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

Tipología y ciclo de vida de los datos

Miguel Ciriano Martín y Pedro A. Pérez Pérez

Tipología y ciclo de vida de los datos

Práctica 1

noviembre de 2024

Información clave:

Nombres de los integrantes: Miguel Ciriano Martín y Pedro A. Pérez Pérez

Enlaces a los sitios web base:

- https://planderecuperacion.gob.es/como-acceder-a-los-fondos/convocatorias (página de búsqueda de ayudas en España, tanto a nivel estatal como autonómico)
- https://www.pap.hacienda.gob.es/bdnstrans/GE/es/inicio (página con el histórico de ayudas concedidas identificadas con granularidad máxima, caso a caso)

Enlace al repositorio de código en GitHub: https://github.com/DataCiriano/Web-Scraping

Enlace al dataset:

https://zenodo.org/records/14060197?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImExZTlmZmZmLTVlYzItNGUwNS04MmEyLTQ3MWFhMmNiZThiMiIsImRhdGEiOnt9LCJyYW5kb20iOiI2MGVjYmQxNjFlY2JmNzEwOTY5ODdlYWU5NDNmZjcyZiJ9.NaJ6zCx2_EApvyV5JrEY2pIbB5q0kOun87jfZteAJwWUu4XYpQJY8cvH9RH-HshmOt8gj1OQ08-hXO94WNmTHA

Enlace al vídeo:

https://drive.google.com/drive/folders/1BFtXEExB9nrwZE6F1CjbCcGgXeXugpHP?usp=sharing

1. Introducción y contexto sobre los *datasets*

1.1. Títulos de los *datasets*

1. Ayudas para empresas del gobierno español central y autonómicos.
2. Concesiones de Ayudas del Estado.

1.2. Descripción

Ayudas para empresas del gobierno español central y autonómicos.

Este *dataset* contiene información de las licitaciones / ayudas estatales y autonómicas publicadas más recientemente junto con su ámbito de aplicación, importe y links a fuentes con mayor detalle, entre otros datos.

Concesiones de Ayudas del Estado

El conjunto de datos extraído mediante web scraping para la Práctica 1 de la asignatura de Tipología y Ciclo de Vida de los Datos contiene información pública sobre las subvenciones y ayudas entregadas por las diferentes administraciones de las comunidades autónomas de España. Algunas de las variables que contiene el dataset son la fecha de la concesión, el ente que concede la subvención o la ayuda y los beneficiarios de las mismas entre otros datos.

1.3. Contexto

Ayudas para empresas del gobierno español central y autonómicos.

Para que las empresas puedan optar a las ayudas y subvenciones públicas, el primer paso es la comunicación. Por este motivo, desde el gobierno, se ha puesto a disposición un sitio web donde poder acceder a los detalles de cada convocatoria existente, ya bien sea una abierta, cerrada o por abrir. La recolección de estos datos navegando a cada uno de los links y obteniendo la información deseada de esas nuevas páginas web para agregarla en una base de datos compacta creemos que aporta mucho valor a la hora de poder encontrar la que mejor se adapte a cada caso particular. La información se extrae del sitio web del Plan de Recuperación, Transformación y Resiliencia del Gobierno de España en conjunto con la Unión Europea.

URL: https://planderecuperacion.gob.es/como-acceder-a-los-fondos/convocatorias

Concesiones de Ayudas del Estado

Las diferentes administraciones autonómicas convocan regularmente diferentes ayudas y subvenciones que pueden ser solicitadas por los ciudadanos y son otorgadas si estos cumplen con los requisitos necesarios. Tanto la convocatoria como la resolución son información de carácter público y como tal queda recogida en los diferentes organismos de comunicación y plataformas, tanto regionales como a nivel nacional.

En este caso particular, para la construcción del conjunto de datos se realizó un web scraping sobre la web pública del Sistema Nacional de Publicidad de Subvenciones y Ayudas Públicas perteneciente al Ministerio de Hacienda del Reino de España.

Url: https://www.pap.hacienda.gob.es/bdnstrans/GE/es/inicio

2. Descripción de los *datasets*

La información de ambos *datasets* es de carácter público y la compartimos bajo la licencia de "Creative Commons Zero v1.0 Universal"

2.1. Contenido

Ayudas para empresas del gobierno español central y autonómicos.

Cada registro se corresponde con una convocatoria abierta para ofrecer ayudas a las empresas o para licitar un servicio.

- **Tipo de convocatoria:** Tipo de oferta que abre el Estado o las autonomías para las empresas.
- **Título:** Descriptivo de la actividad a realizar o ayuda a repartir.
- **Órgano convocante:** Qué entidad concreta lanza la oferta.
- **Localización:** Ubicación, si aplica, de la entidad convocante o la actividad a realizar.
- **Importe:** Cuantía de la oferta
- **Procedimiento:** Estado de la convocatoria (Abierto, Abierto simplificado…)
- **Actividad:** Ámbito de aplicación de la ayuda: Sector, tipo de actividad, objeto de la licitación…
- **Enlace al detalle de la convocatoria:** Autoexplicativo, link de acceso a más detalle de la convocatoria, habitualmente ubicado en las páginas concretas de cada entidad ofertante.

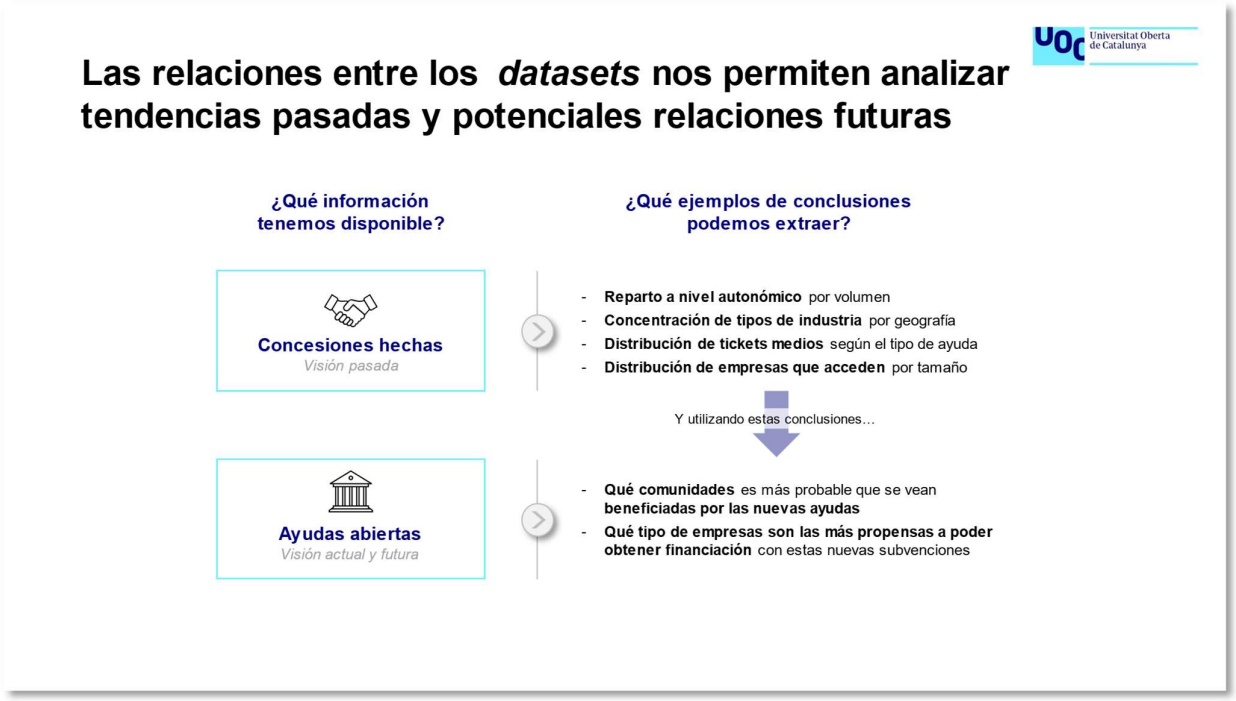
Concesiones de Ayudas del Estado

Cada registro se corresponde con una subvención o ayuda entregada.

- **Convocante:** Entidad o administración pública que ha convocado la ayuda o subvención.
- **Convocatoria:** Nombre o título específico de la convocatoria bajo la cual se ha otorgado la ayuda.
- **Código BDNS:** Identificador único asignado por la Base de Datos Nacional de Subvenciones (BDNS) a cada ayuda o subvención.
- **Reglamento:** Normativa o disposición legal que regula la concesión de la ayuda.
- **Objetivo de la ayuda:** Propósitos o metas que la ayuda pretende alcanzar.
- **Instrumento de la ayuda:** Modalidad o tipo de apoyo proporcionado, como subvenciones directas, préstamos, garantías, etc.
- **Tipo de empresa:** Clasificación de la empresa beneficiaria según su tamaño o características, como microempresa, pyme o gran empresa.
- **Fecha de concesión:** Fecha en la que se aprobó la concesión de la ayuda.
- **Código de concesión:** Número o identificador específico asignado a la concesión individual de la ayuda.
- **Fecha de registro:** Fecha en la que la concesión fue registrada en la BDNS.
- **Beneficiario:** Nombre de la persona física o jurídica que recibe la ayuda.
- Importe nominal: Cantidad total de dinero asignada en la concesión de la ayuda.
- **Ayuda equivalente:** Valor de la ayuda expresado en términos de subvención equivalente, especialmente relevante en ayudas que no son subvenciones directas.
- **Región concesión:** Ubicación geográfica donde se aplica o tiene efecto la ayuda concedida.
- **Sector de actividad NACE:** Clasificación del sector económico del beneficiario según la Nomenclatura Estadística de Actividades Económicas de la Comunidad Europea (NACE).
- **Referencia de la medida:** Código o identificador de la medida específica bajo la cual se enmarca la ayuda.
- **Entidad art. 16 GBER:** Entidad relacionada según el artículo 16 del Reglamento General de Exención por Categorías (GBER) de la Unión Europea.
- **Intermediario art. 21 GBER:** Intermediario financiero o entidad que participa en la concesión de la ayuda según el artículo 21 del GBER.

2.2. Representación gráfica e inspiración

Adjuntamos una *slide* donde se puede apreciar las relaciones entre ambos *datasets* y las posibilidades de análisis posteriores.



3. Código

Proceso de recolección de datos para "Ayudas para empresas del gobierno español central y autonómicos".

1. **Configuración del WebDriver**: El código configura el WebDriver de Chrome en modo headless para ejecutar el navegador sin interfaz gráfica, lo que es muy útil para ejecutar el script en segundo plano sin consumir demasiados recursos. También se añaden opciones para optimizar el rendimiento y evitar problemas de compatibilidad.
2. **Navegación y Extracción de Enlaces**: Para cada página de la lista de convocatorias, el script navega a la URL correspondiente y espera que todos los elementos de las convocatorias carguen correctamente en el DOM utilizando una espera explícita (*WebDriverWait*). A continuación, recoge las URLs individuales de cada convocatoria en la página mediante la clase "*c-convocatoriasResult"*.
3. **Extracción de Datos Específicos**: Una vez que tiene el enlace de cada convocatoria, el script navega a esa página de detalles y utiliza una función para obtener información detallada de cada convocatoria. Para cada uno de los elementos, el script emplea selectores XPath específicos, lo cual permite identificar cada campo en el HTML de manera precisa e inequívoca, ya que al principio no los identificaba correctamente.
4. **Almacenamiento de los Datos**: Tras completar el proceso de recolección, el script organiza los datos en un DataFrame de pandas y los exporta a un archivo CSV, asegurando que los datos recolectados sean fácilmente accesibles y analizables.

Dificultades del Sitio Web y Soluciones Implementadas

1. **Error de Referencia a Elementos Obsoletos (Stale Element Reference)**: Al cambiar de página entre la lista de convocatorias y la página de detalles, los elementos localizados previamente se volvían obsoletos, causando errores. Para solucionar esto, se implementó una lista de URLs recolectadas primero, y luego se accedió a cada enlace de manera individual, evitando que el DOM se actualice mientras se procesan los elementos.
2. **Variabilidad en la Estructura de los Atributos**: El enlace a los detalles de la convocatoria no siempre se encontraba en el mismo atributo o con un texto identificador claro. Para resolver este problema, se creó una función específica para manejar estos enlaces y se usaron diferentes selectores XPath, como buscar enlaces que contuvieran el texto "convocatoria" en el atributo href.
3. **Carga Asíncrona de Elementos**: Para garantizar que el contenido estuviera completamente cargado antes de interactuar con los elementos, se emplearon esperas explícitas (WebDriverWait). Esto permitió que el script esperara hasta que todos los elementos de interés estuvieran presentes en el DOM antes de proceder a extraer los datos.

Descripción del funcionamiento del código de "Concesiones de Ayudas del Estado":

1. **Configuración del WebDriver**:

   La función setup_driver() configura el controlador de Chrome con un User-Agent personalizado para simular una conexión desde un navegador real, evitando bloqueos de acceso por parte del sitio web. El navegador se inicia con un tiempo de espera implícito de 10 segundos y carga la URL deseada.

2. **Cierre del Banner de Cookies**:

Se espera hasta que el botón del banner de cookies sea interactivo y se hace clic en él. Esto garantiza que el banner no interfiera con las interacciones posteriores del script.

3. **Selección del número de elementos por página**:

   Para mostrar 1000 elementos por página, el script desplaza la vista al selector correspondiente y hace clic en él. Después, selecciona la opción de 1000 elementos, esperando 20 segundos para asegurar que la página se cargue completamente.

4. **Navegación a páginas específicas**:

   La función go_to_page(page_number) permite navegar hasta una página determinada. Si la página deseada no está disponible, selecciona la página más alta visible y repite el proceso hasta que la página objetivo esté accesible. Se han implementado intentos múltiples y esperas para mejorar la robustez.

5. **Extracción de datos de la tabla**:

   La función extract_data() extrae datos de la tabla, ignorando la cabecera. Maneja StaleElementReferenceException y TimeoutException para reintentar la extracción en caso de que el DOM cambie o la carga sea lenta. Los datos se guardan en listas para ser exportados a archivos CSV.

6. **Extracción de un rango de páginas**:

   extract_range(start_page, end_page) recorre páginas especificadas y guarda los datos cada 10 páginas en archivos CSV. Si no se encuentran datos en una página, el proceso se detiene.

7. **Combinación y limpieza de datos**:

8. Se combinan todos los CSV generados en un solo DataFrame de pandas, eliminando duplicados y guardando el resultado final en un archivo CSV consolidado.

<u>Dificultades del sitio web y soluciones:</u>

- **Cambio del número de elementos por página**: El sitio requiere hacer scroll en el selector antes de elegir la opción de 1000 elementos. Esto se resolvió usando ActionChains para desplazar la vista y asegurar la visibilidad del elemento.
- **Navegación a páginas altas**: Para alcanzar páginas con números altos, es necesario seleccionar la página más alta disponible repetidamente hasta que aparezca la página deseada. El código implementa un bucle con intentos múltiples para manejar este comportamiento.
- **Carga de la página**: Los tiempos de carga variables pueden causar fallos en la extracción. Para esto, se han agregado tiempos de espera (time.sleep()) y reintentos para garantizar que la página se cargue correctamente antes de continuar.

Contribuciones	Firma
Investigación previa	MCM, PPP
Redacción de las respuestas	MCM, PPP
Desarrollo del código	MCM, PPP
Participación en el video	MCM, PPP