Máster en Data Science

Tipología y ciclo de vida de los datos

Realizado por:
- Janneth Chicaiza Espinosa [https://github.com/jachicaiza]
- Jaime Velandia López [https://github.com/jvelandi]

Fecha de entrega: 15 de abril de 2019

Proyecto GitHub: https://github.com/jachicaiza/tipologia-prac1

Práctica 01

Tabla de contenidos

1. Contexto

1.1 Descripción del proyecto

La Ciencia Abierta se ha constituido en un movimiento capaz de acelerar en gran medida la producción y difusión del conocimiento (Friesike, 2015). Gracias a este y otros paradigmas, los investigadores están publicando en la Web sus conjuntos de datos (*datasets*), así fomentan su reuso y contribuyen a mejorar la colaboración entre los miembros de la comunidad científica.

Esta iniciativa surge con el objetivo de ayudar a los investigadores a encontrar conjuntos de datos que puedan utilizar en sus estudios y experimentaciones. Para conseguir el objetivo propuesto, en este proyecto de la asignatura se han elegido dos métodos de recolección de datos desde la Web: 1) librería scholarly para recoger datos del perfil de investigadores en Google Scholar; y, 2) scrapy para leer los metadatos de los datasets de UCI Learning Machine Repository.

En base a los intereses de un investigador (con un perfil en Google Scholar), el/ella podría ejecutar los scripts desarrollados para descargar los metadatos de los datasets disponibles en el repositorio seleccionado para la tarea de scrapy. Luego, se podría construir un agente que le recomiende los datasets que le ayude a generar nuevo conocimiento (McKiernan et. al. 2016).

1.2 Justificación de la elección de las fuentes de datos

- Google Scholar es un buscador de Google enfocado y especializado en la búsqueda de contenido y literatura científica y académica. Google Académico permite configurar un perfil de autor y realizar seguimiento sobre las citas de trabajos publicados. En este sitio, se puede encontrar información de identificación y los intereses de un investigador o de un conjunto de investigadores asociados a alguna institución.

- UCI Machine Learning Repositories es uno de los catálogos mejor clasificados para encontrar y descargar conjuntos de datos apropiados para ser analizados a través de diferentes algoritmos de aprendizaje automático. Actualmente, el sitio provee alrededor de 468 datasets clasificados de acuerdo a diferentes categorías. Aquí el investigador puede aplicar algunos criterios para filtrar y encontrar los datasets apropiados para sus tareas de análisis, sin embargo, aún no es posible combinar diferentes criterios de búsqueda, ni es posible aplicar consultas más complejas.

2. Título para el dataset

Datasets de UCI LM Repository disponibles para apoyar las tareas de análisis de un investigador.

3. Descripción de datasets

De acuerdo con la descripción proporcionada en el primer apartado, a través de esta práctica, se han generado dos conjuntos de datos:

Dataset	Método de extracción de datos
Researcher's Profile: Describe los datos básicos de un investigador, o de un conjunto de investigadores asociados a una institución, de acuerdo a su perfil disponible en Google Scholar.	Librería scholarly (v. 0.2.4) permite extraer información de autores y publicaciones de Google Scholar.
UCI Datasets: Describe las características de cada uno de los conjuntos de datos disponibles en el sitio de UCI Machine Learning Repository.	Técnica de scrapy utilizando la librería bs4 de python (Lawson, 2015) (*)

(*) No se encontraron impedimentos para realizar scraping puesto que es una práctica con fines académicos

4. Representación gráfica

La **Figura 1,** presenta el procedimiento general aplicado para poder extraer los datos desde las dos fuentes elegidas: 1) el script gScholar.py utiliza la librería *scholarly* para obtener datos desde *Google Scholar*; y 2) el script uciScraper.py utiliza la librería *bs4* para extraer datos desde *UCI ML Repository.*

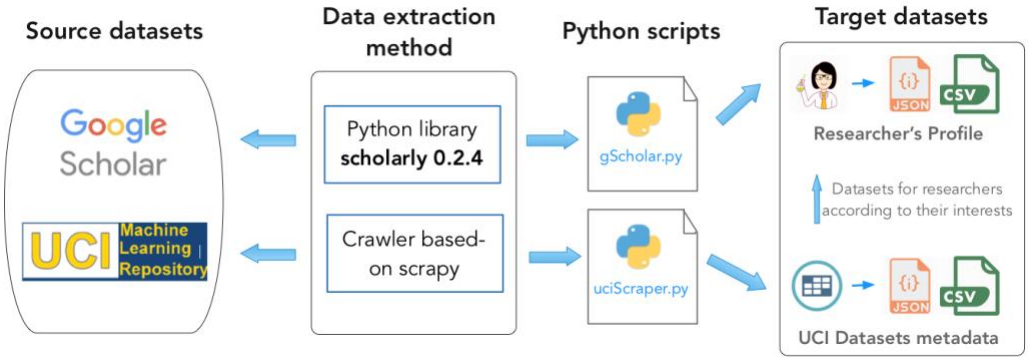

***Figura 1**. Esquema general del proyecto*

Además, como se puede observar en la **Figura 1**, cada conjunto de datos puede ser exportado en dos formatos: CSV y JSON. A partir de la información generada, el siguiente paso sería implementar un motor de recomendación que identifique los datasets que podrían ser útiles a un investigador de acuerdo con sus líneas de interés.

La **Figura 2,** presenta la estructura de cada uno de los conjuntos de datos generados (*Reseacher's profile* y *UCI Datasets*). Al incorporar un tercer conjunto de datos (basado en datos abiertos enlazados -LOD-), el motor de recomendación podría explorar jerarquías de conceptos SKOS para conectar los intereses de los investigadores y las áreas de conocimiento en las que se clasifica cada dataset.

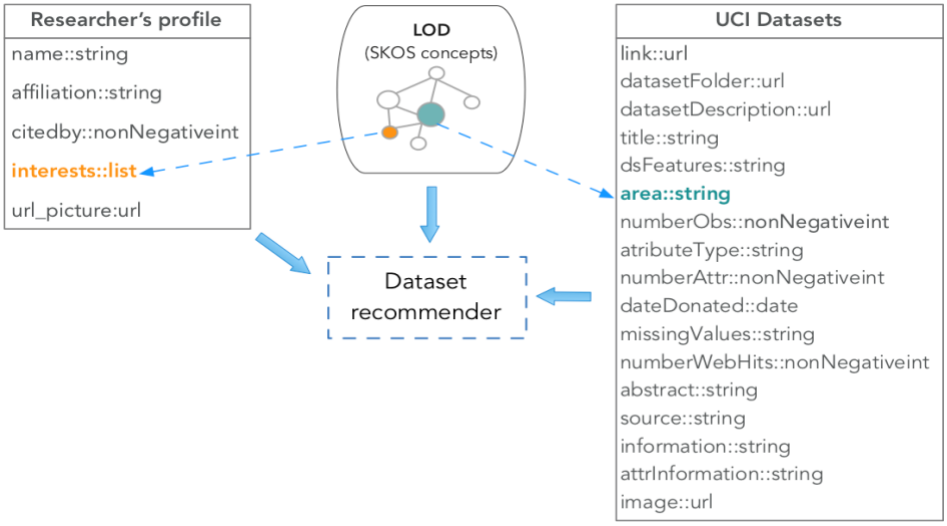

Figura 2. *Estructura de los datasets y proyección para una eventual recomendación*

5. Contenido

A continuación, por cada uno de los datasets creados, se describen sus atributos:

Researcher's Profile:

- name: Nombres del investigador
- affiliation: Nombre de la institución a la que pertenece el investigador.
- citedby: Número de citas del investigador según Google Scholar.
- interests: Conjunto de intereses de un investigador.
- url_picture: Link de la imagen del perfil de un investigador.

UCI Datasets:

- Link: URL de la descripción de dataset disponible en UCI ML Repository
- datasetFolder: Link desde donde se pueden descargar los datos
- datasetDescription: Link del archivo de descripción del dataset
- title: Título del dataset
- dsFeatures: Categoría de las variables del dataset
- area: Área de conocimiento en la que se clasifica el dataset
- numerObs: Cantidad de observaciones
- attributeType: Tipos de datos que conforman el dataset
- numberAttr: Cantidad de atributos del dataset
- dateDonated: Fecha en la que el dataset fue donado a UCI LM Repository
- missingValues: Indica si el dataset contiene valores desconocidos
- numberWebHits: Número de hits del dataset, de alguna manera da una idea de su popularidad.
- abstract: Resumen del dataset
- source: Descripción de los proveedores o autores de los datos.
- information: Provee algunos detalles del dataset
- attrInformation: Descripción de cada uno de los atributos del dataset
- image: Link de la imagen asociada al dataset.

6. Agradecimientos

- A Google, a través de su sitio Scholar y el acceso a la información mediante la librería scholarly.
- A la Universidad de California (UCI) por el acceso estructurado a información de conjuntos de datos habilitados para el análisis.

7. Inspiración

A partir de las dos fuentes generadas, algunos servicios de consulta podrían ser implementados, y así cubrir las siguientes necesidades de información:

- ¿Cuáles son los temas (intereses) en la que los investigadores de una institución dedican sus esfuerzos?
- ¿Por área de investigación, quiénes son los investigadores con mayor impacto?
- ¿Cuáles son los datasets que podrían ser recomendados a los investigadores que trabajan en una determinada temática?
- ¿Qué tipo de análisis se podría aplicar a un determinado dataset?

Al utilizar otro tipo de repositorios o catálogos de datos abiertos basados en CKAN, y explorar las relaciones semánticas entre conceptos disponibles en la Web de Datos, la información inicial recolectada puede ser complementada y enriquecida para así ofrecer servicios más útiles para los investigadores.

8. Licencia

Se ha elegido la licencia **CC BY-NC-SA 4.0 License**. Puesto que este proyecto tiene un fin académico y la idea es contribuir a la generación de conocimiento, la licencia elegida permitirá preservar estos principios, restringiendo el uso comercial de la obra original y de las obras derivadas. Los recursos utilizados serán de consulta libre y no deben generar valor comercial, ya que no fueron creados para este fin.

9. Código

En el proyecto de GitHub [https://github.com/jachicaiza/tipologia-prac1] se puede descargar el código fuente, datos y documentación relacionada al proyecto. En cuanto a los scripts Python, se crearon dos:

- gScholar.py permite buscar, en Google Scholar, investigadores por su nombre o afiliación.
- uciScraper.py obtiene la lista de todos los datasets disponibles en UCI LM Repository y luego accede a la página descriptiva cada uno y extrae sus metadatos.

10. Dataset

Por cada fuente seleccionada, se generan dos formatos de datos JSON y CSV. Los nombres de los archivos se ingresan como parámetro en el script de generación correspondiente.

Nota: las columnas del CSV están separadas por el carácter “|”.

10.1 Datos en formato CSV

Researcher's Profile:

name	affiliation	citedby	email	interests	url_picture
Jordi Conesa i Caralt	Coordinator of the Data Science field at the eHealth Center - Universitat Oberta de ...	808	@uoc.edu	['Analytics', 'e-Learning', 'eHealth', 'Ontologies', 'semantics']	https://scho
Jordi Conesa Caralt	Profesor de Informática, Universitat Oberta de Catalunya	359	@uoc.edu		https://scho

UCI Datasets:

link	datasetFolder	datasetDescription	title	dsFeatures	numberObs	area	atributeType	number Attr	dateDonated	missing Values	number WebHits	image
http://archive.ics.uci.edu/ml/datasets/Auto+MPG	http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/	http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.names	Auto MPG Data Set	Multivariate	398	N/A	Categorical, Real	8	7/7/93	Yes	426201	http://archive.ics.uci.edu/ml/assets/MLimages/Large9.jpg
http://archive.ics.uci.edu/ml/datasets/Economic+Sanctions	http://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/pazzani/	http://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/pazzani/economic-sanctions	Economic Sanctions Data Set	Domain-Theory	N/A	Financial	N/A	N/A	N/A	N/A	52357	http://archive.ics.uci.edu/ml/assets/MLimages/Large153.jpg

10.2 Datos en formato JSON

Para validar el formato de cada dataset, se utilizó el servicio [JSON formatter & Validator](), obteniendo la siguiente vista de cada conjunto.

Researcher's Profile:


```
{ ⊟
   "entities":[ ⊟
      { ⊟
         "name":"Jordi Conesa i Caralt",
         "affiliation":"Coordinator of the Data Science field at the eHealth Center - Universitat Oberta
         "citedby":808,
         "email":"@uoc.edu",
         "interests":[ ⊟
            "Analytics",
            "e-Learning",
            "eHealth",
            "Ontologies",
            "semantics"
         ],
         "url_picture":"https://scholar.google.com/citations?view_op=medium_photo&user=Qx-wNDQAAAAJ"
      },
   ...
}
```


UCI Datasets:

```
{ ⊟
   "dataset":[ ⊟
      { ⊟
         "link":"http://archive.ics.uci.edu/ml/datasets/Auto+MPG",
         "datasetFolder":"http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/",
         "datasetDescription":"http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.names",
         "title":"Auto MPG Data Set",
         "dsFeatures":"Multivariate",
         "numberObs":"398",
         "area":"N/A",
         "atributeType":"Categorical, Real",
         "numberAttr":"8",
         "dateDonated":"1993-07-07",
         "missingValues":"Yes",
         "numberWebHits":"426201",
         "abstract":"Revised from CMU StatLib library, data concerns city-cycle fuel consumption",
         "source":"This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.
         "information":"This dataset is a slightly modified version of the dataset provided in the StatLib library.
mpg.data-original\".\n \"The data concerns city-
cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attri
         "attrInformation":"    1. mpg:           continuous\n    2. cylinders:     multi-
valued discrete\n    3. displacement:  continuous\n    4. horsepower:    continuous\n    5. weight:        continuous
valued discrete\n    8. origin:        multi-
valued discrete\n    9. car name:      string (unique for each instance)",
         "image":"http://archive.ics.uci.edu/ml/assets/MLimages/Large9.jpg"
      },
   ...
}
```


11. Tabla de contribuciones

Contribuciones	Firma
Investigación previa	Janneth Chicaiza Espinosa, Jaime Velandia López
Redacción de las respuestas	Janneth Chicaiza Espinosa, Jaime Velandia López
Desarrollo de código	Janneth Chicaiza Espinosa, Jaime Velandia López

Referencias bibliográficas

Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

Friesike, S.; Widenmayer, B.; Gassmann, O.; y Schildhauer, T. (2015). "Opening science: towards an agenda of open science in academia and industry,"" J. Technol. Transf., 40(4), pp. 581–601.

McKiernan, E. C.; Bourne, P. E.; Brown, C. T.; Buck, S.; Kenall, A.; Lin, J.; McDougall, D.; Nosek, B. A.; Ram, K.; Soderberg, C. K.; Spies, J. R.; Thaney, K.; Updegrove, A.; Woo, K. H.; y Yarkoni, T. (2016). How open science helps researchers succeed," Elife, 5.