfermedad de la baya del café es una enfermedad del café que afecta a la fruta. Este campo indica cómo de resistente es actualmente la variedad de café al CBD. Igual que en el campo anterior, una variedad que es resistente al CBD hoy, puede que no lo sea mañana.

- **Nematodes**: Indica la resistencia actual de la planta a los Nematodos, que son animales microscópicos que infectan las raíces y pueden provocar el marchitamiento y la muerte de la planta. Igual que en el campo anterior, una variedad que es resistente a los Nematodos hoy, puede que no lo sea mañana.

- **Intellectual_Property_Rights**: Indica la propiedad intelectual de esta variedad. Puede ser de dominio público, o estar registrada en la base de datos internacional de variedades, o estar denominada Unión Internacional para la Protección de las Obtenciones Vegetales (UPOV), u otra.

Durante la actividad de Web Scraping y para la obtención del conjunto de datos **se ha tenido que realizar alguna actividad de limpieza y tratamiento**, para evitar tener caracteres no leíbles. En concreto, se ha sustituido el guión '—' y valor vacío por “”.

Si se quisiera analizar el conjunto de datos obtenidos, habría que realizar una fase de limpieza y preprocesamiento. No se ha realizado en este trabajo, cuyo alcance se acota a la actividad de Web Scraping hasta la obtención del conjunto de datos. Al menos deberían tratarse los siguientes casos:

- Las variables Optimal_Altitude_1, Optimal_Altitude_2 y Optimal_Altitude_3 se han tipificado como String. Los valores introducidos implican rangos de alturas o alturas mínimas, con formatos como “700-1300m” o “>1300m”. Para ser analizados cuantitativamente, debería dividirse en 2 variables, una que represente la altura mínima y la otra la altura máxima.

- La variable Planting_Density se ha tipificado como String. Se observa que el valor de la variable (ej, “4000-5000 a/ha (using single-stem pruning)”) está compuesto por un valor de densidad expresado por un rango, junto con un string del tipo de poda. Esta variable podría dividirse en dos variables; la primera de ellas sería de tipo discreto con los rangos de densidad 1000-2000,2000-3000… mientras que la segunda variable sería también de tipo discreto asociado con el tipo de poda *(single stem pruning, multiple stem pruning*).

6 Agradecimientos

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

6.1 Propietario del conjunto de datos

El catálogo "Las variedades del Café Arábica" [11] ha sido publicado por World Coffee Research [10]. Esta organización lleva a cabo investigación básica en el cafeto, la planta del café, que asegure una mayor calidad del café, mayor productividad de la explotación y mejor rentabilidad para los productores.

El catálogo está disponible en dos formatos (un sitio web interactivo[2] y PDF) y dos idiomas (inglés y español). Reúne información esencial sobre 55 variedades de arábica de todo el mundo.

El catálogo ha sido desarrollado en colaboración con expertos en café de toda América Central y el Caribe. Es el resultado de las visitas a cada uno de los ocho países PROMECAFE [13] y de las entrevistas a cerca de 120 personas de alrededor de 70 organismos públicos y privados involucrados en sus sectores nacionales o regionales de café.

El catálogo "Las variedades del Café Arábica" por World Coffee Research se distribuye bajo una licencia Creative Commons[14] Attribution-NonCommercial-NoDerivatives 4.0 International License.

6.2 Estudios similares

La cadena de valor del café engloba fases que van desde la I+D para el desarrollo de nuevas especies, la producción, la distribución y el consumo. Atendiendo a las fases citadas, las variables de análisis alrededor del café pueden ser muy diversas.

Se muestran a continuación estudios de **Web Scraping** disponibles en repositorios públicos de Internet alrededor del café. Los estudios identificados no están orientados al ajuste entre la variedad del café y condiciones agronómicas, que son el objetivo de nuestro trabajo. Se han omitido estudios relacionados con la identificación de cafeterías (*coffee shops*). También se han omitido aquellos repositorios que no ofrecían una explicación mínima de la fuente de datos y trabajo realizado.

Guyot et al [15] muestran en Zenodo una base de datos de especies de café silvestre (WCSdb) alojada por la plataforma Pl@ntNet (http://publish.plantnet-project.org/project/wildcofdb_en), que proporciona información para 140 especies de café. El objetivo de esta base de datos es comprender y caracterizar mejor las especies (identificación, morfología, compuestos bioquímicos, diversidad genética, datos de secuencia). Su interés es biológico.

James LeDoux [16] obtiene un conjunto de datos para la evaluación de la calidad del café con datos obtenidos desde las páginas de evaluación del Coffee Quality Institute's [17] en Enero de 2018. Para cada café muestra datos relacionados con la calidad (aroma, sabor, regusto, acidez, dulzura…), metadatos del grano (procesamiento, color especie), y datos de gestión del cultivo (nombre de la granja, lote, empresa, altitud, región). Un trabajo muy similar sobre la misma web es el realizado en [18].

[2] https://varieties.worldcoffeeresearch.org/es/varieties

7 Inspiración

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

El interés del conjunto de datos reside en la información proporcionada para identificar las condiciones idóneas de cultivo de las especies de café. La selección de la variedad del café está directamente relacionada con la calidad final del café, la productividad y rentabilidad de la explotación y, en consecuencia, la calidad de vida de los productores.

Algunas de las preguntas a las que responde el conjunto de datos son:

- ¿Qué variedades son más resistentes a la enfermedad de la roya?
- ¿Cuál es la altitud en la cual la calidad y el potencial de producción de la variedad es mayor?
- ¿Qué variedades de café producen antes el fruto?
- ¿Qué distancia se debe utilizar para la siembra de una variedad?

8 Licencia

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

El catálogo original está publicado con licencia CC BY-NC-ND 4.0. Por lo tanto, la licencia de nuestro conjunto de datos debe ofrecer las mismas limitaciones. En concreto, esta licencia permite compartir el material original, esto es, **copiar y distribuir el material en cualquier medio o formato**.

Más información sobre el alcance de esta licencia está disponible en Creative Commons [14].

9 Código

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

El código se ha desarrollado en lenguaje **Python**, utilizando el **IDE Spyder** del entorno **Anaconda**. Las librerías que se han importado (y que se deben instalar para ejecutar el código), son las siguientes:

- **requests**: Se ha utilizado para hacer peticiones web.
- **BeautifulSoup**: Se ha utilizado para la terea de scraping.
- **builwith**: Se ha utilizado para adquirir la tecnología de creación de la web.
- **whois**: Se ha utilizado para adquirir información del propietario de la web.
- **pandas**: Se ha utilizado para crear un dataframe y exportarlo a CSV.
- **time**: Se ha utilizado para añadir retardos exponenciales.
- **os**: Se ha utilizado para obtener automáticamente las rutas donde guardar las imágenes y el CSV generado por el programa.

La codificación se ha realizado en Python. Se han definido las siguientes funciones:

- **scrap_all_coffees(main_URL**): Función que tiene como parámetro la URL principal donde aparecen todas las variantes de café. Se encarga de realizar las siguientes tareas:
    - Obtiene la URL de cada variante de café.
    - Implementa también la buena práctica de cambiar user agent del header para no evidenciar que se está realizando web scraping.
    - Llama a la función **scrap_coffe_variety** para cada variante de café**.**
    - Obtiene la URL de la imagen de cada variante de café.
    - Llama a la función **scrap_image.**

- **scrap_coffe_variety(aLink, headers):** Función que tiene como parámetros la dirección URL de cada variante de café y el header con user agent modificado para enmascarar la actividad de scraping. Realiza las siguientes tareas:
    - Realiza el request a la URL de cada variante de café (parámetro aLink).
    - Implementa las buenas prácticas de gestión de timeouts, y de añadir un retraso exponencial para evitar saturar el servidor de peticiones.
    - Para cada variante, realiza las siguientes actividades de scraping:
        - Scrapea variables sueltas (características de apariencia y algunas agronómicas).
        - Scrapea variables en tablas (características agronómicas, genéticas y de disponibilidad).

- **scrap_image(image_link, headers):** Función que tiene como parámetros la dirección URL de la imagen de cada variante y el header con user agent modificado para enmascarar la actividad de scraping. Realiza las siguientes tareas:
    - Realiza el request a la URL de cada imagen de café (parámetro image_link).
    - Implementa las buenas prácticas de gestión de timeouts, y de añadir un retraso exponencial para evitar saturar el servidor de peticiones.
    - Recoge la ruta actual del programa

- **tecnologia(URL):** Función para adquirir la tecnología con la que se ha creado la web.

- **propietario(URL):** Función para adquirir información del propietario de la web.

Las 2 últimas funciones **tecnología(URL)** y **propietario(URL)** son opcionales, ya que se utilizaron al inicio del proyecto para obtener esta información. Actualmente no se utilizan en el código principal, ya que augmentan el tiempo de ejecución y la información que aportan no se introduce en el dataset generado.

En la Figura 2 se muestra un esquema del flujo del programa, en el que se aprecia el flujo de extracción:

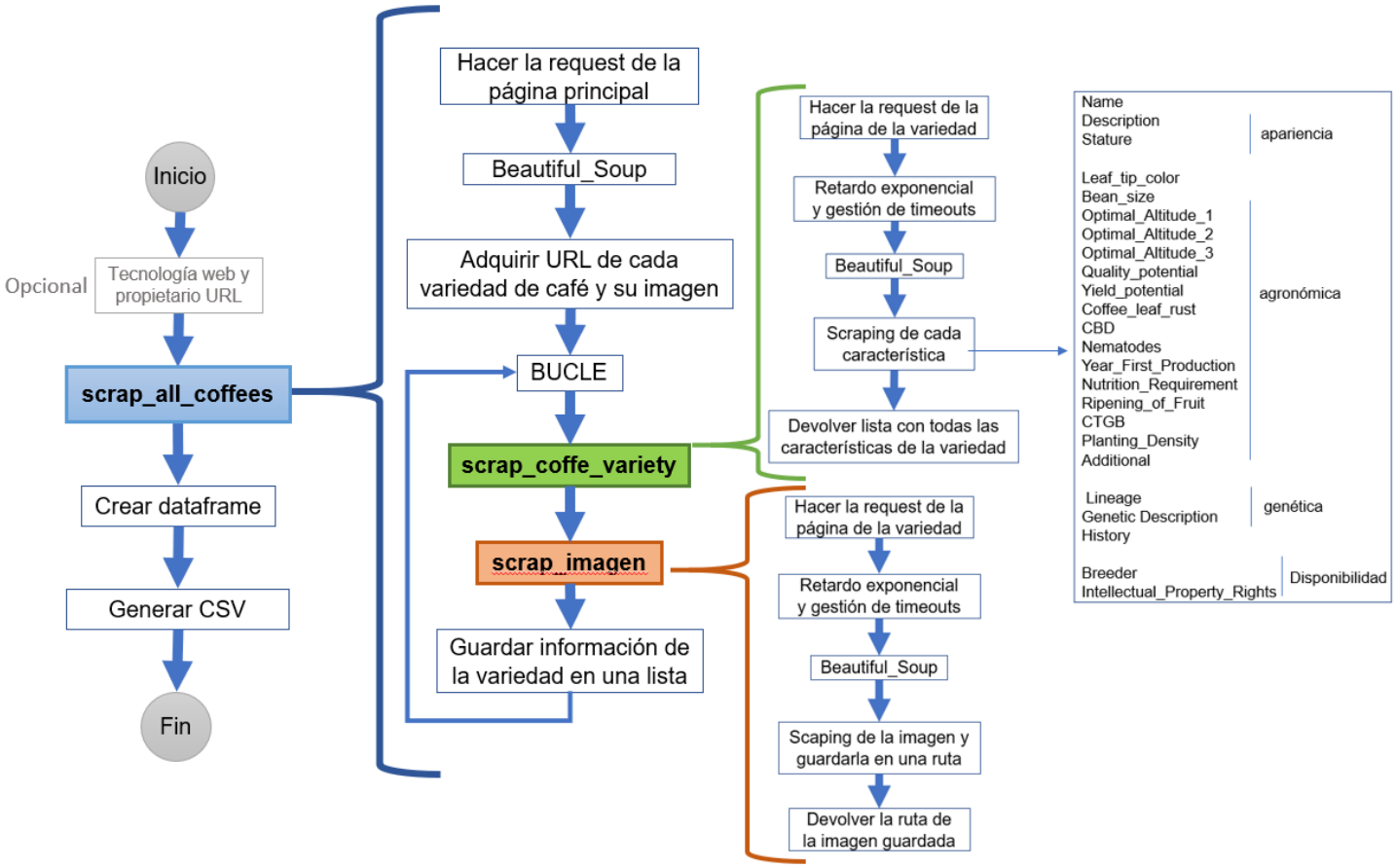

Figura 2: Flujo del programa

A continuación, se anexa la programación realizada:

```
# Importamos las librerias
import requests
from bs4 import BeautifulSoup
import builtwith
import whois
import pandas as pd
import time
import os


# Tecnologia con la que se ha creado la web
def tecnologia(URL):
    technology = builtwith.builtwith(URL)
    print(technology)
```

```
# Información del propietario de la web
def propietario(URL):
    owner = whois.whois(URL)
    print('Domain name: {}'.format(owner['domain_name']))
    print('Creation date: {}'.format(owner['creation_date']))
    print('City: {}'.format(owner['city']))
    print('State: {}'.format(owner['state']))
    print('Zipcode: {}'.format(owner['zipcode']))
    print('Country: {}'.format(owner['country']))


# Escrapea la imagen dada su URL por parámetro y la guarda en una carpeta
# imagenes en el directorio donde se encuetra este programa
def scrap_image(source_url, headers):
    # Hacemos la request y añadimos un retraso exponencial
    # para evitar satura el servidor de peticiones
    t1 = time.time()
    try:
        r = requests.get(source_url, stream = True, headers=headers,
timeout=10)
    except requests.exceptions.Timeout:
        pass
    response_delay = time.time() - t1
    time.sleep(5 * response_delay)
    # Si la request es correcta se scrapea la imagen
    if r.status_code == 200:
        ruta_actual = os.getcwd().replace("\\", "/")
        ruta_imagenes = ruta_actual + '/imagenes'
        # Si no existe el directorio imagenes lo creamos
        if not os.path.exists(ruta_imagenes):
            os.mkdir(ruta_imagenes)
        aSplit = source_url.split('/')
        ruta = ruta_imagenes + "/" + aSplit[len(aSplit)-1]
        output = open(ruta,"wb")
        for chunk in r:
            output.write(chunk)
        output.close()

    return(ruta)


# Scrapea una variedad de café dada su URL
def scrap_coffe_variety(variety_URL, headers):
    # Hacemos la request y añadimos un retraso exponencial
    # para evitar satura el servidor de peticiones
    t1 = time.time()
    try:
```

```
        variety_page = requests.get(variety_URL, headers= headers,
timeout=10)
    except requests.exceptions.Timeout:
        pass
    response_delay = time.time() - t1
    time.sleep(5 * response_delay)
    # Si la request es correcta se scrapea la web
    if variety_page.status_code == 200:
        variety_soup = BeautifulSoup(variety_page.content, "lxml")

        # Obtenemos el nombre y la descripción del café
        name = variety_soup.title.string
        coffee_name = name.split(' | ')[1]
        description = variety_soup.find('p').text
        # Creamos una lista con el nombre y la descripción del café
        variety = [coffee_name, description]

        # Obtenemos las propiedades de apariencia y algunas propiedades
        # agronómicas del café
        values = variety_soup.find_all('div',{'class':'value'})
        for v in values:
            clase = v.parent.get('class')
            # Para la propiedad altitude hay que navegar entre la estruct
ura
            # anidada para obtener los 3 valores de la tabla
            if clase[1] == 'altitude':
                optimal = v.find_all('div',{'class':'altitude-range-
value'})
                for i in optimal:
                    valor = i.text.replace('–','-').strip()
                    variety.append(valor)
            # Para el resto de propiedades se obtiene directamente el str
ing
            else:
                valor = str(v.contents[0]).strip()
                variety.append(valor)

         # Obtenemos el resto de propiedades agronómicas y las propiedade
s
         # de genetics y availability
        values = variety_soup.find_all('td',{'class':'cell value'})
        for v in values:
            clase = v.parent.get('class')
            # Las propiedades agronomics y history tienen párrafos
            # y se extrae el texto
            if clase[1] == 'agronomics' or clase[1] == 'history':
                valor = v.text.strip()
```

```
            # Para el resto de propiedades se obtiene directamente el str
ing
            else:
                valor = v.string.strip()
            # Se reemplazan los valores inexistentes y el caracter '—
' por ''
            if valor == '—' or not(valor):
                valor = ''
            variety.append(valor)

    # Se devuelve la lista variety, que es una lista con todas
    # las propiedades extraidas de esta variedad de café
    return(variety)


# Scrapea todas las variedades de café dada la URL principal
def scrap_all_coffees(main_URL):

    # Crear lista vacía de cafes
    coffees = list()

    # Se define un header cambiando el User-
Agent para no evidenciar que la
    # petición viene de un script
    headers = {}
    headers["Accept"] = "text/html,application/xhtml+xml,application/xml;
\
        q=0.9,image/webp,*/*;q=0.8"
    headers["Accept-Encoding"]= "gzip, deflate, sdch, br"
    headers["Accept-Language"]= "en-US,en;q=0.8"
    headers["Cache-Control"]= "no-cache"
    headers["dnt"]= "1"
    headers["Pragma"]= "no-cache"
    headers["Upgrade-Insecure-Requests"]= "1"
    headers["User-Agent"]= "Mozilla/5.0 (Windows NT 10.0; Win64; x64; \
        rv:82.0)Gecko/20100101 Firefox/82.0"
    # Se realiza la petición
    page = requests.get(main_URL, headers=headers)

    # Si la request es correcta se scrapea la web
    if page.status_code == 200:
        soup = BeautifulSoup(page.content, "lxml")
        # Bucle para encontrar las URL de cada una de las variedades
        for link in soup.find_all('a',href=True):
            aLink=link.get('href')
            # Si la URL corresponde a la de una variedad la scrapeamos co
n
```

```
            # la función scrap_coffe_variety y la añadimos a la lista cof
fees
            if ('varieties/' in aLink):
                coffee = scrap_coffe_variety(aLink, headers)
                # Se obtiene la URL de la imagen del café
                image_link = link.find('img').get('src')
                # Mediante la función scrap_image se guarda la imagen
                # y se obtiene su ruta para añadirla a la lista coffee
                ruta = scrap_image(image_link, headers)
                coffee.append(ruta)
                coffees.append(coffee)
    return(coffees)


# Definimos la URL principal
main_URL = 'https://varieties.worldcoffeeresearch.org/varieties/'
# Definimos la URL raiz
root_URL = 'https://worldcoffeeresearch.org/'

# Mostrar por pantalla la tecnología con la que se ha creado la web
#tecnologia(root_URL)

# Mostrar por pantalla la información del propietario de la web
#propietario(root_URL)

# Crear una lista con todos los cafés scrapeados
print('Begining scraping, wait 7-8 minutes')
t0 = time.time()
coffees = scrap_all_coffees(main_URL)

# Creamos una lista con las columnas del dataset
columnas = ["Name", "Description", "Stature", "Leaf_tip_color", "Bean_siz
e",
            "Optimal_Altitude_1", "Optimal_Altitude_2", "Optimal_Altitude
_3",
            "Quality_potential", "Yield_potential", "Coffee_leaf_rust", "
CBD",
            "Nematodes", "Year_First_Production", "Nutrition_Requirement"
,
            "Ripening_of_Fruit", "CTGB", "Planting_Density", "Additional"
,
            " Lineage", "Genetic Description", "History", "Breeder",
            "Intellectual_Property_Rights", "Image"]

# Crear un dataframe con la lista anterior y añadir columnas
df = pd.DataFrame(coffees, columns=columnas)
```

```
# Exportar el dataframe a CSV en la ruta donde se encuentra este programa
ruta_actual = os.getcwd().replace("\\", "/")
ruta_csv = ruta_actual + '/coffee_varieties.csv'
df.to_csv(ruta_csv, sep=';', index=False)

# Termina el programa
print('Scraping finished')
response_delay = int(time.time() - t0)
print('The program has lasted for {} seconds '.format(response_delay))
```

10 Dataset

Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

Se ha publicado el dataset en format CSV en Zenodo, con el siguiente DOI [19]:

10.5281/zenodo.4252702.

Breve descripción incluida en el dataset subido a Zenodo:

The dataset provides information on the appearance, agronomic information, genetics and availability of each variety of Arabica coffee. The most extensive is agronomic information, data necessary to identify the suitability of a certain location to the growing conditions of the coffee variety.

11 Tabla de contribuciones

Firmas:

- Ivan García, IG
- Itziar Ricondo, IR

Contribuciones	**Firma**
Investigación previa	IG,IR
Redacción de las respuestas	IG,IR
Desarrollo código	IG,IR

12 Referencias Bibliográficas

[1] «Café», *Wikipedia, la enciclopedia libre*. nov. 07, 2020, Accedido: nov. 07, 2020. [En línea]. Disponible en: https://es.wikipedia.org/w/index.php?title=Caf%C3%A9&oldid=130716376.
[2] F. Echeverria-Beirute, S. C. Murray, P. Klein, C. Kerth, R. Miller, y B. Bertrand, «Rust and Thinning Management Effect on Cup Quality and Plant Performance for Two Cultivars of Coffea arabica L», *J. Agric. Food Chem.*, vol. 66, n.º 21, pp. 5281-5292, may 2018, doi: 10.1021/acs.jafc.7b03180.
[3] O. Ovalle-Rivera, P. Läderach, C. Bunn, M. Obersteiner, y G. Schroth, «Projected Shifts in Coffea arabica Suitability among Major Global Producing Regions Due to Climate Change», *PLOS ONE*, vol. 10, n.º 4, p. e0124155, abr. 2015, doi: 10.1371/journal.pone.0124155.
[4] C. Bunn, P. Läderach, J. G. P. Jimenez, C. Montagnon, y T. Schilling, «Multiclass Classification of Agro-Ecological Zones for Arabica Coffee: An Improved Understanding of the Impacts of Climate Change», *PLOS ONE*, vol. 10, n.º 10, p. e0140490, oct. 2015, doi: 10.1371/journal.pone.0140490.
[5] Y. Pham, K. Reardon-smith, S. Mushtaq, y G. Cockfield, «The impact of climate change and variability on coffee production: a systematic review», *Clim. Change*, sep. 2019.
[6] «Situación del Café en Centroamérica». https://www.sica.int/iniciativas/cafe (accedido nov. 07, 2020).
[7] Z. Daggett, «3 Factors Which Impact a Decision on Which Coffee Variety to Plant», *Perfect Daily Grind*, jun. 13, 2015. https://perfectdailygrind.com/2015/06/3-factors-which-impact-a-decision-on-which-coffee-variety-to-plant/ (accedido nov. 08, 2020).
[8] H. Vossen, B. Bertrand, y A. Charrier, «Next generation variety development for sustainable production of arabica coffee (Coffea arabica L.): a review», *Euphytica*, vol. 204, feb. 2015, doi: 10.1007/s10681-015-1398-z.
[9] M. Geleta, I. Herrera, A. Monzón, y T. Bryngelsson, «Genetic Diversity of Arabica Coffee (Coffea arabica L.) in Nicaragua as Estimated by Simple Sequence Repeat Markers», *The Scientific World Journal*, jun. 04, 2012. https://www.hindawi.com/journals/tswj/2012/939820/ (accedido nov. 08, 2020).
[10] «World Coffee Research». https://worldcoffeeresearch.org/ (accedido nov. 07, 2020).
[11] «Arabica Coffee Varieties». https://worldcoffeeresearch.org/work/coffee-varieties-mesoamerica-and-caribbean/ (accedido nov. 07, 2020).
[12] «Arabica Coffee Varieties Catalog». https://varieties.worldcoffeeresearch.org (accedido nov. 08, 2020).
[13] «Promecafe». https://promecafe.net/ (accedido nov. 07, 2020).
[14] «Creative Commons», *Creative Commons*. https://creativecommons.org/ (accedido nov. 07, 2020).
[15] Guyot romain *et al.*, «WCSdb: A database of Wild Coffea Species.» Zenodo, jun. 20, 2020, doi: 10.5281/zenodo.3899717.
[16] J. LeDoux, *jldbc/coffee-quality-database*. 2020.
[17] «Coffee Review - The World's Leading Coffee Guide», *Coffee Review*. https://www.coffeereview.com/review/ (accedido nov. 07, 2020).
[18] «ashish-sharma-as/scrapy_coffeereview», *GitHub*. https://github.com/ashish-sharma-as/scrapy_coffeereview (accedido nov. 08, 2020).
[19] Ivan García y Itziar Ricondo, «coffee_varieties». Zenodo, nov. 06, 2020, doi: 10.5281/zenodo.4252702.