
WEB SCRAPING

Tipologia i cicle de vida de les dades

Pràctica 1

Miquel Piulats
Miquel Muntaner

Índex

Context	2
Títol	2
Descripció del dataset	3
Representació gràfica	3
Contingut	7
Propietari	8
Inspiració	8
Llicència	9
Codi	9
Dataset	10
Vídeo	10

Context

Aquesta pràctica ha estat realitzada pels alumnes Miquel Piulats i Miquel Muntaner per a l'assignatura **"Tipologia i cicle de les dades"** del màster universitari en Ciència de Dades de la UOC.

Per aconseguir una tècnica adequada de *web scraping*, s'ha proposat fer una pràctica on aplicar els coneixements adquirits en l'assignatura. Per a poder fer el treball, s'ha hagut de pensar conjuntament quins eren els objectius per a seleccionar una pàgina web on es pogués desenvolupar la pràctica sense problemes amb el propietari i, alhora, que fos una motivació per a realitzar-la.

Per això s'ha seleccionat la pàgina web de transparència de la *Comunidad de Madrid*, concretament l'apartat de contractes públics. La url és: <https://www.comunidad.madrid/transparencia/informacion-contratos-publicos>

Aquesta pàgina web proporciona, de manera transparent, els contractes que s'han firmat durant els últims anys amb diverses empreses privades per cobrir necessitats públiques de la *Comunidad de Madrid*. La pràctica està centrada en els contractes adjudicats sense procediments de publicitat perquè no era possible extreure tots els registres en un temps raonable.

Títol

El títol ens indica amb claredat el tipus de dataset que ens trobarem: *Contractes públics adjudicats sense procediments de publicitat dins la Comunidad de Madrid*. Amb aquest títol no es pretén ser creatius o artístics, només pretén que sigui concís.

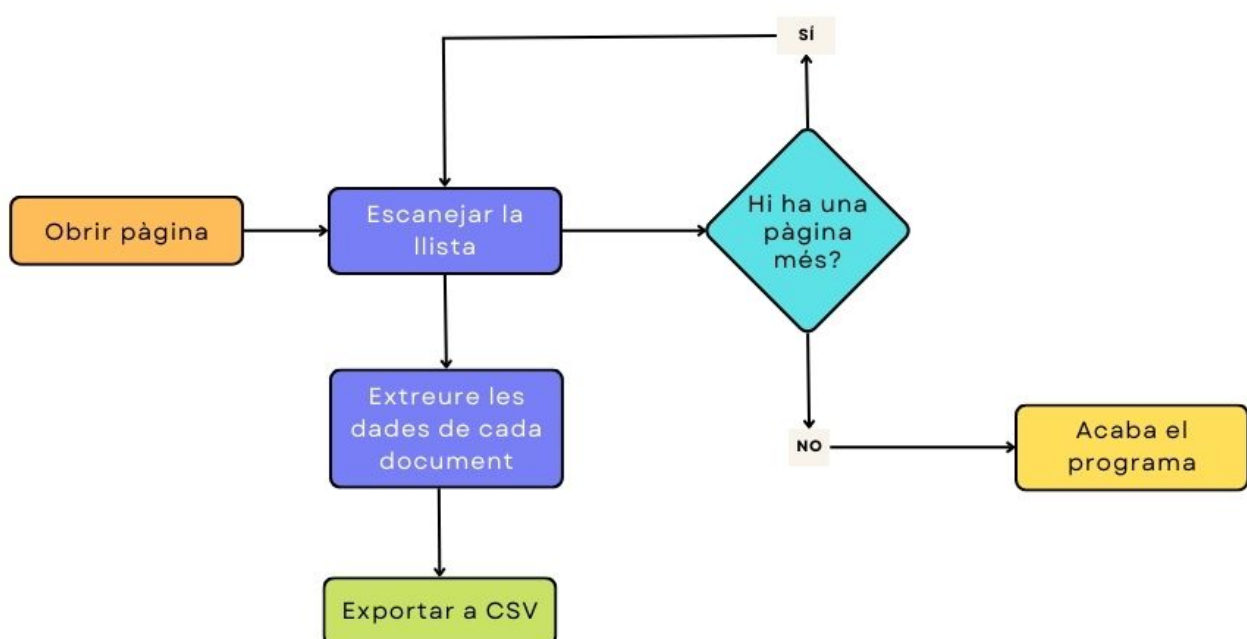
Descripció del dataset

El dataset aconseguit amb el procediment de *web scraping* ens proporciona informació útil sobre les característiques de cada un dels contractes publicats al portal de transparència de la *Comunidad de Madrid*. S'ha filtrat per contractes sense procediments de publicitat, perquè es considera que aquests són uns dels més parcials, però també perquè fer un anàlisi complet de tots els contractes publicats a la pàgina web, hagués portat a fer un script que hagués tardat molts dies en finalitzar. Això, també hagués pogut dur als propietaris o administradors de la pàgina web a sospitar d'un mal ús de la seva pàgina, o que algú té intencions malicioses. Per aquest motiu, per a fer la pràctica, es va optar per fer un anàlisi un poc més simple, però alhora, útil.

Amb la informació que es proporcionarà a través del dataset creat, es podran fer diverses comparacions amb diversos contractes i també es podran agrupar i sumar algunes quanties pressupostàries adjudicades a una mateixa empresa (entre d'altres).

Representació gràfica

Com es pot veure en el diagrama de fluxe, el primer que es farà una vegada s'iniciï l'script,



serà obrir la url que li haurem indicat, en el cas de la pràctica serà la url:

[http://www.madrid.org/cs/Satellite?](http://www.madrid.org/cs/Satellite?c=Page&cid=1224915242285&codigo=PCON_&idPagina=1224915242285&language=es&newPagina=1&numPagListado=5&pagename=PortalContratacion%2FComunes%2FPresentacion%2FPCON_resultadoBuscadorAvanzado&paginaActual=2&paginasTotal=1204&rootelement=PortalContratacion%2FComunes%2FPresentacion%2FPCON_resultadoBuscadorAvanzado&site=PortalContratacion&tipoPublicacion=Contratos+adjudicados+por+procedimientos+sin+publicidad)

[c=Page&cid=1224915242285&codigo=PCON_&idPagina=1224915242285&language=es&newPagina=1&numPagListado=5&pagename=PortalContratacion%2FComunes%2FPresentacion%2FPCON_resultadoBuscadorAvanzado&paginaActual=2&paginasTotal=1204&rootelement=PortalContratacion%2FComunes%2FPresentacion%2FPCON_resultadoBuscadorAvanzado&site=PortalContratacion&tipoPublicacion=Contratos+adjudicados+por+procedimientos+sin+publicidad](http://www.madrid.org/cs/Satellite?c=Page&cid=1224915242285&codigo=PCON_&idPagina=1224915242285&language=es&newPagina=1&numPagListado=5&pagename=PortalContratacion%2FComunes%2FPresentacion%2FPCON_resultadoBuscadorAvanzado&paginaActual=2&paginasTotal=1204&rootelement=PortalContratacion%2FComunes%2FPresentacion%2FPCON_resultadoBuscadorAvanzado&site=PortalContratacion&tipoPublicacion=Contratos+adjudicados+por+procedimientos+sin+publicidad)

Dins aquesta pàgina
escanejarà la llista que hi ha
dins els continguts:
"cajaBlanca", que és on
trobarem els enllaços als
contractes, i els anirem obrint
un per un per a poder
analitzar i extreure els
diferents camps que ens
interessen per a formar el
dataset.

Portal de la Contratación Pública de la Comunidad de Madrid

Buscador

0 1 2 ATENCIÓN AL CIUDADANO Comunidad de Madrid

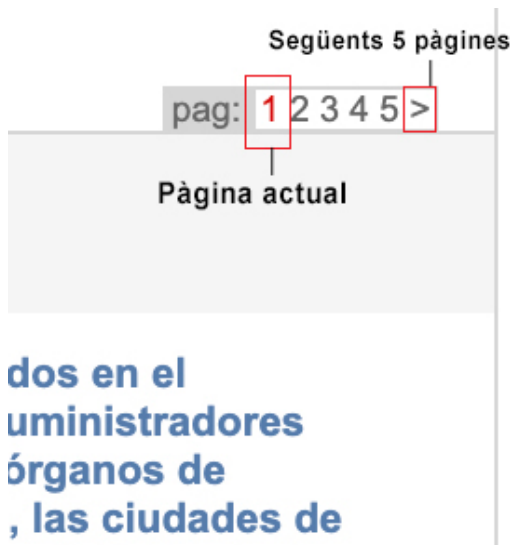
PERFIL DE CONTRATANTE SERVICIOS Y CONSULTAS INFORMACIÓN GENERAL REGISTROS DE LICITADORES OFICINA VIRTUAL ACTUALIDAD

Número de resultados encontrados: 12064. Exportar resultados a Excel

Volver a buscar>







pag: 1 2 3 4 5 >

Situación	Título y descripción
En plazo	Procedimiento de adjudicación de los contratos basados en el Acuerdo Marco 202001AM0004 para la selección de suministradores de vacunas de calendario y otras para determinados órganos de contratación de la Administración General del Estado, las ciudades de Ceuta y Melilla y varias comunidades autónomas (lotes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18 y 20) para los años 2023 y 2024 en la Comunidad de Madrid. - Ref: 4866497 Contratos públicos
Plazo indefinido	Suministro de analizadores portátiles y cartuchos exclus GAAP - Ref: 4866511 Contratos públicos
Plazo indefinido	Contrato de exhibición Dans la mesure de l'impossible - Ref: 4846000 Contratos públicos
Plazo indefinido	Suministro de medicamentos Rezolsta comp 800/150 mg para el Hospital Universitario Príncipe de Asturias - Ref: 4846754 Contratos públicos
Plazo indefinido	Servicio de un sistema de evaluación miocárdica por proceso con resonancia magnética a utilizar por el Servicio de Radiología del Hospital Universitario Ramón y Cajal - Ref: 4866558 Contratos públicos



Una vegada s'hagin identificat i obert cada un dels enllaços de la pàgina actual, es comprovarà si existeix una pàgina posterior i, si és així, s'avançarà pàgina i es tornarà a escanejar tot el llistat dels contractes. En el cas de la pàgina on s'ha fet el *web scraping*, ens trobem amb un símbol que ens passarà a les següents 5 pàgines, i tornarà a començar de nou a fer el cicle amb les pàgines actuals. Una vegada s'arriba a la darrera pàgina, es donaria el programa per finalitzat.

Es pot veure com dins cada contracte hi ha diferents camps amb informació. Aquesta serà la que utilitzarem per a crear el nostre dataset.

Procedimiento de adjudicación de los contratos basados en el Acuerdo Marco 202001AM0004 para la selección de suministradores de vacunas de calendario y otras para determinados órganos de contratación de la Administración General del Estado, las ciudades de Ceuta y Melilla y varias comunidades autónomas (lotes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18 y 20) para los años 2023 y 2024 en la Comunidad de Madrid.	Gestión
En plazo	Actas de las mesas, licitadores y documentación complementaria (2) Acceso
Estado de la licitación Abierto	Pilegos de condiciones (2) Acceso
Tipo Publicación Contratos adjudicados por procedimientos sin publicidad	Licitación electrónica: normativa, información, manual de la aplicación, acceso.
Objeto del contrato Procedimiento de adjudicación de los contratos basados en el Acuerdo Marco 202001AM0004 para la selección de suministradores de vacunas de calendario y otras para determinados órganos de contratación de la Administración General del Estado, las ciudades de Ceuta y Melilla y varias comunidades autónomas (lotes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18 y 20) para los años 2023 y 2024 en la Comunidad de Madrid. Es objeto la adquisición de las dosis necesarias para vacunar de forma rutinaria a todas las personas incluidas en los grupos de población definidos en los respectivos calendarios de vacunación de la Comunidad de Madrid, así como a aquellos colectivos de personas incluidas en ciertos grupos de riesgo.	Preparación Contrato (3) Acceso
Código CPV 33651600-4	Información relacionada
Número de expediente 32LT-2022 (202001AM0004)	 Servicio de alertas >
Referencia 4866497	Si necesitas ayuda...
Tipo de contrato Suministros	 Chatéanos >
Entidad adjudicadora Consejería de Sanidad	 Twitéanos >
Código NUTS ES30	 Llámamos >
Procedimiento Adjudicación Derivado de acuerdo marco	 Escribenos >
Valor estimado sin I.V.A 78.377.156 euros	 Visítanos >
Presupuesto base licitación (sin impuestos) 78.377.156 euros	
Presupuesto base licitación. Importe total 81.512.242,24 euros	
Duración del contrato 2 Años	

El dataset resultant quedaria com el de la imatge inferior. Apareixen alguns camps buits perquè no tots els contractes tenien exactament els mateixos atributs, i perquè alguns no tenien tots els camps emplenats. A la imatge només apareixen alguns dels camps, però s'expliquen al següent apartat.

Adjudicación del contrato publicada el	Compra pública innovadora	Código CPV	Duración del contrato	Entidad adjudicadora	Estado de la licitación	Formalización del contrato publicada el	Formalización del contrato publicada en BOCM el
29 abril 2022	No	33600000-6	24 Meses	Consejería de Sanidad	Abierto	10 junio 2022	22 junio 2022
29 abril 2022	No	33600000-6	24 Meses	Consejería de Sanidad	Abierto	10 junio 2022	22 junio 2022
09 marzo 2022		33600000-6	24 Meses	Consejería de Sanidad	Abierto	07 abril 2022	21 abril 2022
09 marzo 2022		33600000-6	24 Meses	Consejería de Sanidad	Abierto	07 abril 2022	21 abril 2022
23 enero 2022	No	33690000-3	24 Meses	Consejería de Sanidad	Abierto	28 febrero 2022	15 marzo 2022
23 enero 2022	No	33690000-3	24 Meses	Consejería de Sanidad	Abierto	28 febrero 2022	15 marzo 2022
26 octubre 2021	No	33690000-3	12 Meses	Consejería de Sanidad	Abierto	13 diciembre 2021	03 enero 2022
07 octubre 2021		33600000-6	12 Meses	Consejería de Sanidad	Abierto		
		38970000-5	6 Meses	Fundaciones	Abierto		
06 marzo 2022		33600000-6	24 Meses	Consejería de Sanidad	Abierto	04 abril 2022	19 abril 2022

Contingut

Els camps obtinguts mitjançant l'scraping, es fiquen tots dins un dataset nou. La informació que s'ha volgut aportar a aquest dataset, és informació que s'ha considerat interessant per al posterior anàlisi (que encara no s'ha realitzat).

S'ha de tenir en compte també que la base de dades del portal on s'ha extret tota la informació, va actualitzant periòdicament els contractes adjudicats. Així que l'anàlisi que es fes de les dades, hauria de ser amb la informació més actualitzada possible. També podria ser que les polítiques de transparència canviessin en un futur i aquests documents desapareguessin del portal.

Els camps que s'han exportat al dataset final són:

- ❖ **L'estat de la licitació:** Si està encara obert, ja s'ha adjudicat, si ja s'ha resolt o si s'ha hagut de prorrogar el contracte.
- ❖ La **duració del contracte** en dies, mesos o anys.
- ❖ La **data** quan es va **adjudicar** el contracte.
- ❖ La **data formalització** del contracte.
- ❖ L'**entitat adjudicadora** del contracte: Consejería de Cultura, Turismo, deporte, vicepresidencia...
- ❖ **Pressupost base** de la licitació
- ❖ **Valor estimat** del servei contractat.

El període de temps d'aquest dataset està comprès entre el gener del 2011 i l'octubre del 2022

Propietari

El propietari del conjunt de dades és la Comunidad de Madrid qui comparteix, a través del seu portal de transparència, tota la informació referent als contractes públics de Madrid.

Per actuar d'acord amb els principis ètics i legals en el contexte del projecte, s'ha treballat sobre una pàgina web on no s'especificava cap inconveniència al document `robot.txt`, i també s'ha tingut en compte que els documents fossin de caràcter públic.

Inspiració

Ens agrada la transparència i alguns temes d'àmbit polític. Per això vam pensar que seria ideal poder fer una pràctica que ens motivés a trobar informació oculta o informació compromesa per a poder analitzar. La pàgina web de transparència de la *Comunidad de Madrid*, ofereix als seus visitants un document *Excel* amb la informació dels contractes. Però quan ens vam descarregar el document, vam poder comprovar que la informació que es donava estava incompleta i desordenada. Això ens va motivar encara més a poder fer un dataset més net, complet i organitzat que els que ens ofereixen.

Vam trobar un dataset i un anàlisi similar ja fet sobre el govern dels Estats Units, que es va fer per a saber on anaven els diners dels impostos dels ciutadans. Aquest anàlisi està publicat a la pàgina web Kaggle i es pot trobar al següent enllaç: <https://www.kaggle.com/code/skeller/an-introduction-to-government-contract-spend-data>

Es pot veure certa similitud en els camps que contenen els dos datasets, i es podrien fer anàlisis de certa semblansa, però la *Comunidad de Madrid* és més propera a nosaltres i ens sembla més interessant política i econòmicament.

Llicència

La llicència que vam seleccionar pel nostre dataset és una llicència de *Creative Commons*. Està posada a la descripció dins el repositori: https://github.com/lequims/cmadrid_contracts i també al document *Readme.md* dins el projecte. La llicència és la següent: **Reconeixement-NoComercial 4.0 Internacional de Creative Commons** [(<https://i.creativecommons.org/l/by-nc/4.0/80x15.png>)](<http://creativecommons.org/licenses/by-nc/4.0/>)

Vam triar aquesta llicència perquè pensem que la informació que s'ha obtingut ha de ser de caràcter públic i obert a que tothom la pugui utilitzar. Però sí que volem que se'ns reconegui la feina feta, i no volem que ningú en tregui profit econòmic de la nostra feina, ja que perdria l'essència del perquè ho hem fet.

Codi

L'script utilitzat per a fer el *web scraping* s'ha programat amb *Python 3*. La llibreria que s'ha utilitzat per a fer la feina és la de *scrapy*. Aquesta llibreria ens ha permès, de manera relativament sencilla, poder navegar a través d'un html per a poder identificar les etiquetes que ens interessaven per a obtenir les dades.

També s'han modificat alguns ajustos de la llibreria per a que la navegació a través de la pàgina fos el més semblant a la d'un humà. Per això el primer que es va fer va ser canviar l'user agent per defecte de la llibreria, i vam utilitzar-ne un del navegador Mozilla Firefox. També es van modificar les consultes simultànies, que per defecte eren 16, i en vam posar 3. Això fa que tot el procés s'allargui, però es satura menys el servidor i no es molesta tant al propietari. Finalment també vam posar un petit retard en el temps de les consultes, per acabar-ho d'humanitzar, ja que una persona mai farà les consultes tan seguides com una màquina.

Dataset

El dataset obtingut ha estat publicat a zenodo.org i el doi obtingut és el següent: <https://doi.org/10.5281/zenodo.7332098>

Vídeo

En el vídeo mostrem els punts més importants d'aquesta pràctica. L'enllaç al vídeo és: <https://drive.google.com/file/d/1TWpR-4pRnWclsdUUNvBK5wjo8-W5Lyd/view?usp=sharing>

Contribucions	Signatura
Investigació prèvia	Miquel Piulats, Miquel Muntaner
Redacció de les respostes	Miquel Piulats, Miquel Muntaner
Desenvolupament del codi	Miquel Piulats, Miquel Muntaner
Participació del vídeo	Miquel Piulats, Miquel Muntaner