PRA 1 - Web Scraping

Tipología y ciclo de vida de los datos

Universitat Oberta de Catalunya (UOC) - *Máster Universitario en Ciencia de Datos*
Asignatura: Tipología y ciclo de vida de los datos
Alumno: Omar Mendo Mesa / Guzmán Manuel Gómez Pérez
Curso: 2020/21 **Fecha entrega:** 08/11/2020

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

La materia del conjunto de datos corresponde con la información (tanto técnica como legal) de las diferentes misiones de **nanosatélites** que se han realizado o en proceso de realizarse a nivel mundial desde que se tienen datos. Entre dichos datos, podemos encontrar tanto información técnica de cada nanosatélite, como puede se:

- Nombre
- Peso
- Tipo de nanosatélite

Y como información de la misión, haciendo referencia al estado de la misma, las organizaciones e instituciones que están detrás de cada misión o el fin de las mismas.

2. Título del dataset

Siguiendo con la dinámica y el contexto del tema tratado, podemos dar como un posible título el siguiente:

- *Dataset****: Misiones de nanosatélites a nivel mundial***

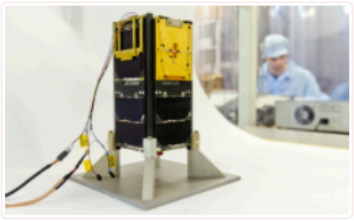
Básicamente, podemos definir este nombre en función del contexto descrito previamente, ya que estamos tratando con la información de cada misión de lanzamiento de un nanosatélite, con sus propios datos técnicos.

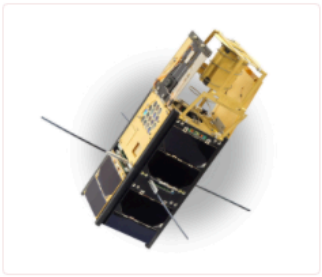
3. Descripción del dataset

Este dataset, como se ha descrito previamente en el contexto, tiene como contenido todos los datos técnicos de cada nanosatélite, así como los datos de su correspondiente misión, pudiendo así tener una visión global del estado, fin e implicados de la misión.

4. Representación gráfica

Como representación gráfica de nuestro dataset, adjuntamos la siguiente imagen:





Name	VZLUSAT-1 (CZ02)
Type	CubeSat
Units or mass	2U
Status	Operational (Official news on 2020-06-30)
Launched	2017-06-23
NORAD ID	42790
Deployer	QuadPack (XL) [ISISpace]
Launcher	PSLV (QB50)
Organisation	Aerospace Research and Test Establishment
Institution	Institute
Entity	Government (Civil / Military)
Nation	Czech
Launch brokerer	ISILaunch
Oneliner	Part of QB50 to study lower termosphere.
Sources	Link-1

Como podemos observar, alguien que no ha indagado o leído sobre el dataset, puede tener una visión global sobre el mismo, observado sobre el objeto del dataset, así como algunos datos sobre los que se están tratando.

5. Contenido

Para cada nanosatélite, el cual se corresponde con un registro en el conjunto de datos, se recoge la siguiente información:

- Name: nombre del nanosatélite.

- Type: Tipo de nanosatélite (Nanosatellite, Picosatellite, Cubesat…)

- Units or mass: Peso del nanosatélite.

- Status: Estado de la misión.

- Launched: Fecha de lanzamiento del satélite en formato YYYY/MM/DD

- NORAD ID: ID de identificación del nanosatélite.

- Deployer: Dispositivo para "instalar" nanosatélites en órbita.

- Launcher: Lanzador de nanosatélites.

- Organization: Organización encargada de la misión del nanosatélite.

- Institution: Institución encargada de la misión del nanosatélite.

- Entity: Entidad encargada de la misión del nanosatélite.

- Nation: País que realiza la misión del nanosatélite.

- Manufacture: Fabricante del nanosatélite.

- Operator: Operador del nanosatélite.

- Launch brokerer: Lanzadera del nanosatélite.

- Costs: Costes de la misión del nanosatélite.

- Oneliner: Objetivo principal de la misión del nanosatélite.

- Failure cause: Causas del fracaso de la misión del nanosatélite.

- Keywords: Palabras clave de la misión del nanosatélite.

- Images: Imágenes del nanosatélite.

6. Agradecimientos

Los datos han sido recolectados desde la base de datos online de **Nanosats EU**. Para ello, se ha hecho uso del lenguaje de programación Python y de técnicas de *Web scraping* para poder extraer la información alojada en el sitio web.

El propietario del sitio web es **Erik Kulu.**

7. Inspiración

El presente conjunto de datos podría utilizarse con gran variedad de fines. Uno de ellos, podría ser, el periodístico, ya que se dispondría, en su totalidad, de la información de los nanosatélites que se han lanzado, que hayan sido pospuestos o cancelados (con sus respectivos motivos), por lo que podría ser de gran utilidad a la hora de recabar información para realizar artículos, ya sea hablando de las tasas de éxito en los últimos años de lanzamientos por un determinado país o del número de universidades alrededor del planeta que están realizando misiones de lanzamiento de estos nanosatélites con fines de investigación.

Otra posible utilidad de ese dataset sería su uso en el campo de la *minería de datos*, con el fin de realizar modelos predictivos para, por ejemplo, determinar si una misión puede resultar exitosa o no en función del país, institución o tipo de satélite del que se esté tratando.

8. Licencia

9. Código fuente y dataset

Tanto el código de la práctica como el dataset generado se encuentran disponibles en el siguiente enlace.

10. Dataset

Una vez se ha subido el dataset generado a Zenodo, obtenemos el DOI del mismo, que es el siguiente:

- **DOI:** 10.5281/zenodo.4233014

Contribuciones	Firma
Investigación previa	Omar Mendo Mesa, Guzmán Gómez Pérez
Redacción de las respuestas	Omar Mendo Mesa, Guzmán Gómez Pérez
Desarrollo código	Omar Mendo Mesa, Guzmán Gómez Pérez

Extras

Con la intención de experimentar con el scrapping de **contenido web dinámico**, se ha ampliado la casuística contemplada en esta práctica a otro escenario adicional. Este consiste en la extracción de datos sobre la mortalidad, recuperación y contagio del COVID19 en todos los países del mundo descargados de una página web de Google. Los campos contemplados son los siguiente:

- Actualmente contagiados (Positives)
- Actualmente contagiados, por millón de habitantes (Positives Per Million)
- Contagiados y recuperados (Recovered)
- Contagiados y fallecidos (Dead)

Su extracción se ha realizado mediante web scrapping, en Python, apoyándonos en el uso de librerías como **Selenium** y Beautiful Soup. El contenido web dinámico se carga conforme el usuario interactúa con la página web, por lo que ha sido necesario descargar en local la totalidad de la página, mediante un Driver del navegador Firefox, antes de poder acceder al conjunto de datos.

El código se encuentra en el siguiente script del repositorio "web_scrapping": **https://github.com/GGP00/web_scrapping/blob/master/src/scrapper.py.** El dataset se ha publicado con el siguiente DOI: **10.5281/zenodo.4260403**

Recursos

- Subirats, L., Calvo, M. (2019). *Web Scraping*. **Universidad Oberta de Catalunya (UOC)**