Mº Begoña Martinez Arribas / Silvia Martín Albarrán

22/03/2020

Tipología/Ciclo de vida de los datos

PRA1- Web Scraping

Práctica nº 1: Web Scraping

22-3-2020

Tipología/Ciclo de vida de los datos

PRA1- Web Scraping

Practica 1 [35% Nota Final]

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes por un proyecto analítico y usar las herramientas de extracción de datos.
Para hacer esta práctica tendréis que trabajar en grupos 2 personas. Tendréis que entregar un solo fichero con el enlace Github (https://github.com) donde haya las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos de vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Podéis mirar estos ejemplos como guía:

- **Ejemplo:** https://github.com/rafoelhonrado/foodPriceScraper
- **Ejemplo complejo:** https://github.com/tteguayco/Web-scraping

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- **Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para resolverlo.**
- **Capacidad para aplicar las técnicas específicas de web scraping.**

Objetivos

Los objetivos concretos de esta práctica son:

- **Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.**
- **Saber identificar los datos relevantes que su tratamiento aportan valor a una empresa y la identificación de nuevos proyectos analíticos.**
- **Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.**

- **Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios) y mediante diferentes mecanismos (tales como queries, API y scraping).**
- **Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.**
- **Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.**

Descripción de la Práctica a realizar.

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

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

Respuesta:

Puesto que el objetivo de la práctica era de extraer un juego de datos a través de web scraping y era permitido el acceso mediante APIs o mediante "rascado", planteamos varios sitios que podrían servirnos

Los criterios han sido diversos: que no fuera demasiado complicado extraer los datos, que no haya problemas de uso de los datos por temas legales y que el juego de datos pudiera servirnos para hacer el gráficos a mostrar.

Se ha analizado distintas opciones , y algunas fueron descartadas porque cuandos se revisaba el fichero robot.txt se veía que no estaba permitido el web scraping y en otros casos, por dificultad de cómo estaba construida la página web o si la empresa no facilita un API Key para acceso a sus datos..Entendemos que el objetivo de la práctica es entender los distintos modos de acceder a datos que se ofrecen en un site y los conceptos básicos de Web Scraping.

Hemos contactado con el AEMET, TripAdvisor, IDEALISTA para sus APIs y consultado páginas de datos abiertos. También webs de listados de catálogos de películas (Netflix, filmaffinity) , transportes (www.renfe.es, www.alsa.es), de revistas , la bolsa del IBEX 35 etc.. Al final y de cara a poder hacer diversos ejemplos ,en vez de fijarnos en un único sitio,hemos propuesto un menú que permitirá explorar diferentes datos y ámbitos:

- OPCIÓN 1 : AGENDA ALCOBENDAS (CULTURA)

En la página web de datos abiertos de Alcobendas (https://datos.alcobendas.org/pages/informacion-datos-abiertos) se ponen a disposición distintos juegos de datos con distintos catalogos y que se puede acceder leyendo el JSON que proporicionan. Después de ver algunos de los juegos de datos, se decidió por la agenda

de eventos culturales, al perfil al que van dedicados y los datos de cuando y donde se ofrecen. También se puede ver la ficha de dicho evento. La tipología de los datos es variable.

El uso que se puede dar a estos datos puede ser para mostrarlos en apps dirigidas a los ciudadanos para informar de los distintos eventos culturales de dicho Municipio o desde el punto de vista análitico, si se fueran recuperado cada cierto tiempo , se podría ver comparativas de a quien se oferta más eventos etc.

- OPCIÓN 2: CURSOS DE LA COMUNIDAD DE MADRID DE FORMACION PROFESIONAL (EDUCACION)

En la página web de datos abiertos de la Comunidad de Madrid (https://www.comunidad.madrid/gobierno/datos-abiertos) se ponen a disposición distintos juegos de datos con distintos catálogos y que se puede acceder leyendo el JSON que proporcionan.

En este caso, existen diversos juegos de datos . Interesante los listados de cursos , por la importancia que tiene para poder publicarlos y que gente desempleada o trabajadores puedan acceder a un catálogo para poder apuntarse o inscribirse a tiempo..

En este caso se ha seleccionado un de cursos de formación profesional, que para la práctica solo sacaremos los primeros cursos .No se ha sacado completo pero si los campos asociados.

- OPCIÓN 3: REVISTAS CIENTÍFICAS (CIENCIA)

En este caso: https://www.clasificacioncirc.es/ . Clasificación de Revistas Científicas

Se seleccionó esa página no por el juego de datos , sino que buscando listados de revistas y con objeto de utilizar la librería Beautiful Soup (BS4) , se vió que la página HTML tenía una estructura muy sencilla de explotar.

Esta clasificación tiene su importancia en el ámbito científico y como en la propia web indican “Clasificación Integrada de Revistas Científicas – CIRC, nace con el fin de generar un instrumento de medida común que sea utilizado por evaluadores, investigadores y grupos de investigación sobre bibliometría, facilitando realizar comparaciones y compartir información.”. [Consultado el día 25/03/2020]

- OPCIÓN 4: MIL ANUNCIOS (CATALOGO/TABLÓN ANUNCIOS Y PRODUCTOS)

En esta opción se lleva a cabo una implementación de web scraping sobre la página web de https://www.milanuncios.com.

Al tratarse de un catálogo/tablón de anuncios nos ha parecido que ofrece una estructura HTML interesante para la realización de la práctica. Al tener diferentes niveles de anidamiento se puede acceder desde la página principal contenedora de todos los productos/anuncios hasta la página de detalle de cada anuncio concreto e incluso dentro de este detalle a sus a datos específicos.

Además, la página contiene varios parámetros que podrían seleccionarse para enriquecer la búsqueda por diferentes conceptos.

A su vez el juego de datos que contiene que incluye tanto cuantitativos como cualitativos permite agrupaciones para su visualización.

Por agilidad y para optimizar la prueba, aunque se recupera el número total de páginas total desde el inicio, se han acotado estas a un número fijo que garantice variabilidad en la información recuperada en un tiempo breve.

- OPCIÓN 5: IDEALISTA (PORTAL INMOBILIARIO)

Mediante la selección de esta opción el objetivo es la extracción de información de la página web https://www.idealista.com/ a través de su API. Para ello y una vez autorizado el acceso y recibida la clave de dicha API, hemos revisado la estructura de los datos que ha confirmado ser una buena opción de extracción por contener una amplia variedad de información.

Ciertamente uno de los puntos que más nos ha interesado además de la riqueza de los datos es el acceso por OAuth 2.0.

Los parámetros de búsqueda permiten flexibilizar esta en base a los diferentes filtros que se consideren. Entre los datos que facilita la API se encuentra de tipo cuantitativo y cualitativo o categórico.

- OPCIÓN 6 : ÍNDICE DE CLASIFICACIÓN SCIMAGO (CIENCIA/RANKINGS)

En este caso https://www.scimagojr.com/journalrank.php pasó lo mismo que con el de las revistas científicas, se buscaba información o rankings para hacer "rascado" (web scraping) e utilizar la librería Beautiful Soup . En esta ocasión tiene más campos que la opción 3 y la estructura HTML es distinta. Admite más filtrados y ordenación y para poder hacer un estudio de cómo varía el índice de impacto de las revistas y el nº de veces que se cita et. son válidos para hacer rankings y ver la evolución de dichos indicadores.

Uso por la comunidad científica.

Ejemplo de uso de una visualización que tienen en la página :

https://www.scimagojr.com/shapeofscience/

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

Respuesta:

- OPCIÓN 1 : AGENDA EVENTOS CULTURALES ALCOBENDAS
- OPCIÓN 2 : CATÁLOGO CURSOS PARA FORMACIÓN PROFESIONAL EN LA COMUNIDAD DE MADRID
- OPCIÓN 3: CLASIFICACIÓN CIRC - REVISTAS CIENTÍFICAS
- OPCIÓN 4: TABLON ANUNCIOS PRODUCTOS 2ª MANO Y PISOS
- OPCIÓN 5: CATALOGO INMOBILIARIO VENTA Y ALQUILER
- OPCIÓN 6 : RANKING DE REVISTAS CIENTÍFICAS DE ESPAÑA EN 2018 EN BASE AL ÍNDICE DE IMPACTO SCIMAGO (SJR)

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).

Respuesta:

- OPCIÓN 1 : AGENDA ALCOBENDAS

  Este juego de datos recoge la información esencial sobre los distintos eventos culturales programados en Alcobendas e indicando a qué perfiles va dirigido, horarios e ubicación de los mismos.

- OPCIÓN 2: CURSOS DE LA COMUNIDAD DE MADRID

  Este juego de datos recoge una parte de los cursos ofertados por la Comunidad de Madrid para la formación profesional, indicando la especialidad , área, duración , si se emite certificado de la especialidad con el fin de informar a los ciudadanos de qué cursos están disponibles y la información del centro donde se impartirá.

- OPCIÓN 3: REVISTAS CIRC

  Este juego de datos (parcial) está basado en la clasificación CIRC de las distintas revistas científicas para uso bibliométrico .

- OPCIÓN 4: TABLÓN DE ANUNCIOS MILANUNCIOS

  Juego de datos basado en un catálogo de anuncios sobre diferentes temáticas (inmobiliaria, empleo, servicios, ocio, etc) así como venta y alquiler de productos de 2ª mano para intercambio. El dataset extraído filtra por la temática de Inmuebles y tipología pisos, con las características de estos (nombre, precio, localización, descripción).

- OPCIÓN 5: PORTAL INMOBILIARIO IDEALISTA

  Juego de datos relacionado con la venta y alquiler de diferentes tipos de inmuebles y anuncios relacionados para este tipo de operaciones/intercambios. El dataset seleccionado filtra por ventas de inmuebles a una distancia determinada dentro de España.

- OPCIÓN 6 :ÍNDICE DE CLASIFICACIÓN SCIMAGO (CIENCIA/RANKINGS)

  Este juego de datos está basado la extracción del listado de revistas/conferencias científicas ordenados por el ranking del índice de impacto y de referencia denominado SJR , que utilizan en la comunidad científica para saber la relevancia de determinados trabajos, revistas o ponencias, en este caso de España y año 2018.

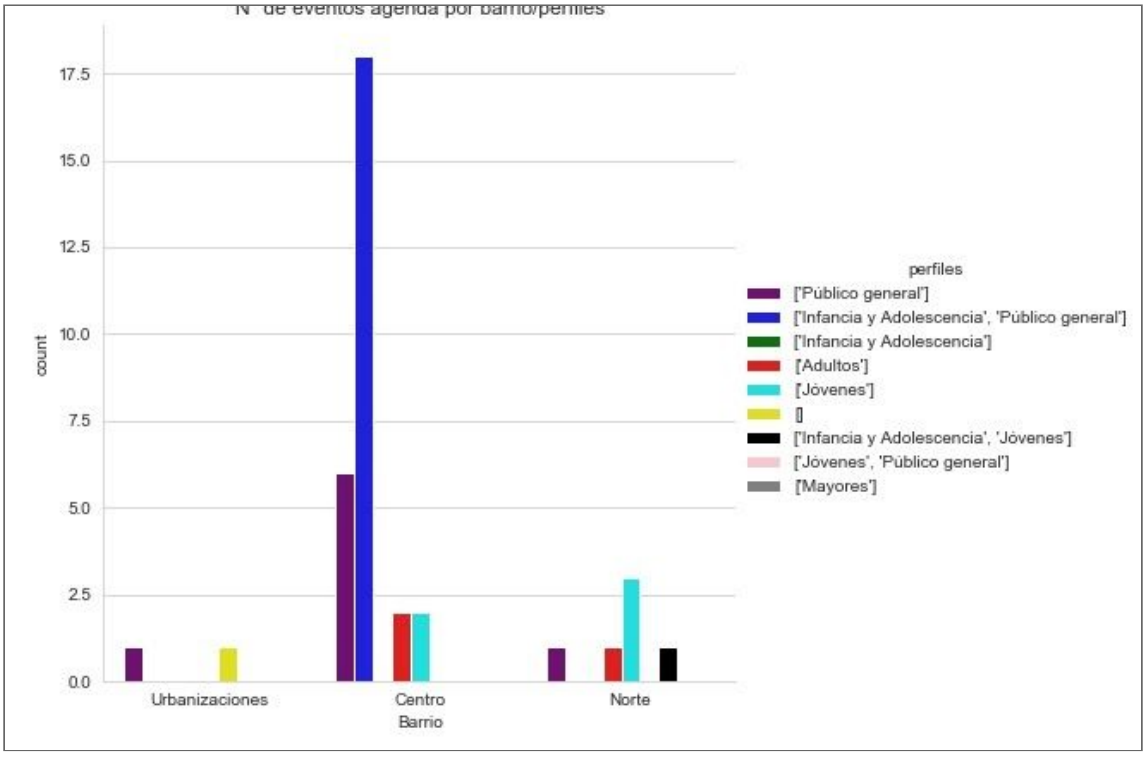
4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.

Respuesta:

- OPCIÓN 1 : AGENDA ALCOBENDAS (CULTURA)

  Se puede por ejemplo extraer la información del nº de eventos en los distintos barrios de Alcobendas y los perfiles a los que están dirigidos.

  Como se observa en el gráfico , el mayor nº de eventos se da en el barrio "CENTRO" y para público general o infancia/Adolescencia, mientras que para los mayores se ofertan menos eventos.

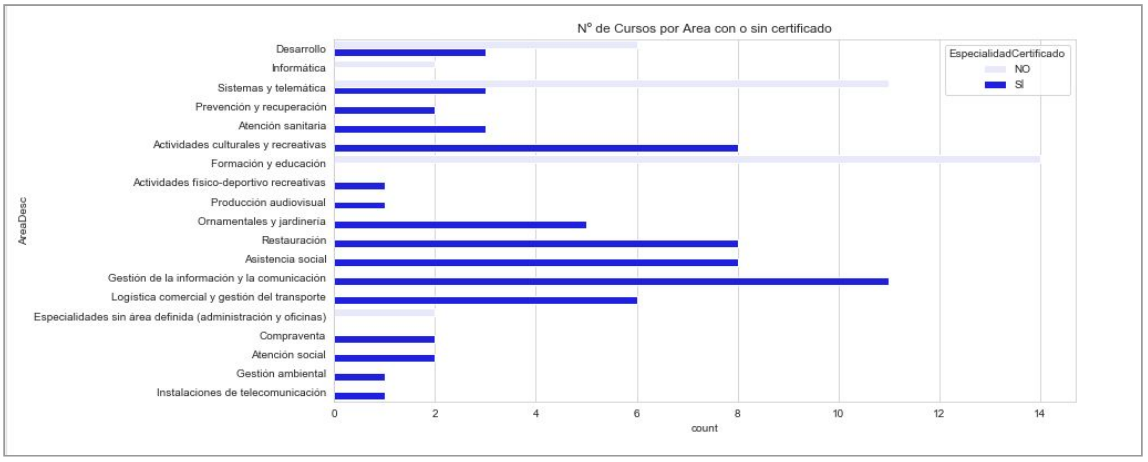


- OPCIÓN 2: CATÁLOGO CURSOS COMUNIDAD DE MADRID PARA FORMACIÓN PROFESIONAL

  NOTA IMPORTANTE: En el juego de datos no se recupera en código todo el listado completo sino unos 100 registros , por lo que el gráfico no es del todo completo.

  Se indican el nº de Cursos por Área y si se emite certificado de especialidad

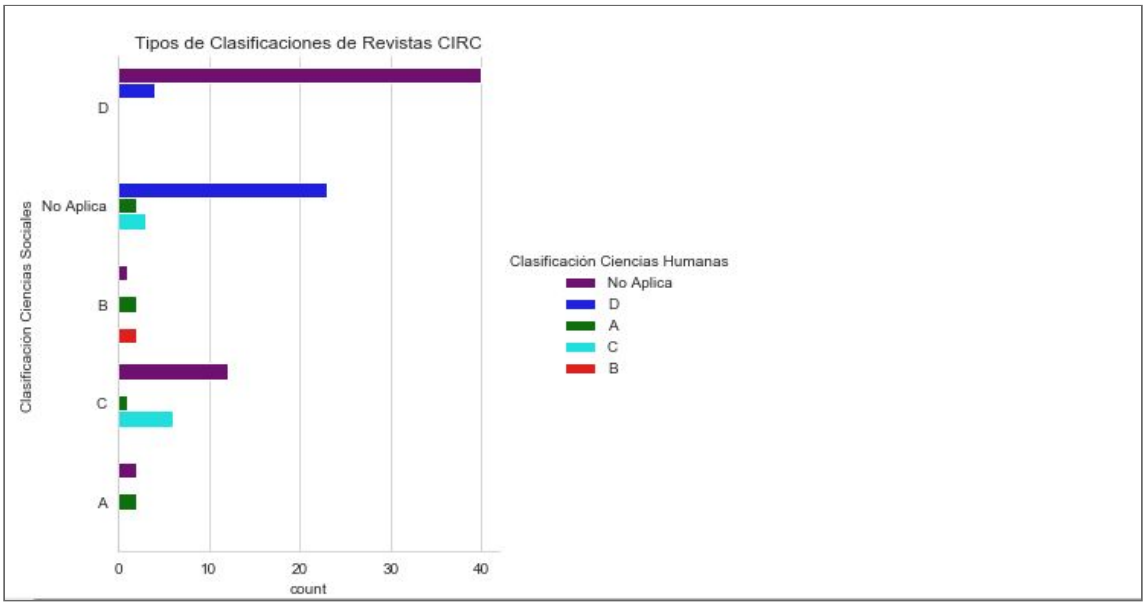
  Según el gráfico, en logística de transporte es en el area donde mas cursos se ofertan con certificado de especialidad.



- OPCIÓN 3: REVISTAS CIRC

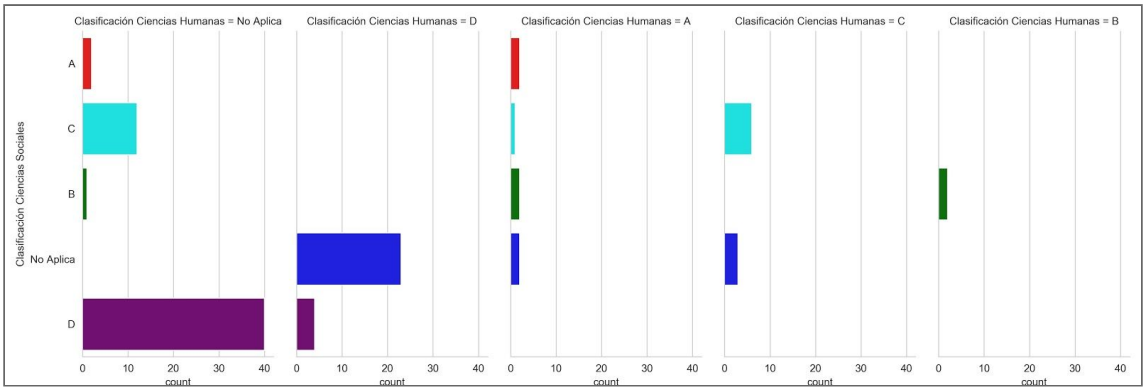
  Este gráfico indica que la mayoría de las revistas que se clasifican para Ciencias Sociales, NO aplica para Ciencias Humanas.

```
grafico= sn.catplot(y = "Clasificación Ciencias Sociales", hue= "Clasificación Ciencias Humanas",kind="count", data = df_dataset,  palette=sn.color_palette(['purple', 'blue','green','cyan','red','gray']), orient = "h", height=5, aspect=1)
```



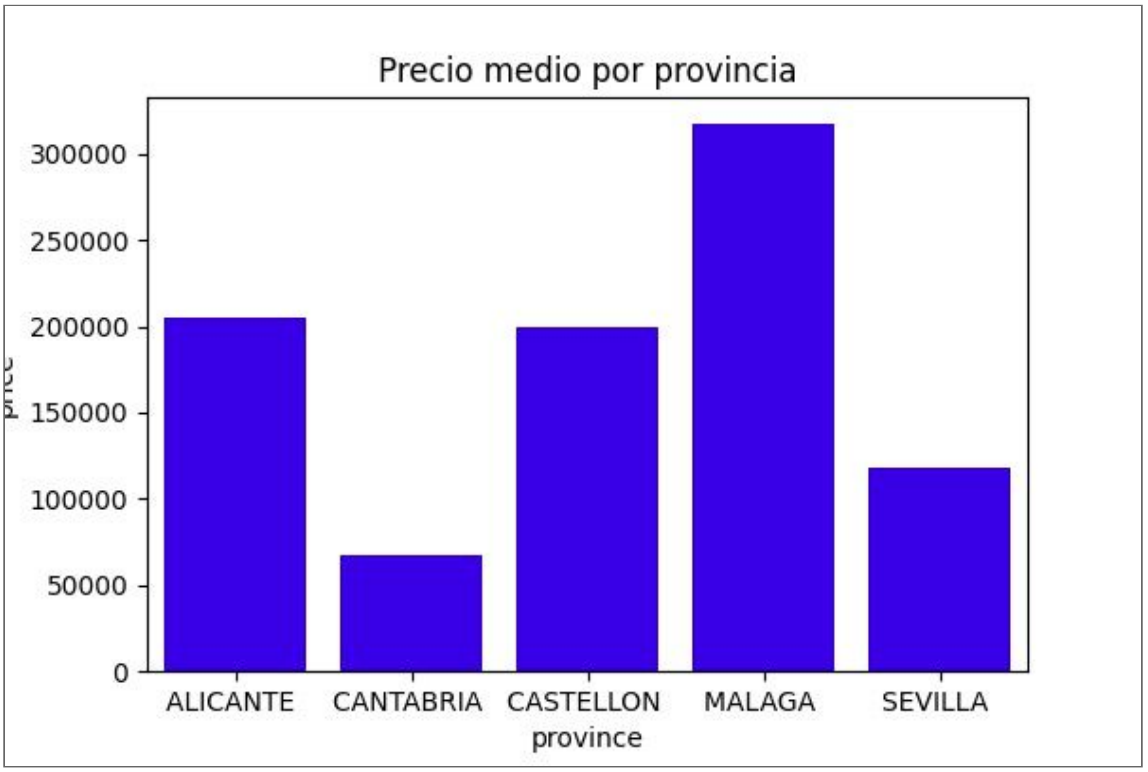
Y otro tipo de gráfico , diferenciando los colores y en columnas distintas las categorías de Clasificación Ciencias Humanas.

```
grafico= sn.catplot(y = "Clasificación Ciencias Sociales", col= "Clasificación Ciencias Humanas",kind="count", data = df_dataset,  palette=sn.color_palette(['purple', 'blue','green','cyan','red','gray']), orient = "h", height=5, aspect=0.6)
```



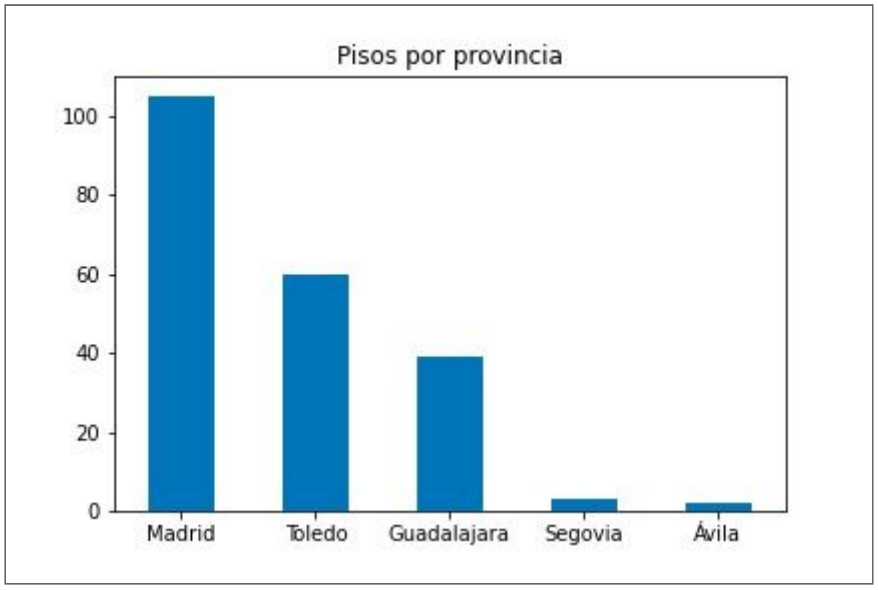
- OPCIÓN 4: TABLÓN DE ANUNCIOS MILANUNCIOS

  El gráfico muestra el precio medio por provincia en los anuncios de venta de inmuebles. Con esta opción podríamos a futuro hacer una comparativa con la opción 5 en cuanto a rangos de precios tratados en ambas webs. Si bien es cierto el subconjunto de datos debe tratar los mismos criterios para realizar esta comparativa.



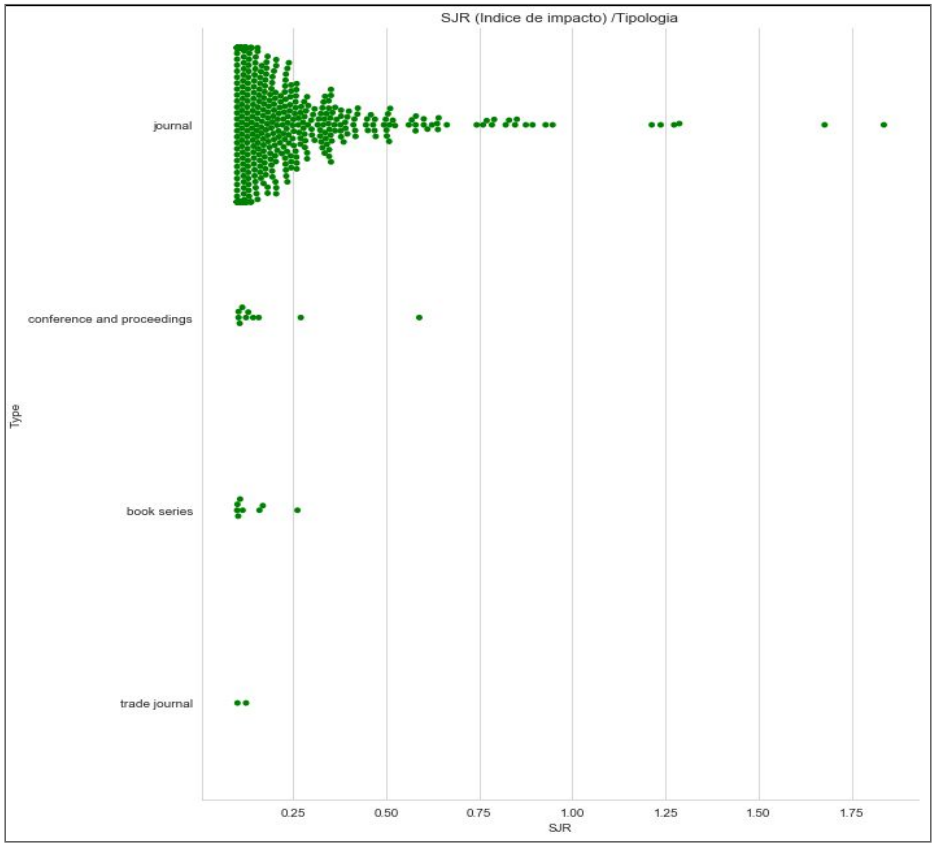
- OPCIÓN 5: PORTAL INMOBILIARIO IDEALISTA

  En el gráfico mostrado a continuación se visualiza la cantidad de inmuebles de tipo 'flat' (piso) en una radio central dentro de España agrupados por provincia.



- OPCIÓN 6 :ÍNDICE DE CLASIFICACIÓN SCIMAGO (CIENCIA/RANKINGS)

  En el siguiente gráfico se ha intentado relacionar la distribución del nº de revistas, en base del índice de impacto SJR y del tipología (Si es revista, conferencia, libros etc). Comos se observa en el gráfico , la mayor parte de las revistas/periódicos tienen un índice de impacto bajo, y sobre todo son de la categoría journal .



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

Respuesta:

- OPCIÓN 1 : AGENDA

El periodo de recogida de estos datos , se corresponden con el año 2020 y al ser el recurso que proporciona la web de datos abiertos, el nº de registros son los que se ofrecen en ese listado.
Si se lanzara cada cierto tiempo la petición , podrían ser que el dataset aumentara o los datos fuesen modificados (por ejemplo : a raíz del Coronavirus muchos eventos se están cancelando).
Se ha utilizado las librería de Python de obtener una petición a través de JSON y luego con el uso de bucles para obtener los campos , una vez que revisa la estructura del JSON. En este caso no hemos extraído todos los campos
El juego proporciona más datos a nivel de equipamiento para el evento de donde se podría extraer la longitud, latitud, y más datos del centro y se podrían por ejemplo pintar en un mapa de para ubicarlos, o subtipos, un texto con formato HTML etc..

CAMPO JSON /CABECERA	DESCRIPCIÓN
temáticas	Se refiere a los distintos tipos de Abonos o categorías que tienen establecidos en Alcobendas: El público, Abono Alternativo etc..
barrio	Indica el barrio de Alcobendas donde se llevará a cabo el evento programado
nombre	Título del Evento programado , por ej:Alcobendas en forma, deporte al aire libre
subtemas	Se refiere a una subcategorías que tienen definidas dentro de las temáticas o oferta cultural, por ejemplo si es de Servicios Deportes o Otros Salud etc.
fechaInicio	Dia inicio en el que está programado el evento.
horaInicio	Hora de inicio del evento: Indica a qué hora exacta comienza el espectáculo o evento en cuestión
fechaFin	Día fin en el que está programado el evento.
horaFin	Hora de fin del evento: Indica a qué hora exacta finaliza el espectáculo o evento en cuestión
perfiles	A qué tipo de público va dirigido Si al al público en general, Adolescentes, Infancias etc..

ficha	Un enlace a una página donde se describe el evento , con más detalle u otros campos.

- OPCIÓN 2: CATÁLOGO CURSOS COMUNIDAD DE MADRID PARA FORMACIÓN PROFESIONAL

Al igual que en el ejemplo anterior, se han extraído los datos a través del recurso JSON que porcionan , accediendo en bucle a cada uno de los registros.

El recurso no permite filtrar por fechas, por lo que son el catálogo que ofrece, aunque en el código lo hemos limitado a 100 registros de cara a un gráfico sencillo , ya que no hemos procedido con la limpieza de los datos. Cursos ofertados de la Comunidad de Madrid.

Si se lanza cada cierto tiempo esta petición , tal vez se comprobaría si cambia dicho listado.

Los campos del csv y su relación con la etiqueta del json

CABECERA	CAMPO JSON	DESCRIPCIÓN
FamiliaDesc	familia_desc	Parece que se trata de una categoría dentro de cada área para tipificar los cursos . Por ejemplo: Informática y comunicaciones dentro del Área Desarrollo
EspecialidadCertificado	especialidad_de_certificado	Indica si se obtiene certificado de especialidad o no. Como son cursos de formación profesional, será conveniente que además se pueda obtener un certificado.
CentroCP	centro_cp	Código postal del centro donde se impartirá el curso
CursoCodigo	curso_codigo	Código del curso que asigna la CM para tener un catálogo
DirigidoA	dirigido_a	A quién va dirigido Desempleados,etc
CentroMunicipio	centro_municipio	Población /Municipio del centro donde se imparte.
CursoDesc	curso_desc	Descripción/Título del curso. Ej: TÉCNICO EN SOFTWARE OFIMÁTICO
Censo	censo	Entiendo que es un nº asignado del curso a nivel de la Comunidad de Madrid

CentroTfno	centro_telefono	Nº de teléfono del centro donde se impartirá el curso
Duración	duracion_formacion	Nº de horas de duración del curso
SectorDesc	sector_desc	Sector del curso.Es una clasifcion que tienen determinada.Ej:Servicios
EspecialidadDesc	especialidad_desc_1	Descripción de la especialidad. Hay dos descripciones, más abajo se define otra.
CentroEmail	centro email	Correo electrónico del centro
Modalidad	modalidad	De qué modo se imparte el curso : si es presencial , o a distancia etc.
CentroDireccion	centro_direccion	Dirección del Centro donde se imparte el curso
SieCodigo	sie_codigo	Código SIE. Debe ser una tipificación de los cursos .
EspecialidadDesc	especialidad_desc	Descripción de la especialidad recibida
AreaDesc	area_desc	Área del curso (Desarrollo , etc.) la clasificación de los cursos
CentroDesc	centro_desc	Descripción del Centro del curso o Nombre del Centro

- OPCIÓN 3: REVISTAS CIRC

Se obtienen los datos por el nº de página y limitado a 10 paginas, con 10 resultados por página. En este caso tampoco va por fechas, por lo que para obtener la evolución de los indicadores, sería necesario que cada cierto tiempo se hiciera la consulta y se podría observar si cambian.

En este caso , se ha utilizado la librería Beautifulsoup , y se observó la página HTML , con una estructura sencilla de filas (<tr>) y celdas (<td>.

Por un lado se obtuvo la cabecera , que a la hora de trasladarlo al csv se deja solo en la primera página, y para el resto de peticiones , solo se saca la tabla de los datos.

CAMPO	DESCRIPCIÓN
Revista	Nombre/Título de la revista clasificada por

	este método CIRC
ISSN	International Standard Serial Number / Número Internacional Normalizado de Publicaciones Seriadas
Clasificación Ciencias Sociales	Se basa en asignar un letra A, A+ B,C, D , No aplica en base a una clasificación de excelencia en Ciencias Sociales según el impacto que generan esas revistas. Más detallles : https://clasificacioncirc.es/clasificacion-circ
Clasificación Ciencias Humanas	Se basa en asignar un letra A, A+ B,C, D , No aplica en base a una clasificación de excelencia en Ciencias Humanas según el impacto que generan esas revistas. Más detallles : https://clasificacioncirc.es/clasificacion-circ

NOTA: A la fecha del 01/04/2020 se ejecutaba correctamente esta opción. Cuando hemos vuelto a probar el día 03/04/2020 preparando la entrega, hemos visto que nos daba un error.Parece que ahora exigen usuario /login y no dejan los resultados en abierto. Aún así,hemos decidido dejar la opción ya que los gráficos con los nuevos colores se han estado ejecutando recientemente.

- OPCIÓN 4: MIL ANUNCIOS

En esta opción existen diversos campos dentro del juego de datos.
Como filtro se ha utilizado el área de Inmobiliaria con los filtros:

CAMPO	DESCRIPCIÓN
id operación	tipo de operación (venta, alquiler)
tipo inmueble	tipo de inmueble (piso, atico, etc)
número dormitorios	número de dormitorios

CAMPO	DESCRIPCIÓN
aditem	ítems dentro del catálogo seleccionado
adlist-paginator-summary	número de páginas devueltas de la selección realizada
aditem-detail-title	título de cada ítem dentro del catálogo seleccionado
ad-detail-title	título del ítem de detalle dentro de la página seleccionada

pagAnuPrecioTexto	precio mostrado en el ítem de detalle
m2 tag-mobile	tamaño del item seleccionado, para el caso tratado metros cuadrados de superficie del inmueble
pagAnuCatLoc	localización del inmueble seleccionado, a partir de este campo se extrae la provincia
pagAnuCuerpoAnu	descripción del ítem de detalle, literal con la descripción del inmueble
pagAnuStatsData	estadísticas del ítem de detalle (veces favorito, veces anunciado, etc)

- OPCIÓN 5:IDEALISTA

Para este práctica y mediante el acceso a través de la API se han utilizado por un lado los siguientes filtros:

CAMPO	DESCRIPCIÓN
country	país seleccionado para la búsqueda
operation	tipo de operación con el inmueble
maxitems	número de ítems por página
order	orden por característica del inmueble
center	coordenadas geográficas
distance	distancia al centro
propertyType	tipo de inmueble
sort	ordenación
numpage	numéro de página
language	idioma

Los datos se han extraído a través del objeto elementList compuesto por los siguientes campos:

CAMPO	DESCRIPCIÓN
address	dirección postal
bathroom	número baños
country	país

distance	distancia
district	distrito
exterior	si es exterior
floor	piso
hasVideo	tiene Video
latitude	latitud
longitude	longitud
municipality	municipio
neighborhood	barrio
numphotos	número fotos
operation	tipo operación
price	precio
propertyCode	código del inmueble
province	provincia
region	región
rooms	número habitaciones
showAdress	dirección
size	tamaño
subregion	subregión
thumbnail	url imagen
url	url
status	estado
newDevelopment	nuevo desarrollo
tenantGender	tipo arrendador
garageType	tipo garage
parkingSpace	parking
hasLift	ascensor
newDevelopmentFinished	terminado
isSmokingAllowed	fumadores
priceByArea	precio/area

detailedType	tipo detalle
externalReference	id referencia

- OPCIÓN 6 : SCIMAGO SJR

Relacionado con el otro de revistas, pero con más campos y también basado en datos del índice de impacto denominado SJR. En esta ocasión los datos existentes del buscador del ranking , si que permite filtrar por país, por un año concreto, y meter los parámetros de orden etc.

Para la práctica, hemos seleccionado el año 2018, y el país ESPAÑA para filtrar las revistas, se ha ordenado por el índice SJR descendente , para recuperar las revistas con mayor índice en la primera fila. En este caso, como el listado permitía hacerlo de 50 en 50 , puesto hemos iterado para obtener todos las revistas de ese ranking.

CAMPO	DESCRIPCIÓN
Number	Nº del registro en la tabla
Title	Título de la revista/conferencia etc
Type	Tipología de si es artículo/conferencia etc
SJR	SCImago Journal Rank. Según se indica en la página web, este índice expresa la media de las referencias (con un peso) recibidas en el año seleccionado por los documentos publicados en dicho periodico en los tres años anteriores Es un índice de cuánto puede impactar
H index	The h index indica el nº de artículos del periodico que ha recibido al menos h referencias/citaciones .Esto cuantifica la productividad y el impacto científico Según wikipedia : El índice h es un sistema propuesto por Jorge Hirsch de la Universidad de California para la medición de la calidad profesional de físicos y de otros científicos, en función de la cantidad de citas que han recibido sus artículos científicos.
Total Docs. (2018)	Nº de documentos de periodo seleccionado, incluidos los que se referencian/citan y los que no.
Total Docs (3 years)	Nº de documentos publicados en los tres años anteriores (los documentos del año seleccionado se excluyen). Se indican los nº de documentos totales, los que se referencias y los que no.
Total Refs. (2018)	Indica todas las referencias bibliográficas del periodico en el periodico seleccionado (en este caso 2018)

Total Cites (3years)	Nº de citaciones recibidas en el año seleccionado por un periodico/revista de los documentos publicados en los 3 años anteriores. Se consideran todos los tipos de documentos.
Citable Docs. (3years)	Nº de documentos citados publicados por una revista en los tres últimos años anteriores(Los del este año seleccionado son excluidos). Se tiene en cuenta : artículos, reviews, y papeles de la conferencias.
Cites / Doc. (2years)	Promedio de citas por documento en un período de 2 años. Se calcula considerando el número de citas recibidas por una revista en el año en curso a los documentos publicados en los dos años anteriores, es decir. citas recibidas en el año X a documentos publicados en los años X-1 y X-2.
Ref. / Doc. (2018)	Nº medio de referencias por documento en el año seleccionado

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

- OPCIÓN 1 : AGENDA ALCOBENDAS

El propietario de este juego de datos es de la web de datos abiertos de Alcobendas (https://datos.alcobendas.org/pages/informacion-datos-abiertos). Queremos expresar nuestro agradecimiento a dicha web, ya que facilita información en formato muy fácilmente explotable para poder añadir información sobre los eventos de dicho municipio y que nos sirve de aprendizaje sobre las distintas formas de presentar los datos. En este caso se trata de OPEN DATA

- OPCIÓN 2: CATÁLOGO CURSOS COMUNIDAD DE MADRID PARA FORMACIÓN PROFESIONAL

Al igual que antes, cada vez más ayuntamientos y comunidades ponen a disposición de los cuidados en formato JSON , conjuntos de datos interesantes para poder consultar o explotar. En este caso , la Comunidad de Madrid, a través de la página (https://www.comunidad.madrid/gobierno/datos-abiertos) nos presenta diversos conjuntos de datos que resultarán interesantes para distintos colectivos.

Se agradece que cada vez existan más entidades que faciliten los datos que pueden servir para difundir datos que son interesantes para diversos colectivos. En este caso, aunque solo haya sido para realizar la práctica, el uso de listado puede ser muy interesante para la gente que no tiene capacitación y desea formarse para poder conseguir un trabajo mejor

- OPCIÓN 3: REVISTAS CIRC

Los datos se recogen en la pagina (https://www.clasificacioncirc.es/).

Al igual que en las anteriores opciones, se agradece al propietario de los datos el haber podido utilizar la página como fuente de obtención de los datos recogidos ya que nos ha permitido aprender con un ejemplo de una tabla muy estructura , el uso de la librería Beautiful Soup.

Los datos que se muestran aquí están basadas en revistas extraídas de la base de datos Scopus y una clasificación de excelencia que hacen en base al índice de impacto.

- OPCIÓN 4: MIL ANUNCIOS

El propietario de la web https://www.milanuncios.com es Adevintia que apuesta por la cooperación para la reutilización de productos, inmuebles y toda clase de objetos de segunda mano, así como por ayudar a la creación de empleo a través del intercambio y contacto de las partes interesadas.

Se agradece al propietario la disponibilidad para la consulta de parte de la página web de cara a poder utilizarla en esta práctica con el único objetivo de alcanzar el aprendizaje requerido.

- OPCIÓN 5: IDEALISTA

El propiatario de estos datos es IDEALISTA , pagina www.idealista.com , a los que agradecemos la rápida contestación a la petición de la API KEY , ya que fué inmediata la respuesta y además nos proporcionaron dos documentos con toda la documentación de uso de la API KEY. El acceso a los datos nos indican además que eran de **100 peticiones al mes**, pero que si era necesario, nos podrían habilitar más accesos.

Se agradece que pongan a disposición de los estudiantes y de la docencia en general, el acceso para poder probar el funcionamiento de manera gratuita con el fin de aprender el uso de las APIs.

- OPCIÓN 6 : SCIMAGO SJR

Los datos se recogen en la página https://www.scimagojr.com/journalrank.php.

SCImago es un grupo de investigación del CSIC , Universidad de Granada , Extremadura Carlos III (Madrid) y Alcalá de Henares dedicados a la investigación y análisis basados en técnicas visuales. El ranking realizado en dicha página está obtenida de la base de datos ae Scopus® (Elsevier B.V.).

Por tanto , se agradece tanto a SCIMago como a la base de datos original el uso de estos datos para la realización de la práctica.

También queremos agradecer , aunque finalmente no lo hayamos utilizado, a AEMET que también facilita su API KEY de manera gratuita y resulta interesante los datos, para su explotación ya que proporciona acceso a los pronósticos de tiempo y recogida de datos de sus estaciones meteorológicas y sensores.

Las referencias están al final, en el apartado REFERENCIAS.

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

OPCIÓN 1 : AGENDA DE DATOS ALCOBENDAS

Este conjunto de datos puede ser interesante para los ciudadanos de Madrid que estén interesados en asistir a algún evento cultural. Como se indicó en el primer apartado , podría servir para que se integrase en apps de información al ciudadano o portales de los centros culturales.

A nivel analítico, es posible que pudiera servir para hacer comparativas del nº de cursos que se ofertan en los distintos niveles y si se debería mejorar en contenido para cubrir a toda la población o barrios.

Preguntas a las que responde : ¿ Existe algún evento cultural hoy?. ¿ A qué público se dirige la oferta cultural de Alcobendas ?¿ Que barrio tiene mayor oferta cultural?

OPCIÓN 2: CATÁLOGO CURSOS COMUNIDAD DE MADRID PARA FORMACIÓN PROFESIONAL

Este conjunto de datos puede ser interesante para las personas que quieran buscar un curso de capacitación profesional y también integrarlo en distintos portales de cursos para informar.

A nivel analítio , nos puede proporcionar información del volumen de cursos existente, cuáles son las áreas en la que se ofertan curso de formación profesional, que duración media tienen los cursos de determinada área, la clasificación de los cursos, que centros ofertan más cursos et. y así se pueden establecer subvenciones a otros centros, para que mejoren por ejemplo la oferta formativa , etc..

Preguntas a las que responde : ¿ Que centros ofertan cursos de Desarrollo?. ¿Que municipios de la Comunidad de Madrid ofertan cursos de una determinada área?¿ Qué duración media tienen los cursos de formación profesional?. ¿ Existen distintas modalidades de cursos? etc.

OPCIÓN 3: REVISTAS CIRC

Este juego de datos tiene interés a nivel científico de cara a como se les clasifica y se puede considerar una revista de renombre / excelencia. Como en la página web indica, se utiliza luego en biblioteconomía

Preguntas a las que puede responder , ¿ Que clasificación obtiene a nivel de Ciencias Sociales determinada revista?. ¿ Y en Ciencias Humanas? .

OPCIÓN 4: MIL ANUNCIOS

El conjunto de datos aportado por esta web permite una variabilidad suficiente para realizar filtros sobre las opciones de búsqueda bajo distintos conceptos.

Contiene datos de diferentes tipos con los que se podrían realizar diferentes analísticas.

La temática de anuncios para extraer información es muy amplia y los criterios dentro de cada temática permiten realizar comparativas.

Podría dar respuesta a preguntas como: ¿Cuál es el tipo de producto (ej coche) que más se anuncia para venta? ¿Rango de precios en una temática concreta (ej precio máximo y mínimo de servicios y a que está asociado?

OPCIÓN 5: IDEALISTA

El conjunto de datos al que se puede acceder a través de la API permite un detalle muy minucioso de las características de los inmuebles que recoge el portal, de tal forma que se puedan realizar filtros y tratamientos específicos para extraer la información de interés o más relevante.

Con ello se permite dar respuesta a consultas que abarquen diferentes puntos de análisis como: ¿tipo de inmueble más común por área? ¿para un tipo de inmueble rango de precios por zonas? ¿relación entre precio y características (calefacción, baños, habitaciones) del inmueble?

Además se podría realizar comparativa del tipo de inmuebles y precios asociados anunciados en una zona concreta en esta opción y la anterior.

OPCIÓN 6 : SCIMAGO SJR

Al igual que el punto 3, este ranking que se obtiene de las revistas a nivel científico de las distintas áreas tiene un interés sobre todo en el ámbito científico. De cara al prestigio / calidad profesional de científicos, en función de la cantidad de citas que han recibido sus artículos. Se usa como índice bibliométrico.

Por tanto este dataset responde a las siguientes preguntas : ¿ Qué posición en el ranking tiene esta revista/artículo?.¿Cuál es la media de las citas de este tipo de revistas en los 3 últimos años?.¿ Cuáles son las revistas mejor posicionadas en l a primer cuartil?. etc.

A nivel analitico, en la propia pagína de SCIMAGO se representa como varía el índice SJR , y permite ver distinta visualizaciones.

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**

- OPCIÓN 1 : AGENDA ALCOBENDAS

  Database released under Open Database License, individual contents under Database Contents License

  Al hacer uso de esta licencia, estamos seguros de que son datos abiertos y que no hay problemas de uso siempre que no se manipulen

- OPCIÓN 2: CATÁLOGO CURSOS COMUNIDAD DE MADRID PARA FORMACIÓN PROFESIONAL

  Creative Commons Attribution Ultimo actualización de los dats Marzo 12, 2019

  Son datos abiertos , estamos seguros de que son datos abiertos y que no hay problemas de uso siempre que no se manipulen al igual que el anterior caso.

- OPCIÓN 3: REVISTAS CIRC

  Unknown License

  En este caso, no he encontrado la opción de el uso de esta web. No tenemos intención de manipulación de estos datos, pero es cierto , que se ha utilizado con el fin de aprendizaje.

- OPCIÓN 4: TABLON ANUNCIOS MIL ANUNCIOS

  Unknown License

  Al igual que la opción anterior no se ha encontrado licencia para el uso de esta web.

- OPCIÓN 5: IDEALISTA

  Unknown License

  En este caso contactamos a través del apartado de labs de la pagina de www.idealista.es y se solicitó el acceso con la descripción del uso (práctica del máster con fines educativos) que íbamos a dar . En el correo de respuesta donde nos proporcionaron el api key y la documentación asociada nos indicaron que teníamos 100 peticiones al mes y que si necesitábamos más. Con las pruebas realizadas las hemos consumido en este tiempo y vamos a solicitar que nos lo amplíen de cara a la entrega final.

- OPCIÓN 6 : SCIMAGO SJR

  Unknown License

  Idem que en la opción 3 .

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

Respuesta:

Enlace a GITHUb .

Es repositorio público.https://github.com/bmartineza01/projects/

El programa es pra1_webscraping_entregaTotal.py (ENTREGA TOTAL)

CSVs : pra1_datasetX.csv donde X es la opción que se seleccione. En Github hemos subido una ejecución realizada de cada opción con el fin de verificar el correcto funcionamiento del programa integrado.

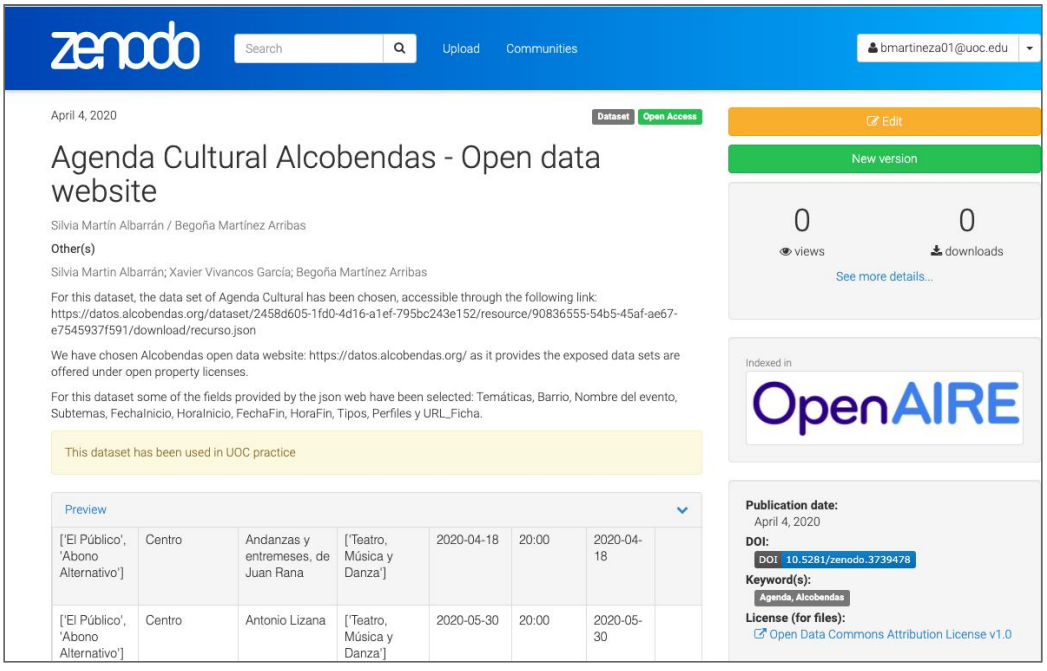
Gráficos : pra1_dataset.jpeg donde X es la opción seleccionada con el gráfico que da como resultado de ejecutar la opción elegida en el menú

10. Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

Respuesta: Se publica finalmente el de datos abiertos de la agenda de Alcobendas.

El dataset publicado es pra1_dataset_agenda.csv que se corresponde con el juego de datos extraído con la primera opción del menú llamado pra1_dataset1.csv.

El repositorio publicado en Zenodo es el siguiente:



11. Entrega. Presentar el trabajo con el DOI del dataset en Github.

Respuesta: Aunque la entrega parcial se hizo en el repositorio de smartinalbar, por motivos de ocultar las claves de apikey de idealista (a pesar de borrar/se veían todos los commits anteriores) , hemos cambiado al repositorio siguiente:
Enlace a GITHUb https://github.com/bmartineza01/projects/

Es repositorio público.
README.ME

Descripción de los ficheros:

El programa es pra1_webscraping_entregaTotal.py (ENTREGA TOTAL)

CSVs : pra1_datasetX.csv donde X es la opción que se seleccione. En Github hemos subido una ejecución realizada de cada opción con el fin de verificar el correcto funcionamiento del programa integrado.

Gráficos : pra1_dataset.jpeg donde X es la opción seleccionada con el gráfico que da como resultado de la ejecución de esa opción.

Referencias.

1.-Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

2.Título:[¿Que es el ISSN?]. Consultado : [28/03/2020].Disponible en: http://www.bne.es/es/LaBNE/CentroEspanolISSN/QueEsElISSN

3 Título[Clasificación CIRC]. Consultado :[28/03/2020]. Disponible en: https://clasificacioncirc.es/clasificacion-circ

4 Título[Rank Journals. Journal Indicators]. Consultado :[28/03/2020]. Disponible en: https://www.scimagojr.com/help.php#rank_journals

5 Título[Agenda e Información de Datos Abiertos]. Consultado :[22/03/2020]. Disponible en: https://datos.alcobendas.org/ y https://datos.alcobendas.org/pages/informacion-datos-abiertos

6 Título[Datos Abiertos]. Consultado :[23/03/2020]. Disponible en: https://www.comunidad.madrid/gobierno/datos-abiertos

7 Título[Idealista -Acceso API Key]. Consultado :[22/03/2020]. Disponible en: http://developers.idealista.com/access-request y https://www.idealista.com/labs/

8 Título[seaborn.catplot]. Consultado :[25/03/2020]. Disponible en: https://seaborn.pydata.org/generated/seaborn.catplot.html

9 Título: [OAuth 2.0].Consultado[23/03/2020]Disponible: https://developer.byu.edu/docs/consume-api/use-api/oauth-20/oauth-20-python-sample-code

10.Titulo:[How to Save a Plot to a File Using Matplotlib] Consultado[27/03/2020] Disponible en : https://chartio.com/resources/tutorials/how-to-save-a-plot-to-a-file-using-matplotlib/

11 Título[Beautiful Soup Documentation] Consultado[24/03/2020] Disponible en: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

12 Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- **Los apartados 1, 2, 3 y 4 valen 0,25 puntos cada uno.**
- **Los apartados 5, 6, 7, 8 valen 1 punto cada uno.**
- **Los apartados 9 y 10 valen 2,5 puntos cada uno.**

Otros criterios que se tomarán en cuenta para la evaluación son:

- **Idoneidad de las respuestas (deberán ser claras y completas).**
- **Complejidad del sitio web elegido para la extracción.**
- **Síntesis y claridad, a través del uso de comentarios, del código resultante.**
- **Presentación adecuada de los datos.**
- **Organización y claridad del documento de entrega final.**

- **Completitud de los documentos requeridos para la entrega final.**

Formato y Fecha de entrega

Durante la semana del 30 de marzo, el grupo podrá entregar al profesor una entrega parcial opcional. Esta entrega parcial es muy recomendable para recibir asesoramiento sobre la práctica y verificar que la dirección tomada es la correcta. Se entregarán comentarios a los estudiantes que hayan efectuado la entrega parcial pero no contará para la nota de la práctica. En la entrega parcial los estudiantes deberán entregar por correo electrónico (xvivancos@uoc.edu) el enlace al repositorio Github con lo que hayan avanzado.

En referente a la entrega final, hay que entregar un único fichero que contenga el enlace a Github donde haya:
1. Una Wiki donde estén los nombres de los componentes del grupo y una descripción de los ficheros.
2. Un documento PDF con las respuestas a las preguntas y los nombres de los componentes del grupo. Además, al final del documento, debe aparecer la siguiente tabla de contribuciones al trabajo, la cual debe firmar cada integrante del grupo con sus iniciales. Las iniciales representan la confirmación por parte del grupo que el integrante ha participado en dicho apartado. Todos los integrantes deben participar en cada apartado, por lo que, idealmente, los apartados deberían estar firmados por todos los integrantes.

Contribuciones	Firma
Investigación previa	Integrante 1, Integrante 2, ...
Redacción de las respuestas	Integrante 1, Integrante 2, ...
Desarrollo código	Integrante 1, Integrante 2, ...

3. Una carpeta con el código Python o R generado para obtener los datos.
4. El fichero CSV con los datos.

Este documento de la entrega final se tiene que entregar en el espacio de Entrega y Registro de AC del aula antes de las 23:59 del día 14 de abril. No se aceptarán entregas fuera de plazo.

Contribuciones y Firmas:

CONTRIBUCIONES	FIRMAS
Investigación previa	SMA/MBMA

Redacción Respuesta 1	SMA/MBMA
Redacción Respuesta 2	SMA/MBMA
Redacción Respuesta 3	SMA/MBMA
Redacción Respuesta 4	SMA/MBMA
Redacción Respuesta 5	SMA/MBMA
Redacción Respuesta 6	SMA/MBMA
Redacción Respuesta 7	SMA/MBMA
Redacción Respuesta 8	SMA/MBMA
Redacción Respuesta 9	SMA/MBMA
Redacción Respuesta 10	SMA/MBMA
Redacción Respuesta 11	SMA/MBMA
Desarrollo Código	SMA/MBMA