Alumnos: **Diego Martín Montoro**
Javier Hernández Hernández
PRAC1

PRÁCTICA 1

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

   Amazon es una plataforma web donde una serie de vendedores anuncian productos de diversa índole. Uno de los departamentos más extensos de dicha plataforma es el de Electrónica, dentro del cual son gestionados datos de miles de dispositivos electrónicos entre los que podemos encontrar teléfonos móviles de distintas marcas y tecnologías en un rango de precios muy amplio.

   Para el objeto de nuestro trabajo, consistente en la obtención de una amplia base de datos con referencias de móviles, y sus precios de venta, hemos considerado que Amazon nos permite la extracción de gran cantidad de información.

   Se han analizado los permisos de la web para su rastreo, en el fichero **robots.txt**.

```
User-agent: *
Disallow: */s?k=*&rh=n*p_*p_*p_
Disallow: /dp/product-availability/
Disallow: /dp/rate-this-item/
Disallow: /exec/obidos/account-access-login
Disallow: /exec/obidos/change-style
Disallow: /exec/obidos/dt/assoc/handle-buy-box
Disallow: /exec/obidos/flex-sign-in
Disallow: /exec/obidos/handle-buy-box
Disallow: /exec/obidos/refer-a-friend-login
Disallow: /exec/obidos/subst/associates/join
Disallow: /exec/obidos/subst/marketplace/sell-your-collection.html
Disallow: /exec/obidos/subst/marketplace/sell-your-stuff.html
Disallow: /exec/obidos/subst/partners/friends/access.html
Disallow: /exec/obidos/tg/cm/member/
Disallow: /gp/cart
Disallow: /gp/content-form
Disallow: /gp/customer-images
Disallow: /gp/customer-media/upload
Disallow: /gp/customer-reviews/common/du
Disallow: /gp/customer-reviews/write-a-review.html
Disallow: /gp/flex
Disallow: /gp/gfix
Disallow: /gp/history
Disallow: /gp/item-dispatch
Disallow: /gp/legacy-handle-buy-box.html
Disallow: /gp/reader
Disallow: /gp/registry/wishlist/*/reserve
Disallow: /gp/richpub/listmania/createpipeline
Disallow: /gp/music/clipserve
Disallow: /gp/recsradio
Disallow: /gp/sign-in
Disallow: /gp/slides/make-money
Disallow: /gp/structured-ratings/actions/get-experience.html
Disallow: /gp/twitter/
Disallow: /gp/vote
Disallow: /gp/voting/
Disallow: /gp/yourstore
Disallow: /ap/signin
Disallow: /gp/registry/search.html
```

```
Disallow: /gp/orc/rml/
Disallow: /gp/dmusic/mp3/player
Disallow: /gp/entity-alert/external
Disallow: /gp/customer-reviews/dynamic/sims-box
Disallow: /review/dynamic/sims-box
Disallow: /gp/redirect.html
Disallow: /gp/customer-media/actions/delete/
Disallow: /gp/customer-media/actions/edit-caption/
Disallow: /gp/dmusic/
Allow: /gp/dmusic/promotions/%m%zMusicUnlimited
Disallow: /gp/customer-media/product-gallery/B007HCCOD0
Disallow: /gp/help/customer/display.html?*nodeId=200534000
Disallow: /gp/feature.html?*docId=1000632623
Disallow: /gp/aag
Disallow: /gp/socialmedia/giveaways
Disallow: /gp/aw/so.html
Disallow: /gp/pdp/profile/
Disallow: /gp/product/product-availability
Disallow: /gp/offer-listing
Disallow: /dp/twister-update/
Disallow: /dp/e-mail-friend/
Disallow: /gp/registry/wishlist/
Disallow: /wishlist/
Allow: /wishlist/universal
Allow: /wishlist/vendor-button
Allow: /wishlist/get-button
Disallow: /gp/wishlist/
Allow: /gp/wishlist/universal
Allow: /gp/wishlist/vendor-button
Allow: /gp/wishlist/ipad-install
Disallow: /registry/wishlist/
Disallow: /local/ajax/
Disallow: /gp/rentallist
Disallow: /gp/video/dvd-rental/settings
Disallow: /gp/rl/settings
Disallow: /gp/video/settings
Disallow: /gp/video/watchlist
Disallow: /gp/video/library
Disallow: /gp/profile/
Disallow: /reviews/iframe
Disallow: /gp/ask-widget/askWidget*
Disallow: /ss/customer-reviews/lighthouse/
Disallow: /gp/aw/ol/
Disallow: /gp/promotion/
Disallow: /hz/leaderboard/top-reviewers/
Disallow: /hz/leaderboard/hall-of-fame/
Disallow: /review/top-reviewers/
Disallow: /review/top-reviewers
Disallow: /review/hall-of-fame
Disallow: /reviews/top-reviewers/
Disallow: /reviews/top-reviewers
Disallow: /reviews/hall-of-fame
Disallow: /hz/help/contact/*/message/$
Disallow: /-/
Disallow: /gp/aw/shoppingAids/

User-agent: EtaoSpider
Disallow: /

# Sitemap files
```

```
Sitemap: https://www.%m%z.es/sitemaps.2307ea63773dfee.SitemapIndex_0.xml.gz
Sitemap: https://www.%m%z.es/sitemaps.d449d7f825f081e.SitemapIndex_0.xml.gz
Sitemap: https://www.%m%z.es/sitemaps.e49bfbf08ac5517.SitemapIndex_0.xml.gz
Sitemap: https://www.%m%z.es/sitemaps.95918f5af3f77b0.SitemapIndex_0.xml.gz
Sitemap: https://www.%m%z.es/sitemaps.02b411f7e4baecc.SitemapIndex_0.xml.gz
```

El propietario de la web, incluye en su fichero robots.txt una serie de páginas sobre las que no permite su acceso a ningún motos de busqueda o rastreador. Esto viene indicado en la primera de las líneas *User-agent:* * (* = para todos los que quieran acceder) y una serie de enlaces a su web relacionados en cada línea *Disallow: xxx.* Incluye una mención especial al buscador "*EtaoSpider"* al que bloquea para su rastreo y acceso a cualquier página de la web.

La primera de las prohibiciones `Disallow: */s?k=*&rh=n*p_*p_*p_` hace referencia a la prohibición de acceso al buscador de la web y a cualquiera de sus artículos. Añade otras muchas páginas, sin acceso como contenido multimedia, imágenes, perfiles de usuarios, datos de vendedores, valoraciones, comentarios, ayudas de venta, etc. Sí que permite el acceso a la lista de deseos.

Por último, incluye 5 Sitemap con url a páginas en las que se indexa todo el contenido que la web permite rastrear para faciliar la labor de los motores de búsqueda (www.sitemaps.org, 2020). Hemos intentado acceder a su contenido, pero nos han denegado su acceso. Se incluye codigo para acceso a la web y lectura. Devuelve código de estado 403.

```
In [8]:
# descargar una página web
import requests
page = requests.get ('https://www.$m$z.es/sitemaps.2307ea63773dfee.SitemapIndex_0.xml.gz')
# Devuelve mensaje de respuesta
print ("el código de estado de la consulta a la página es {}.".format(page.status_code))
page.content

el código de estado de la consulta a la página es 403.

Out[8]: b'<?xml version="1.0" encoding="UTF-8"?>\n<Error><Code>AccessDenied</Code><Message>Access Denied</Message><RequestId>9C2A22
E5D738575A</RequestId><HostId>cTpuAWwK8x4FCVLZ9qtPZw0FgQ1AIFaZYfxtGJ3f+waL/89cfQnLcFyOd1dW4qW2I6+cBpoBtKE=</HostId></Error>
'
```

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

El nombre elegido para el data set es: "Características e Imágenes de los Smartphones en venta en Amazon" puesto que este nombre refleja claramente que datos almacena y de dónde provienen dichos datos.

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).

El conjunto de datos hace referencia a smartphones en venta en la web Amazon. Podemos encontrar smartphones fabricados en los últimos 20 años, desde los clásicos con botones hasta el último Samsung Galaxy con 5G. De cada teléfono se guarda información de sus características técnicas principales, información comercial, imágenes y precio de venta.

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

Alumnos: **Diego Martín Montoro**
Javier Hernández Hernández
PRAC1

Hemos encontrado en internet una imagen que representa fielmente el contenido de nuestro dataset.

Ilustración 1 Imagen de móviles. https://www.bing.com/images/search?view=detailV2&ccid=29JuaTmG&id=8E9607700488F8562324798C7D9C3DC94ECDA486&thid=OIP.29JuaTmGiLVbztOCordbrAHaDt&mediaurl=https%3A%2F%2Fwww.sitiosargentina.com.ar%2Fwp-content%2Fuploads%2F2016%2F05%2Fplan

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

El conjunto de datos se compone de los siguientes campos:

- **Marca**: marca del teléfono, suele coincidir con el fabricante.
- **Fabricante**: empresa fabricadora del teléfono.
- **Modelo**: nombre del modelo del teléfono, suele contener la gama y el número dentro de la misma. Ej: Xiaomi MI 3
- **Sistema operativo**: Indica el nombre del sistema operativo instalado de fábrica en el teléfono.
- **Capacidad de la memoria**: cantidad de memoria secundaria que tiene el teléfono.
- **Memoria extraíble**: indica el nombre de la tecnología de memoria extraíble que tiene el teléfono, en el caso que goce de ella.
- **Resolución del sensor óptico**: indica los megapíxeles de la cámara trasera del teléfono.
- **Valoración**: valoración media de los clientes sobre 5. NOTA: el punto es separador de decimales.
- **Valoraciones**: número total de valoraciones de cliente que ha recibido el teléfono. NOTA: el punto es separador de miles.
- **Precio**: precio actual del teléfono.
- **Imagen**: ruta de la imagen del teléfono en la carpeta Pictures.

NOTA IMPORTANTE: un mismo teléfono puede estar anunciado por varios vendedores con distintas estadísticas de ventas y precio.

Los datos corresponden al intervalo de años desde el 2000 aproximadamente hasta el día de hoy. Es decir, actualmente están todos estos teléfonos en venta. Por otro lado, no podemos saber la fecha de caducidad de los mismos ya que esto depende del mercado.

Los datos de estos teléfonos han sido introducidos en la plataforma por los vendedores que están inscritos en la misma.

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

www.Amazon.es es el portal en España de la web www.Amazon.com. La web está mantenida por **Amazon Europe Core S.à r.l.** y operado por **Amazon Services Europe S.à r.l.** ambas con domicilio social en Luxemburgo.

Es la compañía de comercio electrónico mas grande de Estados Unidos, con la mayor variedad de productos del mundo. Por la cantidad de información, y su complejidad, hemos considera una fuente muy valiosa para poner en práctica nuestro código de web scraping.

Existen diversas entradas en internet con estudios y trabajos sobre como hacer web scraping a la web de Amazon.es. Algunas utilizan el scraping por medio de extensiones del navegador, como web scraper. En su web, incluye videos tutoriales para hacer web scraping (Scraper, 2020). Otros utilizan sotware propio como es el caso de ProWebScraper (ProWebScraper, 2020), ParseHub (ParseHub, 2020) u Octoparse (Octoparse, 2020).

Pero la más utilizada para hacer web scraping, es la programación por python. Existen diversas referencias en internet. Destacar la web de ScrapeHero que resuelve posible bloqueos de la amazon a un scraping de la web (ScrapeHero, 2020). La web towardsdatascience también explica cómo diseñar un código de web scraping para crear una alerta de precio sobre la web de amazon (towardsdatascience, 2020).

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

Este conjunto de datos es muy interesante porque aporta mucha información sobre el mercado existente de un dispositivo que hoy en día es indispensable en la vida diaria.

Mediante este dataset podemos responder preguntas como:

- ¿Qué móviles son los más vendidos?
- ¿Qué móviles son los más valorados por la comunidad?
- ¿Qué móvil recomendar dado un margen económico?
- ¿Cómo afecta el tamaño de la pantalla en el precio?
- ¿Cuánto podría costar este teléfono?
- ¿Cuánto afectó la salida al mercado del nuevo iPhone?
- …

E infinidad de preguntas que se podrían responder con cierta seguridad usando este dataset.

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:**

Alumnos: **Diego Martín Montoro**
Javier Hernández Hernández
PRAC1

La legislación europea señala que tendrán la consideración de “base de datos” las recopilaciones de datos … dispuestos de manera sistemática o metódica y accesibles individualmente por medios electrónicos o de otra forma (europeo, 1996) (art. 1.2).

Partimos del hecho que nuestra base de datos es ‘sui generis’, en el sentido que no es una base de datos original y se ha requerido de una inversión sustancial para la obtención, verificación y presentación de la información a partir de información existente en una base de datos original (Ramos-Simón, 2017). Es importante subrayar que si bien en las bases de datos originales se protege la totalidad de la base de datos, en las bases de datos “sui generis” se protege la inversión sustancial y durante un periodo de quince años (Cámara-Lapuente, 2007).

Desde hace unos años, y cada vez más, organizaciones e investigadores ponen a libre disposición de la comunidad, sus bases de datos para contribuir al progreso y la investigación. Los titulares de estas bases de datos, protegen mediante las distintas licencias abiertas, el uso, forma de explotación y alteración que se pueda dar a sus bases de datos. Surge de este modo, la necesidad de crear licencias de uso que regulen estas limitaciones de uso y favorezcan la compartición de información sin barreras.

Actualmente, las utilizadas con más frecuencia son las licencias Creative Commons (www.creativecommons.org) y de Open Data Commons (www.opendadatacommons.org). Las licencias públicas de Creative Commons (creativecommons.org, www.creativecommons.org, 2020) proporcionan un conjunto estándar de términos y condiciones que los creadores y otros titulares de derechos pueden utilizar para compartir obras originales de su autoría y cualquier otro material sujeto a derechos de autor y a otros derechos que se especifican en cada una de las clases de licencia pública.

Por último, resaltar un par de cuestiones. Las licencias son irrevocables, de modo que con carácter previo a su elección, deben leerse y entenderse los términos y condiciones de cada licencia. Y, al solicitar una licencia, debe contarse con los derechos y autorizaciones que se requieran (potestad legal) para el uso y publicación del material.

- **Released Under CC0: Public Domain License** (creativecommons.org, CC0 1.0 Universal, 2020)

La persona que ha asociado una obra a esta licencia ha dedicado la misma al dominio publico, liberándola de forma mundial y en la medida que lo permita la ley, de todos sus derechos de propiedad intelectual, incluyendo todos los derechos conexos. Al renunciar permanentemente de estos derechos sobre una obra, se alcanza el propósito de contribuir con la publicación de una obra, sobre la que cualquiera puede hacer uso sin miedo a recibir reclamaciones por vulneración de los derechos de autor o propiedad.

Se renuncia a todos los derechos autorales y derechos afines o relacionados que pueda tener en todas las jurisdicciones del mundo. Puede copiar, modificar, distribuir la obra y hacer comunicación pública, incluso para fines comerciales, sin pedirnos permiso para ello.

No se puede renunciar a derechos sobre una obra de la que no se posee al menos permiso de la propietaria. Es necesario por tanto, para publicar bajo esta licencia, contar con la autorización de su propietaria, en el caso de que esta exista.

En definitiva, esta licencia es “No Right Reserved”. Está orientada al uso de datos y metadatos (Ramos-Simón, 2017).

- **Released Under CC BY-NC-SA 4.0 License**

A efectos de esta licencia, se garantiza que el material publicado bajo la misma no se destinará a obtener una ventaja comencial o compensación monetaria (creativecommons.org, CC BY-NC-SA 4.0, 2020).

El material publicado se puede:

- Compartir — copiar y redistribuir el material en cualquier medio o formato
- Adaptar — remezclar, transformar y crear a partir del material

Al compartir, debe reconocerse la autoría del material, incorporando un enlace a la licencia, e indicar si se han realizado cambios. Así mismo, en el supuesto de adaptar, deberá difundirse el nuevo material bajo la licencia original.

- **Released Under CC BY-SA 4.0 License**

Esta licencia coincide con la anterior (BY-NC-SA 4.0) salvo en la restricción de uso no comercial (creativecommons.org, CC BY-SA 4.0 License, 2020) que sí está permitida.

- **Database released under Open Database License, individual contents under Database Contents License.**

La Open Knowledge Fundation regula licencias para la publicación de bases de datos (Open Knowledge Fundations, 2020). Concretamente, la Open Database License (OdbL) recoge los términos y condiciones que regulan la publicación de bases de datos, y los derechos y obligaciones que recaen sobre el publicador y el usuario.

Permite a los usuarios modificar, compartir y utilizar libremente las bases de datos publicadas bajo esta licencia. Sólo son de aplicación a la base de datos, y no a su contenido que pudiera ser en forma de imágenes, material audiovisual o sonidos. Por tanto, para otorgar licencia sobre el contenido deberá incluirse otro tipo de licencia que los regule.

Esta licencia no aplica a los programas informáticos utilizados para la creación de bases de datos, no protege sobre ninguna patente que pudiera existir sobre los contenidos y no cubre ninguna marca comecial asociada a la base de datos.

Al publicar una base de datos bajo esta licencia se otorga una autorización de uso de carácter mundial, gratuita, no exclusiva y rescindible para su uso durante la duración que se haya determinado. Se permite la utilización para uso comercial y la creación de bases de datos derivadas. En este último caso, la nueva base de datos debe incorporar una copia de la licencia original.

- **Other (specified above)**

Dentro de la familia de licencias Open Data Commons, también existen otros dos tipos de licencias:

- Public Domain Dedication and License (PDDL). Es una licencia similar a la CC0 de Creative Commons en la que la con cesión al dominio público, el titular de los derechos de autor ofrece la obra al beneficio público y renuncia a los derechos de autor y derechos *sui generis* que puedan recaer sobre la base de datos, por lo cual la obra es libre y está abierta para otros usos. Asimismo, la renuncia es para todo el mundo y para todos los derechos presentes y futuros en cualquier formato. Del mismo modo, la licencia prevé que si la renuncia resulta inaplicable en alguna jurisdicción, el licenciador autoriza todos los usos sobre la base de datos mediante una licencia no exclusiva, libre de derechos en todo el mundo y por todo el tiempo de protección de la obra. Además, la licencia contiene una renuncia expresa al ejercicio de los derechos morales.

- Attribution License (ODC-By). Esta licencia se equipara a la CC-BY de Creative Commons. La licencia permite a los usuarios compartir libremente, modificar y usar la base de datos con el único requisito de atribuir la paternidad de la obra, libre de derechos, no exclusiva por todo el tiempo que dure la protección de la base de datos, con derecho de extracción y reutilización de la base de datos. La base de datos puede estar protegida por otros derechos que no están contenidos en la licencia (datos personales, contratos privados o marcas…), asimismo recoge la renuncia a los derechos morales del titular.

- **Unknown License**

Según países, la publicación de bases de datos sin una licencia, implica por defecto la asignación de una determinada licencia (Ramos-Simón, 2017).

Tabla 1/ Tabla modalidades y compatibidad de licencias libres, según países

Licencia por defecto	Países	Compatibilidad entre licencias
Sin licencia	Estados Unidos Suecia	-CC-0 y PDDL -(https://theunitedstates.io/licensing/) Datos libres de pago y de licencias
CC0 y CC-BY	Dinamarca	-Atribución del titular: CC-BY -Sin Atribución: CC-0
CC-BY 4.0	Holanda	CC-BY 4.0
OGL	Reino Unido Francia Canadá España	CC-BY 4.0 y ODC-BY ODC-BY y CC-BY 2.0 No explícita No explícita
CC-BY 3.0	Nueva Zelanda Noruega Australia	No se especifica -No base de datos: CC-BY 3.0 -Sí base de datos: ODC-BY No se especifica

- **Conclusión.**

Una vez revisadas las distintas clases de licencia existentes para hacer pública nuestra base de datos, veamos cuál es la que mejor aplica a nuestra base de datos. Indicar que al contener información procedente de otra fuente, en este caso de Amazon.es, nuestra base de datos puede entenderse que es de clase “sui generis”. Debemos conocer la existencia de derechos de autor,

de propiedad intelectual o sobre bases de datos, estén protegiendo la información que pretendemos extraer de la web. En su política de condiciones de uso y venta, incorpora el siguiente punto.

> **3. Derechos de autor, derechos de propiedad intelectual y derechos sobre bases de datos**
>
> Todo contenido alojado o puesto a disposición en cualquiera de los Servicios de Amazon, como por ejemplo texto, gráficos, logotipos, iconos de botones, imágenes, clips de audio, descargas digitales, y recopilaciones de datos son propiedad de Amazon o de sus proveedores de contenido y está protegido por las leyes de Luxemburgo, así como por la legislación internacional sobre derechos de propiedad intelectual, derechos de autor y derechos sobre bases de datos. El conjunto de todo el contenido albergado o puesto a disposición a través de cualquier Servicio de Amazon es propiedad exclusiva de Amazon, y está protegido por las leyes de Luxemburgo e internacionales sobre derechos de propiedad intelectual y derechos sobre bases de datos.
>
> No está permitida la extracción sistemática ni la reutilización de parte alguna del contenido de ninguno de los Servicios de Amazon sin nuestro expreso consentimiento por escrito. En particular, no se permite el uso de herramientas o robots de búsqueda y extracción de datos para la extracción (ya sea en una o varias ocasiones) de partes sustanciales de los Servicios de Amazon para su reutilización sin nuestro expreso consentimiento por escrito. Tampoco le está permitido al usuario crear ni publicar sus propias bases de datos cuando éstas contengan partes sustanciales de cualquiera de los Servicios de Amazon (por ejemplo, nuestras listas de productos y listas de precios) sin nuestro expreso consentimiento por escrito.

Ilustración 2 Condiciones de uso en Amazon.es
https://www.amazon.es/gp/help/customer/display.html/ref=hp_left_v4_sib?ie=UTF8&nodeId=GLSBYFE9MGKKQXXM#GUID-6F014DE6-CB56-4A28-BCAA-38F5D24E6324__SECTION_C77666A6124F4AD6AAA88727F743CBD1

La información extraída de la web Amazon.es está protegida por los derechos de autor, propiedad intelectual y sobre bases de datos. Además, el propietario de la web prohíbe la extracción sistemática y la reutilización del contenido sin expreso consentimiento por escrito. Para poder extraer mediante técnica de web scraping y publicar la información obtenida, bajo el paraguas de common creatives, debemos contar con la autorización de Amazon.es. Se ha enviado email a la web explicando nuestra intención para fines académicos y sin intención comercial, solicitando autorización para extracción y publicación de los datos obtenidos.

No hemos recibido respuesta por parte de Amazon.es a nuestra petición, por lo que hemos entendido que el consentimiento no se ha prestado. Para avanzar en la práctica, se ha decidido publicar un dataset mínimo “dummy” con un subconjunto de filas (smartphones) de los que sólo incluimos información de marca, modelo y precio.

No podemos incluir a nuestro dataset licencia de Creative Commons por no contar con autorización del propietario de la información extraída. Como cita (Ramos-Simón, 2017), en el ámbito de las licencias Creative Commons los dos tipos de licencia que se usan con frecuencia en la apertura de datos son la licencia CC-BY (Reconocimiento) y CC-SA (Compartir igual). En la primera, el titular de los datos exige que se reconozca su titularidad. En la segunda, CC-SA, el titular permite reutilizar y compartir los datos con tal de que la distribución de esos datos resultantes se haga con la misma licencia. Estas dos licencias no son específicas para compartir bases de datos *sui generis*, aunque siguen siendo las más utilizadas para estos propósitos.

Ampliando información relativa al acceso a la información contenida en la web de Amazon.es, en el punto 6 de la Condiciones de Uso, habla de la concesión de una licencia limitada de acceso y utilización para fines personales no comerciales. No incluye el derecho a descargar información

para el beneficio de otra empresa. Consideramos que nuestra intención y finalidad no infringe este punto de Licencia y acceso.

6. Licencia y acceso

Sujeto a tu cumplimiento de estas Condiciones de Uso y las Condiciones Generales de los Servicios aplicables, así como al pago del precio aplicable, en su caso, Amazon o sus proveedores de contenidos te conceden una licencia limitada no exclusiva, no transferible y no sublicenciable, de acceso y utilización, a los Servicios de Amazon para fines personales no comerciales. Dicha licencia no incluye derecho alguno de reventa ni de uso comercial de los Servicios de Amazon ni de sus contenidos, ni derecho alguno a compilar ni utilizar lista alguna de productos, descripciones o precios. Tampoco incluye el derecho a realizar ningún uso derivado de los Servicios de Amazon ni de sus contenidos, ni a descargar o copiar información de cuenta alguna para el beneficio de otra empresa, ni el uso de herramientas o robots de búsqueda y extracción de datos o similar.

Ilustración 3 Punto 6 Licencia y acceso. www.amazon.es

La información contenida en nuestra base de datos puede ser de utilidad para elaborar estudios de mercado. No queremos permitir su alteración para preservar siempre los datos originales. Es nuestro deseo contribuir con la comunidad, facilitando el acceso a esta información, sin posibilidad de modificación bajo una licencia. No podemos regular el contenido del data set, pero sí la inversión sustancial realizada para su creación. Por estos motivos, consideramos apropiada una licencia Public Domain Dedication and License (PDDL) de Open Knowledge Fundation.

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

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

A zenodo hemos subido un dataset Dummy.

El DOI obtenido es el siguiente**: 10.5281/zenodo.4135053**

11. Problemas en el camino.

Durante en desarrollo del scraper, hemos tenido dificultades para evitar ser bloqueados por la web de Amazon, por eso hemos realizado las siguientes acciones:

- Hemos desarrollado un código para rotación de proxies con una lista de más de 1000 direcciones IP distintas.
- También hemos implementado una rotación de User Agents (un total de 20).
- Hemos insertado esperas aleatorias entre peticiones, de manera que se simule el comportamiento humano.
- Y para aumentar la tasa de descarga hemos paralelizado la solución, creando 12 hebras que se reparten los user agents y los proxies.

Como resultado hemos sido capaces de descargar información de 3694 teléfonos y un total de 2434 imágenes en 3 ocasiones sin ser bloqueados, teniendo una tasa de descarga bastante alta.

Alumnos: **Diego Martín Montoro**
Javier Hernández Hernández
PRAC1

El proceso de extracción se divide en dos fases, una primera que nos brinda los enlaces a las páginas detalladas de los teléfonos, y una segunda extracción que usa dichos enlaces para obtener el dataset final.

Al usar los 7000 links obtenidos en la primera fase, resulta que hemos obtenido muchísimos datos menos (en torno a 3600). Por lo que hemos estado observando, creemos que se debe a que:

- Hay una gran mayoría de artículos de telefonía camuflados como teléfonos móviles que salen al realizar la búsqueda en amazon y esto no podemos evitarlo ya que el vendedor lo hace a propósito. Dichos productos carecen de la descripción técnica típica de un teléfono, lo que resulta en obtención de tuplas casi vacías. En realidad creemos que esto no es un problema ya que estas tuplas simplemente corresponden a objetos que están fuera de nuestro interés, y solo ocasionará tener que limpiar más el dataset.
- De los productos que sí son teléfonos pueden ocurrir varias cosas:
    - Descripción por defecto de amazon: no hay problemas la extracción se realiza correctamente.
    - Descripción por defecto pero no introdujo datos, o los introdujo erróneamente: estamos con las manos atadas.
    - La descripción consiste en una imagen que se corresponde con un folleto informativo: creemos que estas son las que más "daño" nos hacen ya que suelen ser de empresas grandes y consolidadas que aportan información de mucha calidad y cuyos dispositivos tienen un gran peso en el mercado.
    - Descripción personalizada: el vendedor ha embebido el código html sin ningún tipo de estándar, datos totalmente desestructurados, en distintos idiomas y muchas veces con errores. Imposible proceder.

Esta serie de situaciones creemos que han llevado a la reducción considerable de la cantidad de información que en una primera instancia creíamos que íbamos a poder extraer. Hemos llegado a esta conclusión estudiando la lista de teléfonos ofertada por amazon manualmente. De las 400 páginas que aparecen (30 teléfonos por página), en las últimas 150 páginas todos son productos que carecen de información, algunos sin precio y otros que nunca llegaron a estar en venta realmente y no disponen de foto, suponemos que los vendedores los pusieron con el objetivo de ganar algún tipo de ventaja en la indexación de amazon. De las 250 restantes, hay gran cantidad de anuncios que sufren alguna de las situaciones comentadas con anterioridad.

Me temo que esto es típico del web Scrapping y hay que acostumbrarse a ello.

CONTRIBUCIONES	FIRMA
Investigación previa	DMM, JHH
Redacción de las respuestas	DMM, JHH
Desarrollo código	DMM, JHH

Referencias

Cámara-Lapuente, S. (2007). *Derecho sui generis sobre las bases de datos.* Editorial Thomson-Civitas.

creativecommons.org. (octubre de 2020). *CC BY-NC-SA 4.0*. Obtenido de https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.es

creativecommons.org. (octubre de 2020). *CC BY-SA 4.0 License*. Obtenido de https://creativecommons.org/licenses/by-sa/4.0/deed.es_ES#

creativecommons.org. (octubre de 2020). *CC0 1.0 Universal*. Obtenido de https://creativecommons.org/publicdomain/zero/1.0/legalcode.es

creativecommons.org. (octubre de 2020). *www.creativecommons.org*. Obtenido de https://creativecommons.org/

europeo, P. (1996). *Directiva 96/9/CE del Parlamento Europeo y del Consejo, de 11 de marzo de 1996, sobre la protección jurídica de las bases de datos.* http://www.boe.es/buscar/doc.php?id=DOUE-L-1996-80413.

Octoparse. (2020). *Octoparse*. Obtenido de https://www.octoparse.es/blog/como-scrape-datos-de-productos-de-amazon

Open Knowledge Fundations. (2020). *Open Data Commons*. Obtenido de https://opendatacommons.org/licenses/

ParseHub. (2020). *ParseHub*. Obtenido de https://www.parsehub.com/blog/scrape-amazon-product-data/

ProWebScraper. (2020). *www.medium.com*. Obtenido de https://medium.com/prowebscraper/how-to-scrape-amazon-product-data-adb07ea7cc12

Ramos-Simón, L. F. (2017). *El uso de las licencias libres en los datos públicos abiertos. .* Obtenido de Revista Española de Documentación Científica, 40 (3): e179: http://orcid.org/0000-0003-2267-8405

ScrapeHero. (2020). *www.scrapehero.com*. Obtenido de https://www.scrapehero.com/tutorial-how-to-scrape-amazon-product-details-using-python-and-selectorlib/

Scraper, W. (2020). *Web Scraper*. Obtenido de https://www.webscraper.io/tutorials

towardsdatascience. (2020). *www.towardsdatascience.com*. Obtenido de https://towardsdatascience.com/scraping-multiple-amazon-stores-with-python-5eab811453a8

www.sitemaps.org. (2020). Obtenido de https://www.sitemaps.org/es/protocol.html

Tablas

Tabla 1/ Tabla modalidades y compatibidad de licencias libres, según países

Alumnos: **Diego Martín Montoro**
Javier Hernández Hernández

PRAC1

Ilustraciones