Práctica 1: Web Scraping

Tipología y ciclo de vida de los datos

JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

Tipología y ciclo de vida de los datos

Práctica 1: Web Scraping

1. Wiki donde haya los nombres de los componentes del grupo y una descripción de los ficheros:

   https://github.com/jgomezgar/Tutor-UOC/wiki

2. Un documento Word, Open Office o PDF con las respuestas a las preguntas y los nombres de los componentes del grupo.

Miembros del equipo:

Este desarrollo ha sido realizado de manera individual por **José Luis Gómez García** (jgomezgar@uoc.edu).

1. Título del dataset. Poned un título que sea descriptivo.

   Tutor-UOC

2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.

   Seguimiento de alumnos UOC tutorizados

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente.

4. Contexto. ¿Cuál es la materia del conjunto de datos?

Este script desarrollado en python 3 (leer README), recorre *las webs de herramientas de tutoría*, produciendo una serie de conjuntos de datos relacionados, que **incorporan la información que necesita un tutor de la UOC para el seguimiento de sus alumnos tutorizados**, estos conjuntos de datos son:

- Planes Tutorizados.
- Ficha simple de Alumno.
- Última Conexión al Campus.
- Asignaturas Cursadas en el semestre.
- Fechas y calificaciones de PECs.

Las webs de herramientas de tutoría: Presentan varias dificultades para web scraping directo:

1. Protegidas por usuario y contraseña.
2. Necesidad de un token de sesión válido.
3. Código HTML generado por JS durante la ejecución de la web.
4. Acceso solo mediante perfil de tutor.

Ha sido necesario usar selenium y el driver de Chrome para poder gestionar estas dificultades.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?

Planes Tutorizados:

Dataset: PLANES.csv.

Periodo de Validez: Semestral, coincidiendo con el curso lectivo.

Frecuencia/Momento de actualización: Al inicio del semestre Lectivo.

Campos:

- PLAN : único campo, contiene los código de los planes tutorizados.

Ficha simple de Alumno.

Dataset anonimizado: ALUMNOS.csv.

Periodo de Validez: Semestral, coincidiendo con el curso lectivo.

Frecuencia/Momento de actualización: Al inicio del semestre Lectivo. Debido a incorporaciones tardías, es posible que se necesiten varias actualizaciones en los primeros meses del curso.

Campos:

- Ficha_code: Código del Alumno, identificador principal del alumno.
- Semestre: Semestre cursado.
- Plan_code: Código del plan cursado en el Semestre.
- Alt_email: email alternativo (anonimizado debido al carácter público de este dataset).
- email: email facilitado por la UOC (anonimizado debido al carácter público de este dataset).
- Ap_Nom: Apellidos y Nombre (anonimizado debido al carácter público de este dataset).
- Ingreso: Semestre de Ingreso en el Plan.
- Nuevo: Flag con valor 1 si es nueva incorporación en el semestre.

Última Conexión al Campus.

Dataset: CONEXIONES.csv.

Periodo de Validez: Semanal.

Frecuencia/Momento de actualización: Semanal, despues del fin de semana. El comportamiento habitual de nuestros estudiantes es aprovechar el fin de semana para ponerse al día con sus estudios.

Campos:

- Ficha_code: Código del Alumno, identificador principal del alumno.
- Ultima_Conex: Momento (Fecha y hora) de la última conexión al Campus Virtual.
- Ultima_Accion: Momento (Fecha y hora) de la última acción/desconexión al Campus Virtual.

Asignaturas Cursadas en el semestre.

Dataset anonimizado: ASIGNATURAS.csv.

Periodo de Validez: Semestral, coincidiendo con el curso lectivo.

Frecuencia/Momento de actualización: Al inicio del semestre Lectivo. Debido a incorporaciones tardías, es posible que se necesiten varias actualizaciones en los primeros meses del curso.

Campos:

- Ficha_code: Código del Alumno, identificador principal del alumno.
- User_Id: Identificador alternativo del alumno.
- Semestre: Semestre cursado.
- Asign_Nom: Nombre de la Asignatura.
- Asig_Code: Código de la Asignatura.
- Aula: Número de aula de la Asignatura.
- Asign_Prof: Profesorado del aula (anonimizado debido al carácter público de este dataset).

Fechas y calificaciones de PECs.

Dataset: PECS.csv.

Periodo de Validez: Semanal.

Frecuencia/Momento de actualización: Semanal, despues del fin de semana. Atendiendo a las fechas de PECs, las fechas de entrega y publicación de Evaluaciones, se suelen fijar entre la última hora del Fin de semana y los primeros días de semana.

Campos:

- Ficha_code: Código del Alumno, identificador principal del alumno.
- User_Id: Identificador alternativo del alumno.
- Semestre: Semestre cursado.
- Asig_Code: Código de la Asignatura.
- Aula: Número de aula de la Asignatura.
- PEC: Nombre de la Prueba de Evaluación Continua.
- PEC_Fecha: Fecha tope de entrega de PEC.

- PEC_Entrega: Fecha real de entrega de PEC.
- PEC_Nota: Calificación de PEC.
- PEC_Publicacion: Fecha de publicación de calificaciones de la PEC.

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

Todos estos datos son propiedad de la UOC. Están organizados por medio de las herramientas facilitadas al tutor, una serie de webs, que concentran listados de alumnos con enlaces a:

- Ficha del estudiante.
- Seguimiento académico del estudiante.
- Seguimiento de la actividad del estudiante.

Sin estas herramientas sería imposible el correcto seguimiento del alumno.

Análisis Anteriores:

He participado activamente en un grupo para "*La mejora de herramientas y espacios de tutoría*", no tengo constancia de la existencia de análisis previos ni por parte de la UOC, ni de terceros, ni de otros tutores, siquiera, a título personal, de ningún sistema ,web scraping, para mejorar el acceso a la información de UOC, enfocada a las necesidades de los tutores.

Por otra parte, hace ya algunos años, cree un sistema de web scraping, sobre estos mismos datos, desarrollado en VBA, mediante macros en excel. Al poco tiempo el sistema quedó inservible, debido a los cambios en la seguridad y sobretodo a la inclusión de HTML generado en tiempo de ejecución por código Java Script (JS-Rendered).

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

Inspiración.

Llevo varios años trabajando como tutor en la UOC, específicamente, como tutor del máster de Inteligencia de Negocio y Big Data, y unos de las cosas que más me llamó la atención fué que no existiese un sistema BI, que facilitase el análisis de la actividad de los alumnos.

Dentro de este análisis, uno de los temas más preocupantes es el **abandono**, teniendo en cuenta que existen indicadores, muy conocidos, para detectar el abandono, como son:

- Largos periodos sin conectarse al campus.
- Retrasos en las PECs.
- Suspensos en las PECs.

Sin embargo, detectar alumnos con estas características, es muy complicado. Por ejemplo, personalmente, cada semestre, suelo tener más de 100 alumnos, y estos suelen cursar 4 asignaturas, Por lo que para recabar esta información es necesario realizar más de 500 consultas en la web de la UOC. Y este es un seguimiento que para realizarse correctamente, debería ser semanal. Por otro lado, este tipo de alumnos, no suelen revisan el correo electrónico de la UOC, por lo que utilizar el correo alternativo del alumno, puede marcar la diferencia a la hora de ayudar a un alumno. Pero este dato, no es fácil de gestionar con las herramientas facilitadas por la UOC.

Por tanto son básicos los siguientes datos:

- Email externo
- Última conexión al Campus.
- PECs

Cuestiones:

- **¿Por qué es interesante este conjunto de datos?:**

  Es interesante porque unifica en un único lugar, toda la información necesaria para la monitorización de los estudiantes.

- **¿Qué preguntas le gustaría responder la comunidad?:**

1. La principal pregunta con la que responder con estos datos es "*qué alumnos están en peligro de abandono*", además permite el contacto con estos usuarios al facilitar el correo electrónico personal de estudiante.
2. Otra interesante cuestión, es cómo poder enviar mensajes personalizados a los estudiantes según sus patrones de comportamiento y otros datos de su perfil. Teniendo esta información es posible realizar segmentaciones, por notas, por retrasos, por frecuencia de conexión, por asignaturas, por país,.... Una vez segmentados, es fácil enviar un mensaje personalizado a cada grupo: Felicitando por notas, preocupándose por retrasos, ...

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado.

El **Software** (código) está protegido por "*MIT License*", que básicamente permite cualquier uso sin restricciones: Software LICENSE

- Aunque existe una limitación impuesta por el propio portal web, que ya requiere usuario, contraseña y permisos de Tutor en el campus de la UOC, la motivación principal de este tipo de Licenciamiento es que cualquier persona (tutor UOC) pueda utilizar, mejorar y compartir libremente este código.

Los **datasets** están protegidos por "*CC0 1.0 Universal*", que no limita el uso de estos datos: Dataset LICENSE

- El motivo de este tipo de licenciamiento "Universal", es debido a que por un lado, la anonimización de datos sensibles y la falta de resultados PEC, debido al momento inicial del curso, ofrecen pocas posibilidades frente a la generación de los datos personalizados para cada tutor.

3. Una carpeta con el código Python o R generado para obtener los datos.

https://github.com/jgomezgar/Tutor-UOC/tree/master/CODIGO

4. El fichero CSV con los datos.

https://github.com/jgomezgar/Tutor-UOC/tree/master/CSV