Pregunta 1. Contexto.

La web elegida ha sido la perteneciente a la Cámara de Cuentas de Andalucía, que es un órgano técnico dependiente del Parlamento de Andalucía, al que corresponde la fiscalización externa de la gestión económica, financiera y contable de los fondos públicos de la Comunidad Autónoma de Andalucía.

Durante todo el año, esta Institución realiza fiscalizaciones que se materializan en informes que son aprobados, tras unos controles de calidad y el principio de contradicción[1], y enviados a otros organismos y publicados en el Boletín Oficial de la Junta de Andalucía (en adelante, BOJA).

Estas fiscalizaciones nacen del Plan de Actuaciones Anual, y éste es basado en el Plan Estratégico, que suele abarcar 5 ejercicios.

Existen actualmente cuatro departamentos de fiscalización:

- Junta de Andalucía: Cuyo ámbito es el gobierno autonómico.
- Organismos y Empresas Públicas.
- Entidades Locales.
- Coordinación: Los informes emanados directamente de Presidencia.

El objetivo de esta práctica es la obtención de los informes definitivos aplicados junto con sus códigos de identificación. Como valor añadido, en los casos donde existía esta información, se ha añadido el número de BOJA junto con su contenido, y el acceso a los informes completos y resumidos.

La herramienta genera las siguientes salidas:

- Un dataset por cada departamento en formato CSV.
- Un dataset de todos los departamentos en formato CSV.
- Un dataset de todos los departamentos separados por hojas en formato Excel.

Esta información resulta útil para dar cumplimiento de las necesidades de las propias entidades fiscalizadas, y a los ciudadanos y las organizaciones que velan por la transparencia, la lucha contra el fraude y la corrupción, y que hacen uso de la legislación de transparencia en vigor[2].

En cuanto al aspecto técnico, se debe destacar que:

- El archivo de robots.txt facilita la búsqueda desde un buscador tipo Google.
- La información es accesible directamente en la web sin usuario/contraseña. Sin embargo, no sigue un criterio de ordenación claro, lo que dificulta la localización de un informe en concreto.

  Por ejemplo, la página principal de informes de auditoría del departamento de Junta de Andalucía es: https://www.ccuentas.es/junta-de-andalucia. Se paginan los informes de 10 en 10, donde cada página es https://www.ccuentas.es/junta-de-andalucia/1 hasta la última. Sin embargo, si se escribe otro número fuera del rango, también responde. Si se emplea cualquier tipo de localizador de enlaces, entra en bucle porque el redirector lo acepta y lo selecciona el contenido del final.

- Existe un sistema de redirección que entorpece el empleo de herramientas de búsqueda de enlaces tanto simples como complejas. Debido a esto, se ha tenido que realizar un "buscador" y un "iterador" ad hoc.
- Para facilitar la mejora de la explotación de los datos, se han sometido a un proceso de curación y normalización.

Pregunta 2. Título de DataSet.

La herramienta lanza seis DataSets. Cuatro de ellos recolectan los informes por departamento en formato CSV y los otros dos, todos los informes tanto en formato CSV como en Excel.

El título para los DataSet completos es:

CCA_**All_Departments**_Reports_Details_Dataset

Los títulos para los DataSet por departamento son:

- CCA_**Coordinación**_Reports_Details_Dataset
- CCA_**Corporaciones_Locales**_Reports_Details_Dataset
- CCA_**Junta_de_Andalucía**_Reports_Details_Dataset
- CCA_**Organismos_y_Empresas_Publicas**_Reports_Details_Dataset

Pregunta 3. Descripción de DataSet.

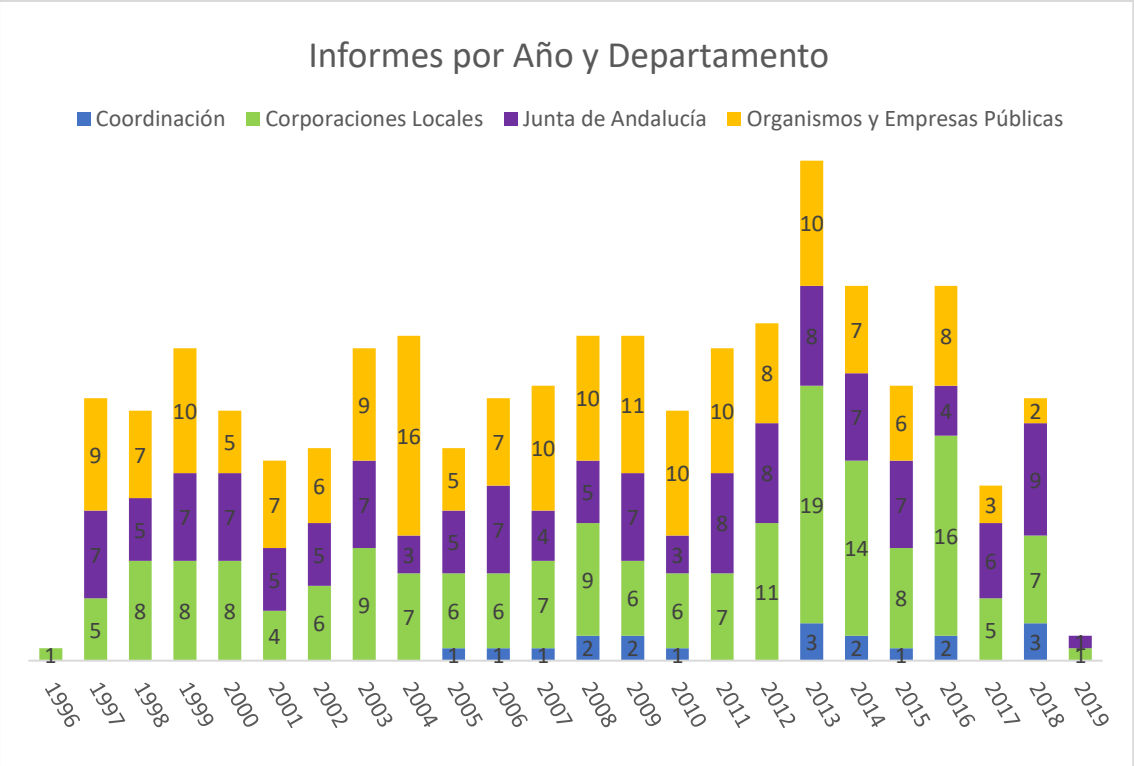
Los DataSet recolectan información sobre los informes de fiscalización publicados de la Cámara de Cuentas de Andalucía, junto con: su código identificativo, su título, el número de BOJA que le corresponde[3], el tiempo que se estima ha sido necesario para su realización, el acceso a los informes completos, resumidos y al BOJA correspondiente[4], y el año que empezó la fiscalización.

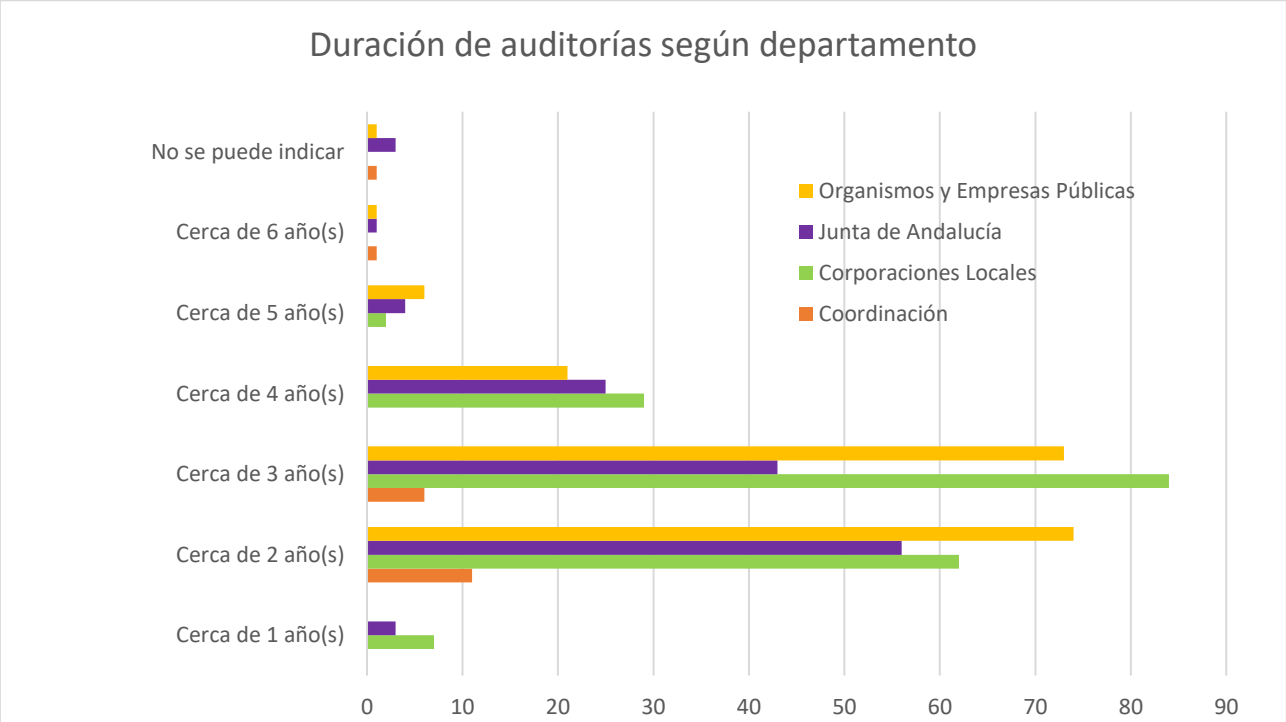
De ahí que el nombre de los DataSet siga el siguiente esquema:

- CCA: Cámara de Cuentas de Andalucía (el organismo en cuestión).
- En negrita: Los departamentos en cuestión o todos los departamentos en el caso de *All_Departments*.
- Reports_Details: Detalles de los Informes.

[1] Comúnmente llamado alegaciones.
[2] Estatal: Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno.
Autonómica: Ley 1/2014, de 24 de junio, de Transparencia Pública de Andalucía.
[3] Cuando éste aparezca en la web.
[4] Si están accesibles vía la web.

Pregunta 4. Representación gráfica.





	Cerca de 1 año(s)	Cerca de 2 año(s)	Cerca de 3 año(s)	Cerca de 4 año(s)	Cerca de 5 año(s)	Cerca de 6 año(s)	No se puede indicar
Organismos y Empresas Públicas		74	73	21	6	1	1
Junta de Andalucía	3	56	43	25	4	1	3
Corporaciones Locales	7	62	84	29	2		
Coordinación		11	6			1	1

Pregunta 5. Contenido.

A continuación, se describe los campos de los DataSet.

Campo	Descripción
Index	Número y correlativo.
Department	Nombre del departamento.
Code	Código identificativo del informe de fiscalización.
Title	Nombre descriptivo del informe de fiscalización.
BOJA_Published	Número y fecha de publicación en BOJA[5].
Time_Auditing	Un cálculo aproximado de tiempo necesario desde el comienzo de la actuación hasta su publicación en BOJA[6].
Complete	Acceso al informe completo de fiscalización[7].
Resume	Acceso al resumen del informe de fiscalización[8].
BOJA_Link	Acceso al BOJA donde se publica dicho informe[9].
Year_of_Audit	Año correspondiente al inicio de las actuaciones.

Teniendo en cuenta que estos datos:

- Son relevantes para:
  - Las entidades auditadas y aquellas relacionadas con éstas que requieran de su información,
  - los equipos de auditoría de la propia Cámara de Cuentas de Andalucía, del Tribunal de Cuentas del Estado, y en ocasiones, por el Tribunal de Cuentas de las Comunidades Europeas, y
  - para los ciudadanos en general;

Y sin perjuicio de la relevancia legal que tiene todo documento oficial publicado en el boletín correspondiente, el período de tiempo cubre desde la obligación legal de publicación en los medios de exteriorización de la Institución hasta el último.

Para facilitar su búsqueda, se han ordenado todos los *DataSets* por el año de publicación de forma descendente, es decir, desde el más reciente al más antiguo.

Seguidamente se describe cómo se han recogido dichos datos:

Primer Paso: Responder a ¿dónde estaban los datos de los informes?

Se ha investigado dónde se encontraban los datos, y el resultado fue que estaban ubicados en las cuatro *urls* principales siguientes, una por cada departamento:

- https://www.ccuentas.es/junta-de-andalucia/
- https://www.ccuentas.es/corporaciones-locales/
- https://www.ccuentas.es/organismos-y-empresas-publicas/

[5] En algunos casos, no aparece la publicación del BOJA.
[6] En algunos casos, no aparece la publicación del BOJA, por lo cual, no puede calcularse.
[7] En algunos casos, no está accesible.
[8] En algunos casos, no está accesible.
[9] En algunos casos, no está accesible.

- https://www.ccuentas.es/coordinacion/

Mediante la investigación, se descubrió que los resultados eran paginados por cada 10 elementos irremediablemente, dando como resultado que a la *url* principal se le añadía "/número de página", por ejemplo: https://www.ccuentas.es/junta-de-andalucia/1.

Prosiguiendo por esta línea de actuación, se empleó código extraído del libro *Web Scraping with Python*, en relación a la búsqueda de enlaces y de mapas de la web, descubriendo que, dicha web, está sustentada por un sistema de redireccionamiento que causa el efecto de bucle continuo. En detalle, en una inspección visual de la web, en el caso de https://www.ccuentas.es/junta-de-andalucia/ aparecía que el límite era la página 14, sin embargo, se podría escribir https://www.ccuentas.es/junta-de-andalucia/55 que daba resultados. Lo que ocurre es que tras la página 14, los resultados eran un calco de ésta.

Así que se construyó la función `LookForDelimeter` donde se procesa la página principal de cada departamento en búsqueda de la información visual que indica cuál es la última página web de la paginación.

Para ello se emplea `BeautifulSoup`, buscando al principio el **`id`** de `pagination`[10], y posteriormente, dentro de ese objeto, se recolecta todas las etiquetas de tipo **`a`**. A través de un proceso de transformaciones, se obtiene el número límite de las paginaciones.

Por último, se diseñó un iterador (`Iterator`), de tal forma que se genera una lista con todas las *urls* necesarias por departamento.

Llegado a este momento, se puede responder dónde están los datos.

Paso segundo: Responder a ¿cómo recojo los datos de cada página de informes de fiscalización?

Para responder esta pregunta, se ha desarrollado una función que genera un CSV por departamento (`GeneratingCSVbyDepartment`) a partir de la información extraída de la web mediante la función `WebProcessing`. Finalmente se agrupan los DataSet de cada departamento para obtener un DataSet único y global.

En detalle:

Para el proceso de la web, se emplea el módulo `BeautifulSoup`, a través de un *parser* de *html*.

- Para obtener el nombre del departamento, se busca la etiqueta H1 que lo almacena (`department_name = soup.h1.text.strip()`)

Para el resto de elementos, se hace una búsqueda de la tabla (`table = soup.find('table')`), dentro de ella, de la fila (`table_rows = table.find_all('tr')`) y en el interior de cada fila, se recolectan los datos de cada celda (`td = tr.find_all('td')`). A continuación, se indica cómo se obtienen:

- Para el código de la actuación (*Code*), directamente se invoca al contenido del primer elemento de la lista de **`td`** (`code_name = DataCuration(td[1].contents)`)

[10] Se supo esta etiqueta gracias a la inspección del código web.

- Para el año de comienzo de la actuación (*Year_of_audit*), se aplica la búsqueda de un patrón sobre el valor de *Code*. (`code_year_pattern = re.compile(r'\d{1,4}$')` `code_year = code_year_pattern.findall(code_name)`)
- Para los informes (*complete, resume, BOJA_link)* se buscan los enlaces existentes en cada fila y se crea una lista con los valores del atributo **`href`** (`for a_items in tr.find_all("a"): a_item.append(a_items.attrs["href"])`), luego a través de una serie de búsquedas en las cadenas dentro de la lista **a_item**, se identifica a qué documento se refiere.
- Para comprobar si ha habido publicación de BOJA (*BOJA_Published*), ha sido más laborioso. Requiere comprobar si existe como tal en **`td[2],`** empleando patrones de búsqueda de los dos formatos más habituales (número y fecha, o año/número), y luego un proceso de transformación para que si es el primer caso, aparezca siguiendo este patrón (núm. NUM, DIA/MES/AÑO) con un año formado por cuatro dígitos.
- Para obtener el título del informe (*Title*), simplemente se encuentra en **`td[2]`** y se obtiene así:
  `title_name = DataCuration(td[2].contents)`. En los casos en que el BOJA no está disponible como enlace, la información del BOJA forma parte del texto del título por lo que se procede a eliminarla empleando un patrón específico `title_name = re.sub("\(BOJA n.m. \d{1,3},?\s{0,1}\d{1,2}[-,/]\d{1,2}[-,/]\d{1,4}\)", "", title_name)`
- Para la obtención del tiempo de duración (*Time_auditing*), se aplica un patrón al campo *boja_published* extrayendo el año y restándolo a la variable *code_year*[11] y sumándole uno. Ya que, si empieza y acaba en el mismo año, no tendría sentido que indicara que la duración sea 0.

En todo momento se ha procedido a:

- Eliminar espacios en blanco.
- Dar la información lo más homogénea posible, por ejemplo, las fechas.
- Y en caso de no obtener el dato, por su carencia, proceder a indicarlo con un mensaje que aporte valor al usuario de la herramienta.

Pregunta 6. Agradecimientos.

Para realizar esta práctica, se ha pedido permiso al Jefe de Servicio de Informática de la Cámara de Cuentas de Andalucía como responsable del mantenimiento de la plataforma donde se encuentra. Siendo otorgado éste, sin problemas.

Los datos, al ser documentos oficiales y de carácter público, no han requerido permiso para obtenerlos.

Previamente a esta herramienta, toda investigación o análisis ha sido manual.

Para no causar "ruidos" a los sistemas de análisis de tráfico, se ha cambiado el *user-agent* a uno descriptivo para esta práctica: `user_agent = {"User-Agent": "Practica-Web-Scraping"}`.

[11] Es obtenida de aplicar un patrón al campo *Code*.

Pregunta 7. Inspiración.

Los informes de fiscalización de la Cámara de Cuentas de Andalucía, como el resto de Instituciones de Control Externo, proporcionan información de alto valor añadido ya que comprenden:

- Cumplimiento legal.
- Análisis financiero (desde una perspectiva no sólo económica).
- En ocasiones, un análisis operativo si se están realizado los procesos, los programas, las acciones, etc. bajo los principios de economía, eficiencia y eficacia. Además, de mostrar el alineamiento con los principios de buen gobierno y transparencia.
- Capacidad de mejora continua para las entidades auditadas.

El acceso a dichos informes de forma cómoda, es interesante:

- A las entidades auditadas.
- Al Parlamento de Andalucía.
- A los Tribunales de Cuentas del Estado y Europeo.
- A otros instrumentos de control.
- A la ciudadanía general y a las asociaciones de transparencia.

Este conjunto de datos, responden a:

- ¿Qué informes han sido publicados y, por lo tanto, son definitivos y aprobados?
- ¿Dónde puedo consultarlos?
- ¿En qué BOJA se encuentran?
- Y a título estadístico, ¿aproximadamente cuánto tiempo transcurrió entre su inicio y su definitiva publicación en BOJA?
- Además, de un detalle que facilita las búsquedas y que consiste en ordenarlos en el año de su inicio (del más reciente al más antiguo).

Pregunta 8. Licencia.

Para la publicación del dataset hemos seleccionado una licencia de Creative Commons, todas las licencias de esta organización aseguran que los creadores o licenciadores de una obra sean reconocidos como autores. Además es vigente en todo el mundo y su duración equivale a la duración de los derechos de propiedad intelectual aplicable.

Dentro del abanico de licencias que se ofrecen hemos escogido la conocida como Reconocimiento-CompartirIgual 4.0 Internacional (**CC BY-SA 4.0**) por considerar que es la que mejor se adecúa a nuestras necesidades.

Esta licencia permite las siguientes acciones:

- Compartir la obra. Se permite copiar y redistribuir el material en cualquier medio o formato.
- Adaptar la obra. Se permite reutilizar, transformar y construir sobre el material para cualquier propósito, incluso comercial.

Bajo las siguientes condiciones:

- Atribución al autor. Se debe otorgar el crédito apropiado, proporcionar un enlace a la licencia e indicar si se realizaron cambios. De ninguna manera se debe sugerir que el autor respalda a la persona o al uso que hace dicha persona de la obra.
- Compartir igual. Si se reutiliza la obra se deben distribuir las contribuciones bajo la misma licencia que el original.
- Sin restricciones adicionales. No se puede aplicar términos legales o medidas tecnológicas que impidan legalmente a otros hacer todo lo que la licencia permita.

Consideramos que siempre es positivo contribuir al crecimiento del patrimonio digital, mediante la divulgación y la reutilización de las obras para crear valor añadido. Además, al tratarse de documentos oficiales y de carácter público consideramos básico que exista la máxima accesibilidad a los datos, con el fin de promover la transparencia.

Esta licencia obliga a distribuir la obra bajo la misma licencia por lo que nos aseguramos que la información se siga redistribuyendo, tal y como queremos.

Además, consideramos de gran relevancia proteger los derechos de autor en cualquier obra. Esto implica reconocer la autoría del creador y exonerarlo de cualquier responsabilidad si un tercero hace un uso inapropiado del material publicado.

Pregunta 9. Código.

Se adjunta con la práctica el fichero `CCA_Reports_Details_Scraper.py` en el repositorio de GitHub.

Pregunta 10. DataSet.

Tal como se ha indicado, se genera un DataSet por departamento y dos generales y se adjuntan en el repositorio de GitHub. Se ruega que, a efectos de evaluación, se contemple el DataSet de todos los departamentos en formato de CSV:

CCA_All_Departments_Reports_Details_Dataset.csv.

El resto queda a efectos de valor añadido.

Contribuciones

Contribuciones	Firma
Investigación previa	Inés Caro Molina, Ángel Carrasco Núñez
Redacción de las respuestas	Inés Caro Molina, Ángel Carrasco Núñez
Desarrollo de código	Inés Caro Molina, Ángel Carrasco Núñez