Tipología y ciclo de vida de los datos
PRÁCTICA 1

Alberto Lendínez Gutiérrez y David López de la Fuente

14 Abril 2020

Índice

1. Práctica 1

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

1. **Contexto**. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.
2. **Definir un título para el dataset**. Elegir un título que sea descriptivo.
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).
4. **Representación gráfica**. Presentar una imagen o esquema que identifique el dataset visualmente.
5. **Contenido**. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.
6. **Agradecimientos**. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).
7. **Inspiración**. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.
8. **Licencia**.
9. **Código**. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.
10. **Dataset**. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.
11. **Entrega.** Presentar el trabajo con el DOI del dataset en Github.

1.1. Contexto

Se requiere estudiar la totalidad de datasets disponibles en el repositorio de Machine Learning UCI, una web con una colección de bases de datos y generadores de bases de datos open source que pueden ser utilizados por cualquiera que desee realizar determinados estudios definidos en un campo concreto. Así pues, encontramos una gran cantidad de tipos de datos (desde médicos pasando por automovilísticos, arte, banderas y hasta una base de datos de hongos y setas comestibles).

La extracción del dataset que contenga la lista completa de las bases de datos que la página web contiene y su posterior estudio y manipulación es el objetivo principal de este trabajo, y trabajaremos con la URL principal del directorio de datasets: `http://archive.ics.uci.edu/ml/datasets.php`

Figura 1: Logo de UCI

1.2. Título del dataset

Repositorio de Datasets de UCI

1.3. Descripción del dataset

El dataset contiene la lista de bases de datos que se puede encontrar en el repositorio web de UCI y lo conforman 7 columnas:

- **Nombre:** nombre del dataset.
- **Tipo de dato:** tipos de datos que conforman el dataset.
- **Default Task:** tarea o finalidad con la que se realizó el dataset.
- **Atributos:** el tipo de los datos que conforman el dataset (strings, integers...).
- **Instancias:** numero de filas de datos que conforman el dataset.

- **Numero de atributos:** número de atributos diferentes que conforman el dataset.

- **Año:** año en el que se donó/subió a la web el dataset.

1.4. Representación gráfica

La siguiente foto es una representación gráfica del dataset:

	A	B	C	D	E	F
1	Nombre	Tipo de dato	Default Task	Atributos	Instancias	Numero de atributos
2	Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8
3	Adult	Multivariate	Classification	Categorical, Integer	48842	14
4	Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38
5	Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294
6	Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279
7	Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7
8	Audiology (Original)	Multivariate	Classification	Categorical	226	
9	Audiology (Standardized)	Multivariate	Classification	Categorical	226	69
10	Auto MPG	Multivariate	Regression	Categorical, Real	398	8
11	Automobile	Multivariate	Regression	Categorical, Integer, Real	205	26
12	Badges	Univariate, Text	Classification		294	1
13	Balance Scale	Multivariate	Classification	Categorical	625	4
14	Balloons	Multivariate	Classification	Categorical	16	4
15	Breast Cancer	Multivariate	Classification	Categorical	286	9
16	Breast Cancer Wisconsin (Original)	Multivariate	Classification	Integer	699	10
17	Breast Cancer Wisconsin (Prognostic)	Multivariate	Classification, Regression	Real	198	34
18	Breast Cancer Wisconsin (Diagnostic)	Multivariate	Classification	Real	569	32
19	Pittsburgh Bridges	Multivariate	Classification	Categorical, Integer	108	13
20	Car Evaluation	Multivariate	Classification	Categorical	1728	6
21	Census Income	Multivariate	Classification	Categorical, Integer	48842	14
22	Chess (King-Rook vs. King-Knight)	Multivariate, Data-Generator	Classification	Categorical, Integer		22
23	Chess (King-Rook vs. King-Pawn)	Multivariate	Classification	Categorical	3196	36
24	Chess (King-Rook vs. King)	Multivariate	Classification	Categorical, Integer	28056	6
25	Chess (Domain Theories)	Domain-Theory				
26	Bach Chorales	Univariate, Time-Series		Categorical, Integer	100	6
27	Connect-4	Multivariate, Spatial	Classification	Categorical	67557	42
28	Credit Approval	Multivariate	Classification	Categorical, Integer, Real	690	15
29	Japanese Credit Screening	Multivariate, Domain-Theory	Classification	Categorical, Real, Integer	125	
30	Computer Hardware	Multivariate	Regression	Integer	209	9
31	Contraceptive Method Choice	Multivariate	Classification	Categorical, Integer	1473	9
32	Covertype	Multivariate	Classification	Categorical, Integer	581012	54
33	Cylinder Bands	Multivariate	Classification	Categorical, Integer, Real	512	39
34	Dermatology	Multivariate	Classification	Categorical, Integer	366	33
35	Diabetes	Multivariate, Time-Series		Categorical, Integer		20

Figura 2: Representación gráfica

Como vemos, el dataset está dividido en el número de columnas establecido en el apartado 1.3 y consta de los mismos headers en sus columnas.

1.5. Contenido

Los campos que contiene el dataset son aquellos enumerados en el apartado 1.3: Nombre, Tipo de dato, Default Task, Atributos, Instancias, Número de atributos y Año y han sido recogidos mediante la conexión a la URL de la página de UCI facilitada en el apartado 1.1 del presente trabajo y su posterior volcado a un objeto del tipo BeutifulSoup para luego ser transformado en un archivo de extensión .csv. El periodo de tiempo de los datos comprende las fechas 1987 a 2020 y se corresponden con la época en los que los datasets fueron cedidos a la web. Existen algunos datos sin fecha que podrían ser filtrados en un procesamiento de limpieza de datos posterior y que muy probablemente se normalicen en futuras entregas del proyecto. Este suceso también incluye el resto de campos a excepción del de "Nombrez en futuros análisis se procederá a establecer filtros o soluciones generalizadas.

1.6. Agradecimientos

Los datos han sido recuperados a través del repositorio de Machine Learning UIC (Center for Machine Learning and Intelligent Systems). La información extraída para la realización de esta práctica proviene de su web `http://archive.ics.uci.edu/ml/datasets.php`. En esta web podemos encontrar un repositorio con información de todos los datasets que poseen.

Para extraer la información que se encuentra sobre el HTML de sus páginas web, hemos realizado un programa en Python utilizando diferentes librerías:

- **urlib:** permite la conexión a webs HTML.
- **beautifulsoup4:** permite analizar documentos HTML y extraer su información.
- **pandas:** permite el análisis y manipulación de los datos. Con esta librería crearemos el dataset y el CSV.
- **unicodedata:** proporciona acceso a la base de datos de caracteres Unicode. Nos servirá para modificar los datos Unicode del código.

Se ha realizado un análisis previo de la información encontrada en el repositorio web, viendo su formato y observando la información que se encontraba en este. Un gran número de usuarios facilitan los datasets que pueden ser descargados para realizar estudios concretos sobre ellos.

1.7. Inspiración

El conjunto de datos tiene como cometido facilitar el análisis de los datasets existentes en el repositorio y establecer algún tipo de orden que permita su clasificación por tipo o cualquier característica común que pueda apreciarse entre su basto catálogo de datasets.

Esto permitiría en un rápido vistazo establecer métricas sobre qué campo de datos es el que posee mayor contenido o sobre qué tipo de atributos o normalización se encuentra con más frecuencia entre los datasets. Todo esto para establecer el valor real que contiene el repositorio web de UCI y poder catalogar qué punto de referencia de datasets puede ofrecer según su contenido.

1.8. Licencia

Released Under CC0: Public Domain License puesto que estos datos se podrían extraer del propio repositorio web y además han sido subidos y donados por usuarios que no esperan ningún tipo de remuneración o beneficio con su uso. Además,

son accesibles para todo aquél interesado en su contenido. Esta licencia permite copiar, modificar, distribuir los datos y hacer comunicación pública.

1.9. Código

El código fuente creado para la realización de la práctica se encuentra subido al GitHub relacionado a la misma:

`https://github.com/alendinezuoc/PRAC1-Dataset-UCI/blob/master/Dataset_Machine_Learning_Repository.py`

Este código permite la extración de los datos del repositorio web UCI y la creación de un dataset con los datos recuperados. Se ha optado por la utilización de las librerías "Urllib"para la conexión con páginas web HTML y PHP, "BeautifulSoup"para volcar el contenido HTML de la web de UCI en un objeto BeautifulSoup que nos permita trabajar con él para extraer los datos en un dataset.

Además de esta breve explicación, en el repositorio se puede encontrar un archivo README con la descripción del proyecto así como diversa información de utilidad en su Wiki que complementaría todo lo ya aportado en el presente documento.

1.10. Dataset

El dataset se ha subido correctamente al repositorio Zenodo:

`https://zenodo.org/record/3748994#.XpL-qMgzZPY`

En el dataset se pueden encontrar los datos recuperados de la web de UCI y el volcado de su contenido en columnas renombradas según las especificaciones del apartado 1.3.

1.11. Entrega

En en siguiente enlace se encuentran el código utilizado para la realización del presente trabajo en formato .py, el dataset generado a través del código y un fichero README y una Wiki con información acerca del estado del proyecto y del funcionamiento del código:

`https://github.com/alendinezuoc/PRAC1-Dataset-UCI/`

2. Contribución al trabajo

En este apartado se adjunta la tabla de contribución al trabajo firmada por los integrantes del grupo con sus propias iniciales, constatando así la participación de ambos en todo momento durante la realización del trabajo expuesto.

Contribuciones	Firma
Investigación previa	Integrante 1: ALG; Integrante 2:DL
Redacción de las respuestas	Integrante 1: ALG; Integrante 2:DL
Desarrollo de código	Integrante 1: ALG; Integrante 2:DL

3. Conclusiones

- Se ha logrado extraer satisfactoriamente un dataset de una página web a través de Web Scrapping
- El código realizado es fácilmente entendible y está bien documentado
- El dataset resultante debería ser optimizado en el proceso de limpieza y análisis de datos

4. Bibliografía (libros)

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Mitchell, R. (2015) Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly media.
- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.

5. Bibliografía (webs)

- https://aukera.es/blog/web-scraping/
- https://hackernoon.com/
- https://jarroba.com/scraping-python-beautifulsoup-ejemplos/