Práctica 1: ¿Cómo podemos capturar los datos de la web?

FutureLearn

Estudiantes: Eloy Pérez González y Soulaiman el Hamri

1. Contexto

Se ha conseguido un dataset con información de cursos extraída del sitio https://www.futurelearn.com/ . FutureLearn es una plataforma de educación digital que se ha elegido por la gran variedad y cantidad de cursos que ofrece, que nos ha permitido extraer información de más de 1500 cursos.

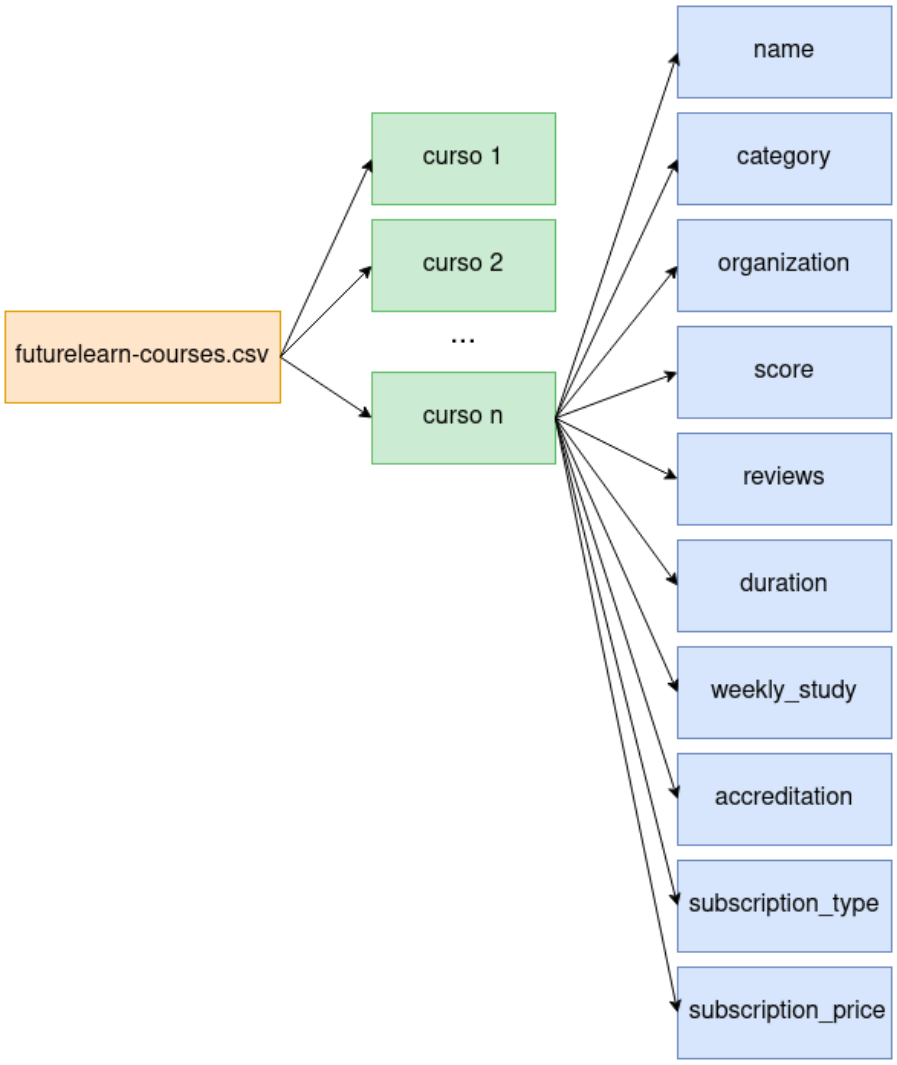
2. Título

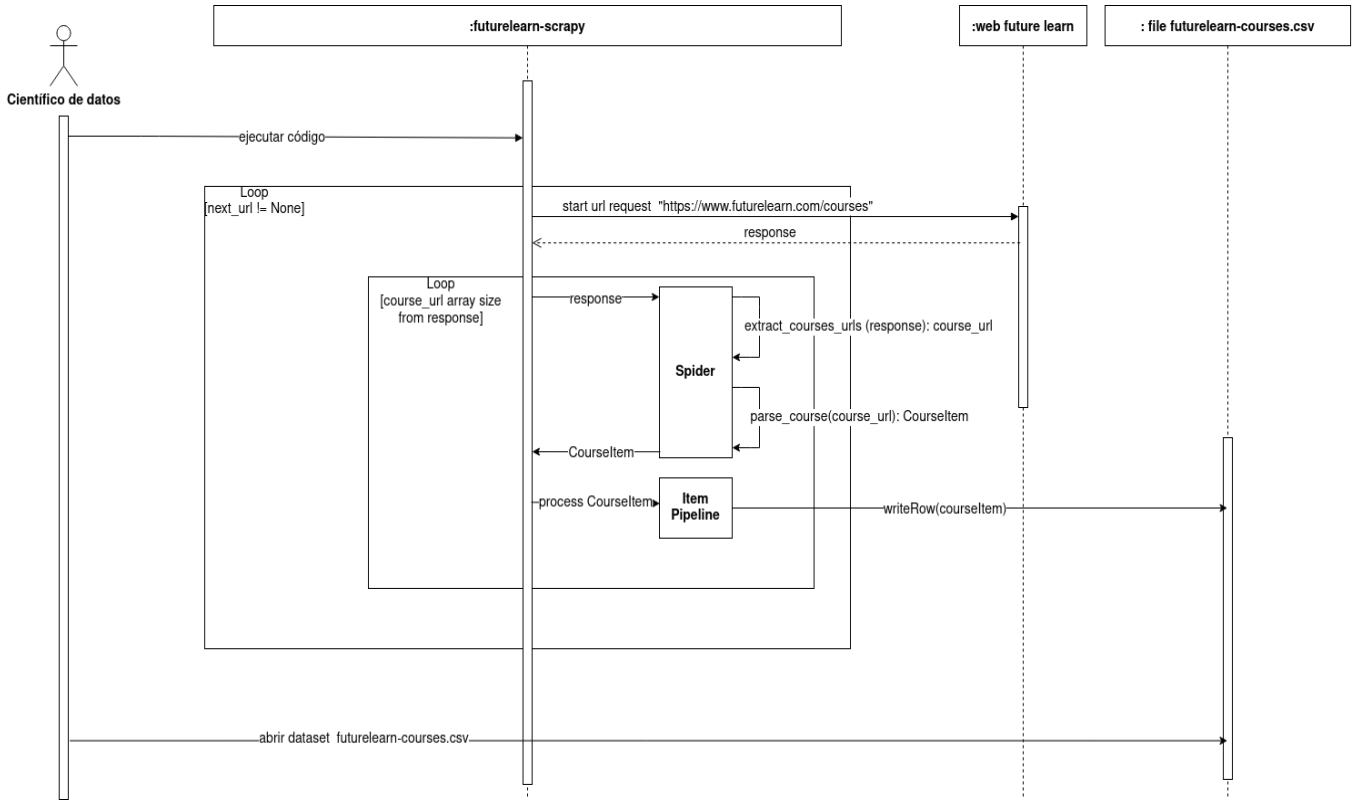
Damos al dataset el nombre de “futurelearn-courses” en referencia al contenido y sitio de donde se ha obtenido la información. Se ha elegido presentar los datos en formato CSV, por lo que el nombre de fichero usado es “futurelearn-courses.csv”.

3. Descripción del dataset

El dataset futurelearn-courses.csv contiene un conjunto de cursos, para los cuales se recoge información como el nombre, la institución que lo organiza, la valoración o el precio.

4. Representación gráfica





5. Contenido

Los campos del dataset son los siguientes:

1. **Accreditation**: Indica si se da una acreditación al realizar el curso.
2. **Category**: La rama de conocimiento a la que pertenece el curso, como puede ser historia, informática o artes.
3. **Duration**: El tiempo que dura el curso.
4. **Name**: El nombre del curso.
5. **Organization**: La institución que imparte el curso.
6. **Reviews**: La cantidad de evaluaciones por parte de los usuarios que tiene el curso.
7. **Score**: La puntuación media que los usuarios han dado al curso.
8. **Subscription_type**: El tipo de subscripción necesaria para el curso. Algunos entran dentro de una tarifa plana en la plataforma y otros se pagan exclusivamente.
9. **Subscription_value**: El precio de la subscripción asociada al curso. Puede ser un valor único para los cursos que se pagan individualmente o por mes para las tarifas planas.
10. **Weekly_study**: El tiempo requerido de estudio cada semana.

6. Propietario

El conjunto de datos contiene información extraída de Futurelearn, que es una plataforma de educación digital. Fundada en 2012 por varias universidades británicas como la universidad de Birmingham o el King's College London. Actualmente es propiedad de The Open University y Seek Ltd.

Para ver si el sitio web tiene algún método para prevenir el web scraping lo primero que se ha hecho es consultar el archivo robots.txt, donde se ha visto que no hay ningún problema para el contexto del proyecto. A continuación se ha visto que no tienen ningún mecanismo de control de acceso como por ejemplo CAPTCHA, y por último se han inspeccionado los elementos HTML para ver que no habría ninguna dificultad añadida para obtener la información.

7. Inspiración

El conjunto de datos elegido nace de la curiosidad por responder preguntas sobre la efectividad y valoración de los cursos online, que se pueden contraponer en ocasiones a otros cursos más formales como podría ser una carrera universitaria.

El dataset obtenido permite responder a preguntas como que tipo de cursos son los más valorados, y como influyen las distintas variables como el precio, la duración o la temática del mismo.

8. Licencia

En los términos de uso de la página, apartado "Website use" se explicita:

"You are not allowed to commercialise our website or the content on it (i.e. you are not allowed to make money or attract advertising to another business by using our website)."

Por tanto se elige para el dataset la licencia **Creative Commons 4.0 BY-NC-SA**, que especifica:

- Atribución: Se debe hacer referencia al autor.
- Uso no comercial: No se puede usar el dataset para fines comerciales.
- Compartir igual: Los trabajos derivados del dataset deben usar la misma licencia.

9. Código

El código se puede encontrar ubicado en la carpeta /source del repositorio https://github.com/SulaimanUOC/futureLearnScraper.

Para realizar el proceso de recolección de datos de la web se ha usado la librería Scrapy (Phyton).El código consta de diversos componentes que interaccionan entre sí. Destacar que en items.py se define el modelo de datos que se va a obtener, en nuestro caso CourseItem. En los spiders se implementa el código que se encarga de parsear y extraer los datos html

obtenidos de la web y por último, en pipelines.py es donde se define el procesamiento del item una vez extraído por un spider.

La mayor dificultad que presenta el sitio web es que no todos los cursos siguen una estructura homogénea en el código html/css, ya que hay particularidades que hacen un poco más difícil la obtención de los datos en función del curso que se trate.

10. Dataset

A continuación el link del dataset publicado en Zenodo:
https://doi.org/10.5281/zenodo.7332123

El dataset se encuentra en la carpeta
https://github.com/SulaimanUOC/futureLearnScraper/tree/main/dataset del github.

11. Vídeo

El vídeo se encuentra en el siguiente enlace:

https://drive.google.com/file/d/1EWtNuMOP-YGJalzpqcm1VpxF1We1Adpa/view?usp=sharing

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	Eloy; Soulaiman
Redacción de las respuestas	Eloy; Soulaiman
Desarrollo del código	Eloy; Soulaiman
Participación en el video	Eloy; Soulaiman