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

El sitio elegido es https://www.uoc.edu. Se trata de la web de la UOC (Universitat Oberta de Catalunya). Esta universidad ofrece una amplia gama de cursos, masters universitarios y no universitarios, formación de postgrado, etc …

El propósito de su web es doble y depende del área de la web:

- Área pública: Accesible sin restricciones. Su principal función es dar a conocer la oferta formativa del centro.
- Área privada (campus): Se trata de la principal herramienta de acceso que la UOC pone al alcance de sus alumnos para recibir la formación e interactuar con el centro.

Para el desarrollo de la presente práctica y la elaboración del dataset, hemos hecho uso, únicamente, del área pública de la web.

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

UOC Educational Offer.

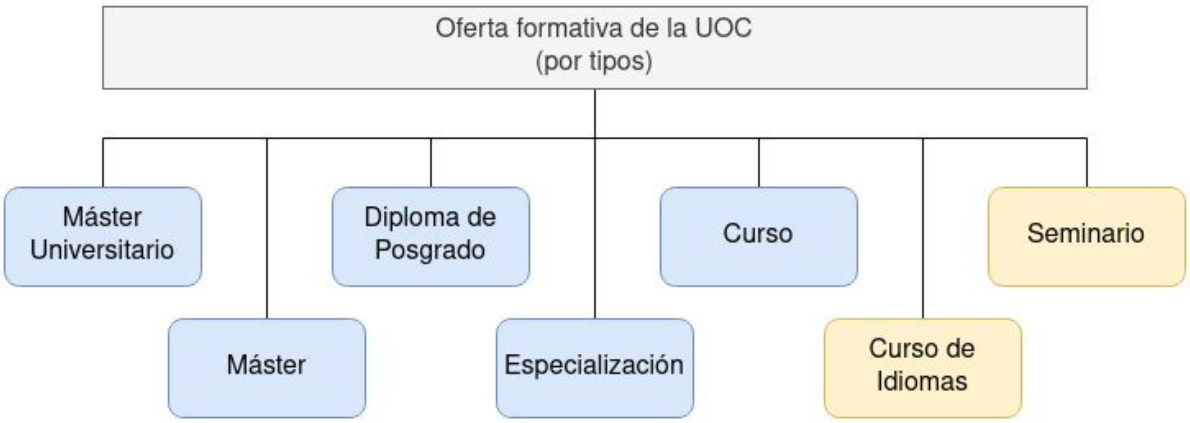
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

Se trata de un fichero CSV que contiene los distintos cursos, másters, másters universitarios, postgrados, especializaciones y seminarios que ofrece la UOC dentro de su oferta formativa pública.

El dataset incluye información relevante con respecto a dicha oferta: títulos, descripción corta, precio, duración, fechas de inicio, etc...

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

El dataset recoge la oferta formativa de la UOC. Esta se puede resumir, por tipos, según el siguiente esquema:



Fuente:

- https://estudios.uoc.edu/ ...
- https://x.uoc.edu/ ...

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

Cada registro del dataset (recurso) se compone de los siguientes campos. En rojo los campos requeridos (todos los registros tienen algún valor válido en estos campos) y en negro los que podrían estar vacíos:

- **type**: tipo de recurso (*máster universitario, máster, diploma de posgrado, especialización, curso, curso de idiomas, seminario)*.
- **name**: nombre del recurso (p.ej: *Máster en dirección de empresas*).

- **description**: breve descripción del recurso.
- **duration**: duración del mismo (1 mes, semestral, etrc...).
- **title**: Título (diploma que otorgan al completarlo, no confundir con el campo *name*).
- **ects**: Cantidad de créditos ECTS que otorga el recurso al ser completado.
- **price**: Precio (€).
- **url**: Url en la que se puede consultar toda la información completa del recurso.
- **date_init**: Fecha de inicio más cercana.

Los datos recogidos componen la oferta formativa de la UOC para el año en curso. Por lo tanto, la validez de los mismos es efímera y cambiará con cada nuevo curso académico.
Los datos se han recogido mediante técnicas de webscraping. Haciendo uso de un programa escrito en Python. La metodología que sigue el programa, para cada uno de los tipos de recurso a obtener (máster, seminario, etc...) es la siguiente:

1. Cargamos la url de entrada al tipo de recurso.
2. Obtenemos las urls individuales a cada uno de los recursos, haciendo uso de la librería lxml y localizando dichas urls mediante Xpath.
3. Cargamos, una por una, las url de cada recurso individual. Cada una de estas páginas contiene los datos importantes que salvamos (título, precio, fecha de comienzo, etc…). Los leemos haciendo uso de lxml, localizando cada uno de ellos mediante Xpath.

Una vez recogida la información de los recursos de interés, volcamos la información al dataset (fichero CSV).
Para más detalles de implementación, véase el README.md del repositorio de GitHub.

NOTA:

Aunque la librería lxml y Xpath están pensados, en principio, para el lenguaje XML, resultan muy útiles y eficientes para analizar y extraer información del código HTML.

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

Los datos presentados son propiedad de la UOC - Universidad Oberta de Catalunya.

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

La oferta formativa de la UOC es rica y muy completa. Su web está bien estructurada, desde el punto de vista de la usabilidad, pero a menudo un mismo recurso es accesible desde varios puntos y es imposible ver a vista de pájaro la oferta completa.
El presente dataset permite filtrar y catalogar sus diferentes recursos. También puede ser útil para elaborar estadísticas sobre el catálogo formativo que se oferta.

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

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License

La licencia elegida para el trabajo realizado es CC BY-NC-SA 4.0 License.
Los motivos por los que elijo esta licencia es:

- Se permite compartir y adaptar los datos.
- Se debe atribuir la propiedad de los datos a su legítimo propietario (la UOC en este caso) se deja claro que el propietario no me cede los derechos sobre los mismos.
- El uso del presente dataset es para uso no comercial. Parece lógico que nadie debería hacer uso de los datos de la UOC con fines comerciales, salvo la propia UOC.
- El resultado de modificar o tratar los presentes datos, debe ser distribuido bajo la misma licencia.

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

El código completo del proyecto, junto con la documentación necesaria, se encuentra en un repositorio de git, en mi cuenta de GitHub:
https://github.com/serfer2/uoc-web-scraper

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

El dataset es accesible en Zenodo, con el DOI *10.5281/zenodo.4205919*.
Url: https://zenodo.org/record/4205919#.X6CSC66CE9I

NOTA: el separador de campo del CSV es el ";" y la *quote policy* del mismo es la *minimal* (solo se entrecomillan aquellos campos de texto que contienen el separador de campo ";"). Es por ello que la previsualización del dataset en zenodo no se ve bien.