prÁctica 1

Realización de Web Scraping con Python en la página web de Mil Anuncios y posterior almacenamiento en soporte CSV-Excel

Tipología y ciclo de datos

Contenido

[1. Componentes del grupo 2](#_Toc5971697)

[2. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información. 3](#_Toc5971698)

[3. Definir un título para el dataset. Que sea descriptivo. 4](#_Toc5971699)

[4. Descripción del dataset. Desarrollar una breve descripción de la información aportada. 4](#_Toc5971700)

[5. Representación gráfica. Presentar una imagen o esquema que identifique el *dataset visualmente.* 5](#_Toc5971701)

[Una vez ejecutado el script, el Dataset Original es el mostrado a continuación: 5](#_Toc5971702)

[6. Contenido. Explicar los campos del Dataset Obtenido. 8](#_Toc5971703)

[7. Agradecimientos. Presentar al propietario del conjunto de datos. 10](#_Toc5971704)

[8. Inspiración. Explique por qué es interesante este conjunto de datos y que preguntas puede responder. 10](#_Toc5971705)

[9. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección 12](#_Toc5971706)

[10. Contribuciones 13](#_Toc5971707)

1. Componentes del grupo

David Quiles Gómez

Iván López-Baltasar Benito

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

En la red disponemos de varias webs y portales que contienen gran información relativa a la compra-venta de vehículos de segunda mano, km 0 y de ocasión. Realizando Web Scraping a las principales web de este tipo podríamos obtener información muy rápidamente de gran cantidad de vehículo de cualquier provincia de España faciliando además todas aquellas carácteristicas que suelen buscar los usuarios como Kilometraje, precio, estado, color, etc.

Hemos elegido la pagina Web MilAnuncios.com debido a que contiene gran cantidad de información bien estructurada y responde prácticamente a todas las preguntas que suelen formularse los usuarios. En esta Web, se dispone de información no sólo de particulares sino también de concesionarios que publican semanalmente gran cantidad de información de todos sus vehículos. Además, esta web también nos ofrece información de cualquier parte de España.

Para la obtención de la información se ha utilizado un script en Python. En principio hemos tenido bastantes problemas a la hora de ejecutar el script ya que en la primera conexión, el sistema detectaba que las peticiones no las estaba haciendo un humano y nos mostraba una página donde se solicitaba un captch o similar. Suponemos que dada la naturaleza de la información que MilAnuncios proporciona, es un sitio muy susceptible de recibir peticiones automáticas.

Hemos probado bastantes cosas, lo primero de todo fue establecer unas cabeceras en la petición donde entre otras cosas, se informara del “Agent User”. Así nos mostraba dos peticiones, a la tercera el sistema volvía a detectar que las peticiones eran realizadas por un “bot”. Probamos bastantes cosas más como espaciar las peticiones mediante ejecuciones “sleep” de duración aleatoria, así como intentar salir por diferentes proxys (esto lo vimos en diferentes sitios de internet donde daban consejos a la hora de evitar baneos al realizar web scraping). También probamos utilizando Selenium, alternando incluso el navegador a utilizar, pero como máximo obteníamos 3 peticiones antes del baneo.

Finalmente lo que mejor resultado nos ha dado ha sido realizar una conexión previa con un navegador, revisar las cabeceras enviadas en la petición (a través de un inspector de red incorporado en FireFox Developer Edition –fig 1), obtener el código de la cookie establecida y ese código meterlo en las cabeceras de la petición (fig 2).

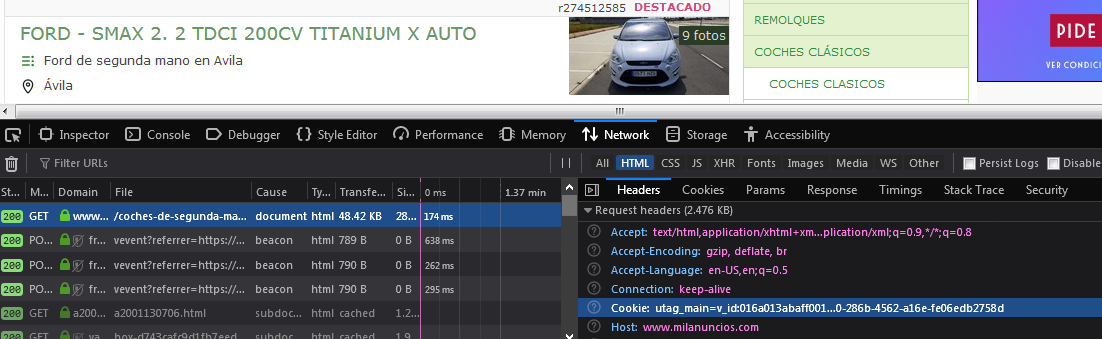
**

Fig 1 Revisión de las cabeceras enviadas por el navegador en la petición

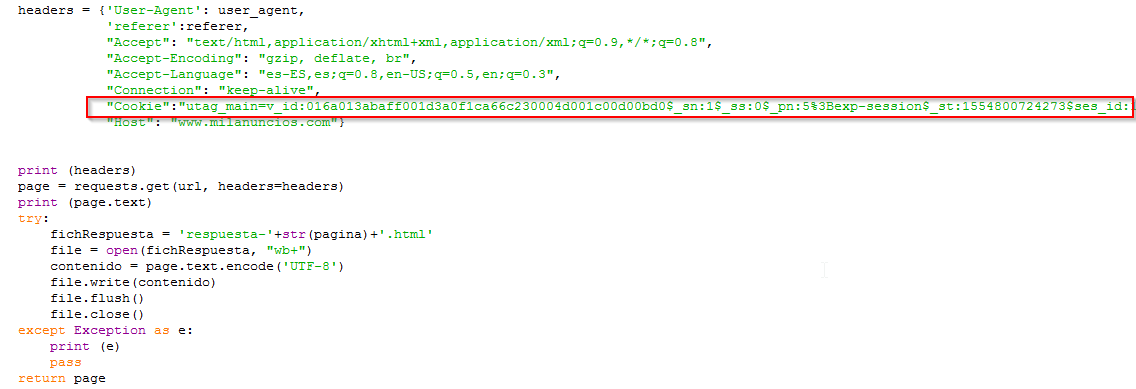
**

Fig 2 Fragmento del código Python donde se incluye la cookie en la cabecera

1. Definir un título para el dataset. Que sea descriptivo.

El título elegido para el dataset será “FindMyCar”.

Obviamente elegimos un título que pueda dar idea rápida de cuál será su utilidad.

1. Descripción del dataset. Desarrollar una breve descripción de la información aportada.

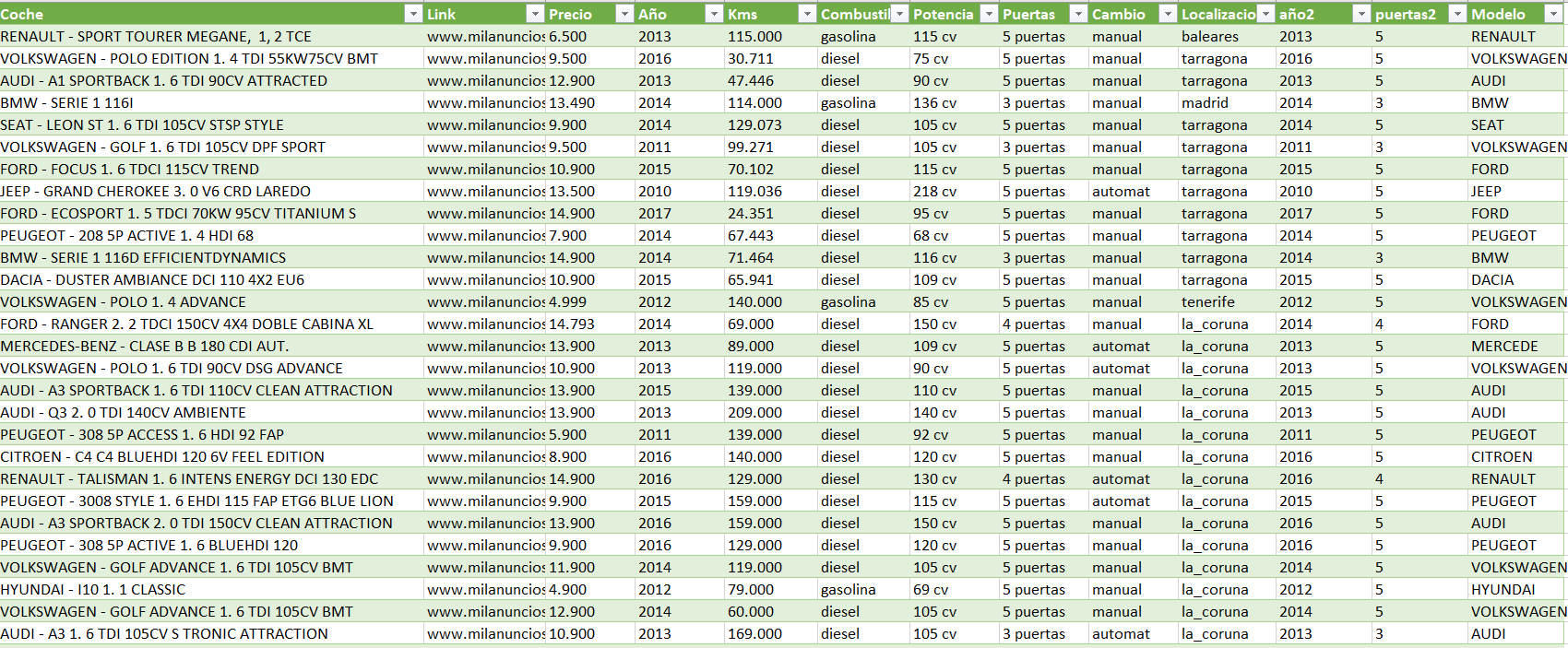
El Dataset obtenido muestra toda la información que un usuario corriente podría considerar importante a la hora de comprar un vehículo. El dataset se muestra en filas fácilmente manipulable para posteriormente poder ser fitrado, manipulado, almacenado en una BBDD o bien ser tratado con alguna técnica de Minería de datos como Clustering o Clasficación por ejemplo, si los consideramos necesario y nuestro objetivos va más allá que del simple hecho de comprar un coche. Quizás nos interese conocer en qué ciudad se vende más un tipo de coche o en que provincia disponemos de mayor stock, etc.

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

Una vez ejecutado el script, el Dataset Original es el mostrado a continuación:

**

Tras unos pequeños arreglos que nos facilitarán la realización de las gráficas, el dataset obtenido para poder obtener conclusiones es el siguiente:



A continuación se muestran unos sencillos gráficos que nos permitirán utilizar los datos y obtener interesantes conclusiones. El objetivo de las siguientes gráficos es mostrar la utilidad del D

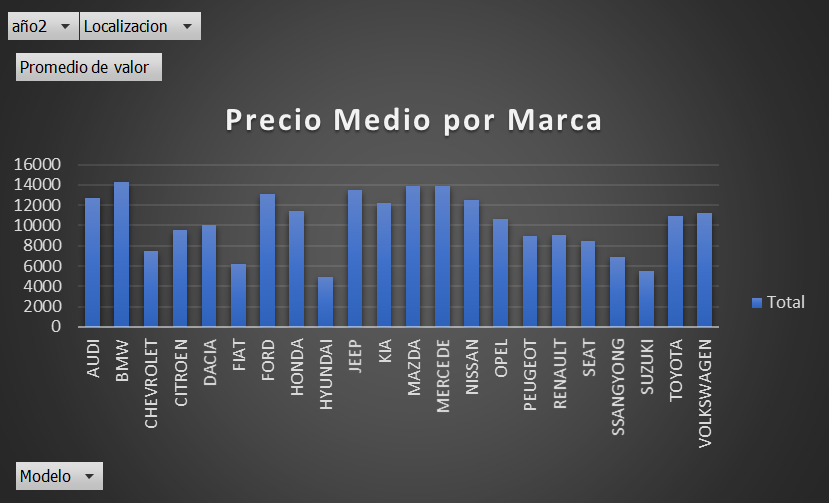
En este primer gráfico podemos observar como la gran mayoría de coches disponibles en MILANUNCIOS tienen entre 4 y 6 años.

Por tipo de combustible en el que la gran mayoría de coches que se venden por diesel.

En un gráfico algo más complejo con más variables, podremos obtener en que provincia se venden los coches más baratos.

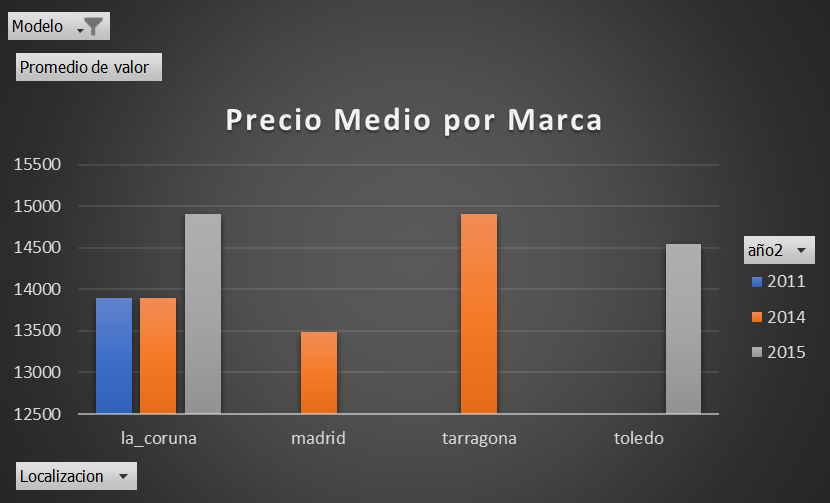
A la vista del gráfico se podría concluir que en Málaga y Tenerife es donde más barato se venden los coches.

También podremos obtener fácilmente, cuáles son las marcas que se venden a precios mas económicos o más elevados.



Supongamos un ejemplo algo más complejo: Si queremos comprar un BMW, podriamos obtener fácilmente qué provincias son las que tienen mejores precios .

Una rápida conclusión, nos llevaría a pensar que los BMW que se venden en Madrid y que son del 2014 tiene el mejor precio medio por coche.



1. Contenido. Explicar los campos del Dataset Obtenido.

El dataset obtenido tiene la siguiente información:

* MODELO 🡪 Modelo y marca del vehículo.
* LINK 🡪 El link para acceder directamente a la web donde está todo el contenido.
* PRECIOS 🡪 El precio del vehículo
* AÑO MATRICULACIÓN.
* KILOMETROS
* TIPO DE COMBUSTILES 🡪 Gasoil, diesel, etc.
* POTENCIA 🡪 en CV
* PUERTAS 🡪 3 o 5 Puertas
* CAMBIO 🡪 Manual o automático.
* LOCALIZACION 🡪 Provincia donde se encuentra localizado el vehículo

6.1. VALIDEZ DE LOS DATOS

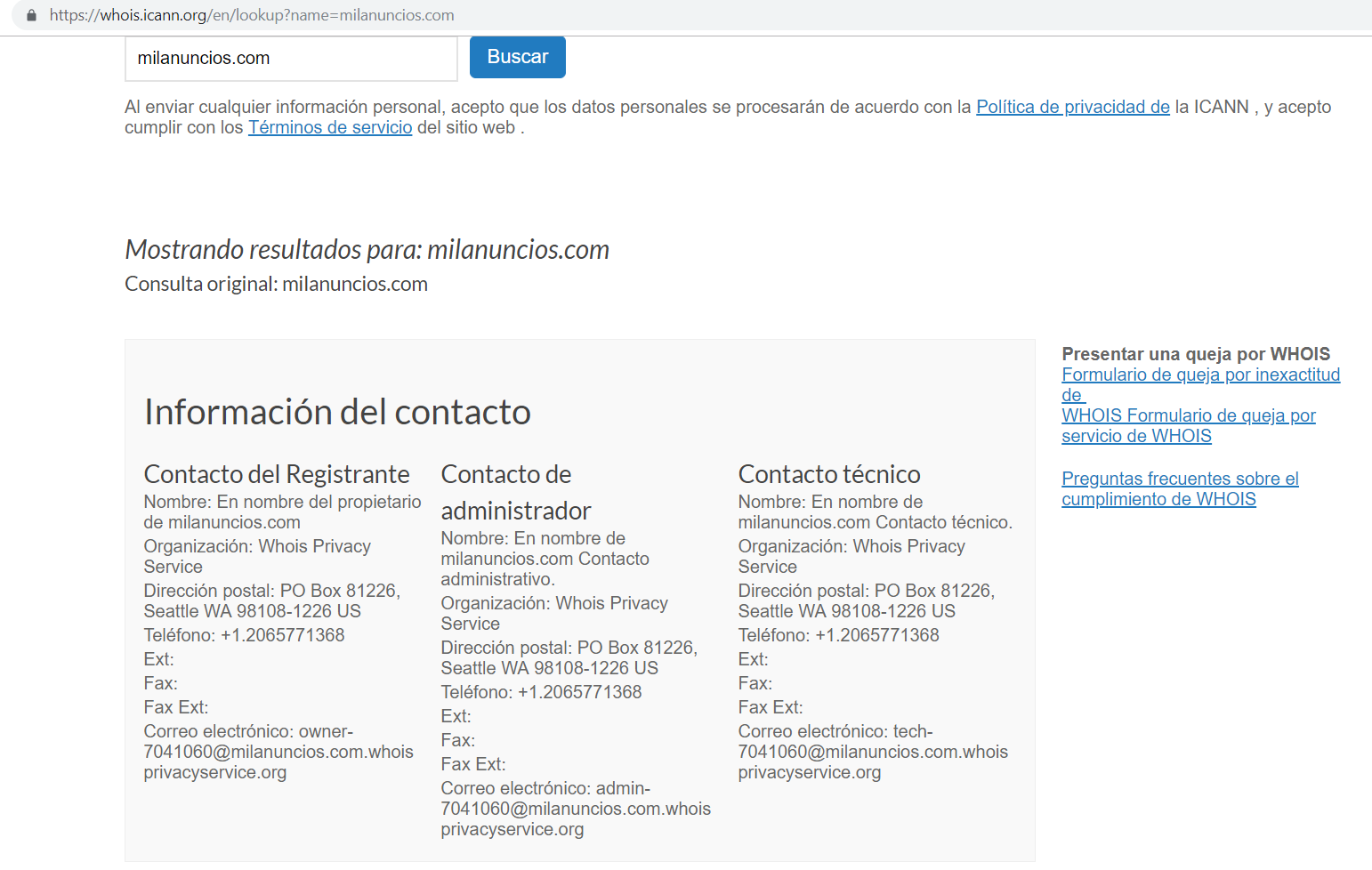
No somos expertos en el business de la compra-venta de vehiculos usados, pero es obvio que los datos deberán ser refrescados bastante a menudo aunque algunas métricas puedan ser bastante estables. La compra-venta de vehículos está en constante evolución y semanalmente se traduce en cientos de operaciones que podrían modificar nuestras conclusiones.

Un ejemplo podría ser el fenómeno reciente en el que los vehículos diésel han sido penalizados de un mes para otro con nuevos impuestos que han hecho que su precio y su demanda se reduzcan considerablemente.

1. Agradecimientos. Presentar al propietario del conjunto de datos.

Como no podía ser de otra manera, agradecemos a todo el equipo de milanuncios.com su extraordinario trabajo. Acceder a su web a sido sencillo y sus datos han sido fácilmente manejables.

El propietario al que debemos el éxito de nuestro trabajo es:



Los propietarios de la web no están identificados. Hemos utilizado la herramientas whois para intentar averiguar quien son los propietarios pero su identidad está oculta.

1. Inspiración. Explique por qué es interesante este conjunto de datos y que preguntas puede responder.

Una de las inversiones más complejas y difíciles que tiene cualquier persona es la adquisición de un vehículo. La búsqueda de uno que cubra todas nuestras necesidades y expectativas suele ser cuestión de semanas o incluso meses, visitando diferentes Webs de coches de ocasión, concesionarios, mercado particular, etc.

El proyecto puede ser interesante tanto del lado del comprador como por parte del vendedor

Por parte del comprador, a través de un proyecto de Web Scraping podemos obtener información precisa de todas esas Webs de manera rápida y concreta y que sin duda nos ahorrará decenas de horas de busqueda por las diversas webs que se dedican a la compraventa de vehículos.

No obstante creemos que la información obtenida es mucho más interesante desde el punto de vista del vendedor puesto que con los obtenidos y obteniendo los datos de coches vendidos podríamos establecer relaciones entre las ventas y las compras en las diferentes provincias de España. Podríamos contestar, como ejemplos, a las siguientes preguntas

* Que coches se venden más en cada una de las provincias
* Qué tipo de coches se venden más en cada una de las provincias (usados, nuevos, etc)
* Que rango de precio tienen los coches vendidos (en cualquier provincia)
* En qué fechas se venden más coches
* Duración media de los stock de los modelos de los diferentes marcas, nuevamente por provincias, etc, etc.
* Que concesionarios tienen precios más competitivos
* Precios medio de los modelos vendidos en cada provincia para comprender si (p.ej.) si la marca Peugeot tiene precios medios más económicos en una provincia que en otra.

Hay inmobiliarias que están comenzando a incorporar a Científicos de Datos para establecer relaciones entre los bienes inmobiliarios y los compradores, de manera que la inmobiliaria puede saber si un piso de 120 m2 con un precio de 145.000 en Getafe será vendido antes de 60 días. La contestación a todas estas preguntas nos podría llevar a sacar conclusiones de si nos interesa adquirir, para después vender, un bien inmobiliario en una determinada ciudad.

De manera análoga, este interesante ejercicio, podría trasladarse a los coches, camiones, etc para tomar decisiones relevantes y optimizar nuestros recursos económicos.

Por supuesto, disponer de este conjunto de datos nos permitirá utilizar técnicas de minería de datos capaces de responder a todas estas preguntas. Este dataset y debido a su claridad es fácilmente manipulable, escalable para poder crear modelos útiles.

1. 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

Cabe recordar inicialmente que estas licencias hacen referencia a las licencias Creative Commons y que cada una de ellas tiene posibilidades para configurarlas y que permite a los autores de los dataset poder decidir de qué manera va a circular en internet su trabajo. Dependiendo de esta configuración, los usuarios podrán publicar, citar, reproducir y crear obras derivadas utilizando el trabajo publicado generalmente bajo ciertas restricciones.

Explicamos rápidamente algunas siglas (llamados módulos oficialmente) de estas licencias:

CC🡪 Creative Commons (abreviatura)

BY 🡪 requiere la referencia al autor original.

SA 🡪 permite obras derivadas bajo la misma licencia o similar.

NC 🡪 obliga a que la obra no sea utilizada con fines comerciales.

ND 🡪 no permite modificar la obra de ninguna manera.

CCO 🡪 Sin derechos reservados.

Estos módulos asimismo se combinan entre si para formar las seis licencias de Creative Commons:

* Atribución (CC – BY)
* Atribución Compartir Igual (CC-BY-SA)
* Atribución No Derivadas (CC-BY-ND)
* Atribución No Comercial (CC-BY-NC)
* Atribución No comercial Compartir Igual ( CC-BY-NC-SA)
* Atribución No comercial No derivadas (CC-BY-NC-ND)

Todas las licencias permiten el derecho fundamental de redistribuir la obra con fines NO comerciales y sin modificaciones, aunque las opciones NC y ND hacen que la obra no sea de libre acuerdo para tal redistribución.

Una vez definidas y recordadas las posibilidades, nuestra licencia tendrá las siguientes características:

* BY, puesto que requerirá al autor original
* NC, puesto que no podremos utilizarla para fines comerciales. Entendamos que milanuncios recibe dinero debido a la publicidad que insertan en su web y que se presupone que ven las personas que la visitan y que obviamente la información que obtenemos de su web es información en propiedad y protegida posiblemente por la LPD.
* ND, puesto que no podremos modificar los registros bajo ningún concepto

Así pues, el trabajo podría tener la licencia más abajo indicada y que no está indicada en las opciones del enunciado:

* Atribución NO comercial NO derivadas (CC-BY-NC-ND)

Una licencia alternativa sería

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

Que permitiría derivar el trabajo bajo la misma licencia y también sin fines comerciales.

Fuente: <https://es.m.wikipedia.org/wiki/Creative_Commons>

Las licencias Creative Commons están compuestas por cuatro módulos de condiciones:

* Attribution/Atribución (BY), requiere la referencia al autor original.
* Share Alike/Compartir Igual (SA), permite obras derivadas bajo la misma licencia o similar (posterior u otra versión por estar en distinta jurisdicción).
* Non-Commercial/No Comercial (NC), obliga a que la obra no sea utilizada con fines comerciales.
* No Derivative Works/No Derivadas (ND), no permite modificar la obra de ninguna manera.

1. Contribuciones

|  |  |
| --- | --- |
| Contribuciones | Firma |
| Investigación previa | David Quiles / Iván López-Baltasar |
| Redacción de las respuestas | David Quiles / Iván López-Baltasar |
| Desarrollo código | David Quiles / Iván López-Baltasar |