

PRÁCTICA 1: WEB SCRAPING









Componentes del grupo: Nerea Calderón Mulero y David Cisneros Martínez

1) EJERCICIO 1:

1. CONTEXTO. EXPLICAR EN QUÉ CONTEXTO SE HA RECOLECTADO LA INFORMACIÓN. EXPLICAR POR QUÉ EL SITIO WEB ELEGIDO PROPORCIONA DICHA INFORMACIÓN:

Para esta práctica de **Tipología y Ciclo de Vida de los Datos**, se ha planteado el estudio e investigación de la mundialmente conocida plataforma de contenido audiovisual **YouTube**. Se trata de una plataforma en la que una serie de individuos llamados “Youtubers” comparten todo tipo de **contenido audiovisual** (como videos, streamings...) y de todo tipo de temáticas. En la otra cara nos encontramos a los consumidores, quienes a través de sus interacciones con el contenido en Youtube generan **remuneraciones** a los “Youtubers” a través de múltiples vías (para más detalle: <https://support.google.com/youtube/answer/72857?hl=es>). Estas remuneraciones son obviamente, mayores cuanto mayor sea el número de usuarios, así como de interacciones con el contenido correspondiente.

Esto ha supuesto que Youtube haya logrado consolidarse como el líder en el sector de la creación de contenido audiovisual, así como de entretenimiento, pero no sólo eso, sino que resulta ser la **segunda página web más visitada del mundo** como podemos ver en el siguiente ranking:

Rank ①	Website ①	Category ①	Change ①	Avg. Visit Duration ①	Pages / Visit ①	Bounce Rate ①
1	 google.com	Computers Electronics and Technology > Search Engines	=	00:11:24	8.77	28.16%
2	 youtube.com	Arts & Entertainment > TV Movies and Streaming	=	00:21:48	12.61	19.63%
3	 facebook.com	Computers Electronics and Technology > Social Networks and Online Communities	=	00:10:08	8.46	32.59%
4	 twitter.com	Computers Electronics and Technology > Social Networks and Online Communities	=	00:10:38	10.04	31.69%
5	 instagram.com	Computers Electronics and Technology > Social Networks and Online Communities	=	00:07:51	11.24	34.81%
6	 baidu.com	Computers Electronics and Technology > Search Engines	=	00:06:02	8.05	20.31%
7	 wikipedia.org	Reference Materials > Dictionaries and Encyclopedias	=	00:03:55	3.11	57.00%
8	 yandex.ru	Computers Electronics and Technology > Search Engines	-1	00:11:18	9.44	22.23%

<https://www.similarweb.com/top-websites/>

No es de extrañar pues, que debido al potencial que tiene, muchos usuarios busquen o vean en esta plataforma la **oportunidad de montar un negocio**, pero al mismo tiempo e inherente tanto a su masividad como a su dinamismo fruto de la sociedad frenética a la que nos enfrentamos, la **competencia es enorme** y el **contenido** demandado o de éxito es muy **cambiante/dinámico**, complejizando el éxito/viralización. Por consiguiente, resulta de gran interés conocer las tendencias más actuales que se encuentran en auge en una plataforma tan masiva e influyente como es Youtube, de cara a desarrollar contenido en torno a dichas tendencias o en línea con lo que los usuarios demandan, facilitando así la captación del público y aumentando las probabilidades de éxito.

Dependiendo de los objetivos de un “Youtuber”, le interesará generar un contenido u otro en función a su **estrategia**, por ejemplo:

- Estrategia basada en generar contenido que **fidelice** (nº suscriptores).
- Estrategia basada en conseguir una gran **viralización** (nº de visualizaciones por video).
- Estrategia basada en originar un gran **impacto** (nº de comentarios).
- Cualquier otro tipo de estrategia o combinación de estas.

Como **sitio web**, contamos con **YouTube Tendencias** <https://www.youtube.com/feed/trending?bp=6gQJRkVleHBsb3Jl>, no hemos querido hacer uso de la **API**, ya que consideramos interesante para esta práctica hacerlo a través de técnicas de **WebScraping**. En esta página, encontraremos ordenados por Tendencia un **set limitado de videos** de los que podremos rescatar información útil además de la que obtendremos accediendo a cada uno de ellos gracias a técnicas de WebScraping.

2. TÍTULO. DEFINIR UN TÍTULO QUE SEA DESCRIPTIVO PARA EL DATASET:

Contenido audiovisual en tendencia en YouTube: datos destacados.

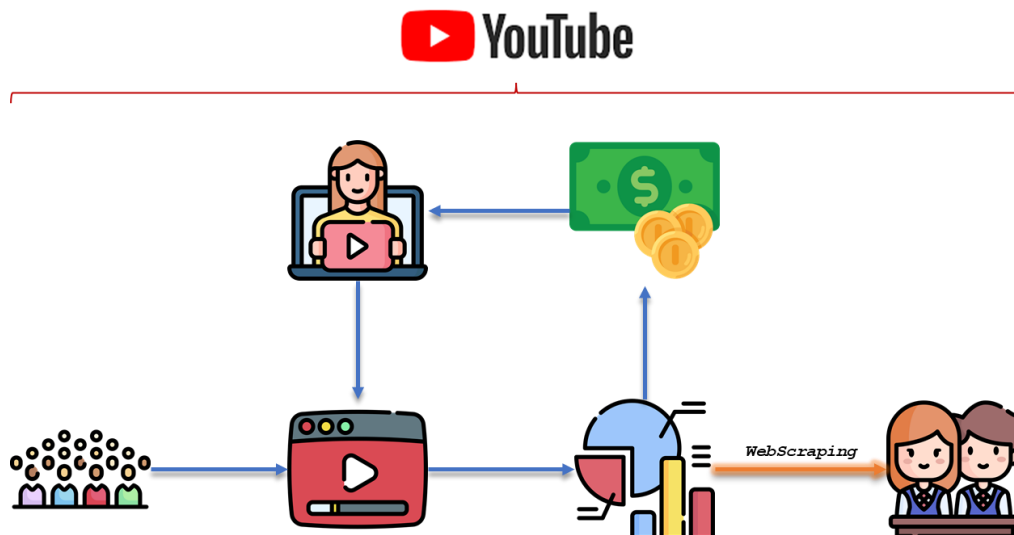
3. DESCRIPCIÓN DEL DATASET. DESARROLLAR UNA DESCRIPCIÓN BREVE DEL CONJUNTO DE DATOS QUE SE HA EXTRAÍDO. ES NECESARIO QUE ESTA DESCRIPCIÓN TENGA SENTIDO CON EL TÍTULO ELEGIDO.

Así como indica el título del dataset, este comprende una serie de campos con los datos más relevantes a la hora de evaluar un video que se encuentra en el momento del Scraping en tendencias. Las magnitudes de las columnas “numéricas” son por lo general cantidades (visualizaciones, suscriptores...) que, al no haber sido preprocesadas o transformadas, continúan siendo string a pesar de su naturaleza.

Un ejemplo de esto último es el número de “Me gusta” en un video, nos interesa tener un dato numérico entero (2, 5, 40...) pero vemos que, en el dato extraído, al número entero le sigue “Me gusta” teniendo un formato tipo string, no siendo el más idóneo.

Hemos extraído la información de 50 de los videos que se encuentran en Tendencias, un número cercano al límite de la web, al no tener más páginas u opciones de carga dinámica. El formato del conjunto de datos se almacenará en uno de los formatos interoperables más usados como es CSV.

4. REPRESENTACIÓN GRÁFICA. DIBUJAR UN ESQUEMA O DIAGRAMA QUE IDENTIFIQUE EL DATASET VISUALMENTE Y EL PROYECTO ELEGIDO.



5. CONTENIDO. EXPLICAR LOS CAMPOS QUE INCLUYE EL DATASET, EL PERIODO DE TIEMPO DE LOS DATOS Y CÓMO SE HAN RECOGIDO.

En nuestro dataset, contamos con una serie de campos que describiremos a continuación. Hay que comentar que, en cuanto al periodo, al ser tendencias es muy breve, siendo que en los primeros videos (tendencias más recientes) apenas estamos hablando de horas, mientras que en el resto hablamos de días.

Vamos con la descripción de las variables:

- i. **Url_Video:** dirección URL (*Uniform Resource Locator*) del video correspondiente.
- ii. **Title:** Título del video.

- iii. **Trending_Position:** Posición que ocupan en el Ranking de Tendencias en Youtube.
- iv. **Visualizations:** número de visualizaciones totales.
- v. **Date:** Fecha/momento de publicación del video.
- vi. **Likes:** Número de Likes/MeGusta del video.
- vii. **Subscribers:** Número de suscriptores al canal del video.
- viii. **Channel:** Canal (persona, organización...) del video.
- ix. **Url_Channel:** Dirección URL (*Uniform Resource Locator*) del canal correspondiente.
- x. **Comments:** Número de comentarios.
- xi. **Type:** Tipo/categoría del video.
- xii. **Length:** Duración del video en minutos y segundos.
- xiii. **#Tag:** Hastag con el que se etiqueta el video.

La recogida de datos se llevó a cabo a través de librerías y herramientas de **WebScraping** en el lenguaje **Python**. En primer lugar, se procedió a acceder al **Website de Tendencias** accediendo **secuencialmente** y de manera **ordenada** a cada video, una vez dentro de cada uno de estos, se extrajo y añadió la información relevante en un **DataFrame**. Finalmente, este set de datos lo almacenaríamos en un formato interoperativo como lo es **CSV**.

6. **AGRADECIMIENTOS.** PRESENTAR AL PROPIETARIO DEL CONJUNTO DE DATOS. ES NECESARIO INCLUIR CITAS DE ANÁLISIS ANTERIORES O, EN CASO DE NO HABERLAS, JUSTIFICAR ESTA BÚSQUEDA CON ANÁLISIS SIMILARES. JUSTIFICAR QUÉ PASOS SE HAN SEGUIDO PARA ACTUAR DE ACUERDO A LOS PRINCIPIOS ÉTICOS Y LEGALES EN EL CONTEXTO DEL PROYECTO.

YouTube es un sitio web perteneciente a una empresa privada dentro de la Industria de Internet, dedicado principal y exclusivamente a la subida de cualquier tipo de contenido audiovisual, dentro de los marcos de la ley. Todos los videos a los que podemos acceder, donde incluiríamos los videos en tendencia, han sido subidos a la nube por diversos canales de manera pública por lo que, cualquier persona esté o no registrada en la página, puede acceder a verlos de manera gratuita. Está la opción de privatizar los videos, lo que haría que sólo el propietario del canal, o aquellos usuarios que tuvieran acceso a la URL del vídeo, podrían ver el contenido.

Además, YouTube nos ofrece algunas estadísticas del vídeo, así como el número de visitas, las reacciones positivas que ha tenido o incluso el número de comentarios que ha recibido, lo que nos ha facilitado en gran medida la recogida de datos de popularidad de los videos de tendencia.

No obstante, es importante comentar que, estos canales ofrecen su contenido de manera pública a cambio de asumir la totalidad de los derechos de autor, editor o concesión del contenido. Queremos aclarar que, en nuestro caso, no se han vulnerado estos derechos ya que solo se ha accedido a la información que la plataforma YouTube nos ofrece de estos videos, sin violar los derechos de copyright de los autores del contenido.

Es por esto que nuestros agradecimientos no se centran en un organismo en particular si no que, queremos agradecer a todos esos canales que han publicado sus videos de manera pública y a YouTube por ofrecernos la posibilidad de conocer algunas de sus estadísticas sobre sus videos, haciendo posible nuestra recogida de información.

Existen diversos análisis que ofrecen información a cerca de la popularidad de los videos en Youtube, como por ejemplo:

- <https://mediakix.com/blog/most-popular-youtube-videos/>
- <https://www.socialmediatoday.com/news/new-study-looks-at-the-most-popular-youtube-content-and-highlights-key-tre/559556/>

Estos artículos nos ofrecen una explicación con valores y estadísticas de los motivos de popularidad en los videos, basados en el tipo de contenido y los canales más populares. No obstante, no explican paso a paso la metodología de obtención de estos valores y no es posible utilizar esa información para realizar un estudio de popularidad en tendencias.

7. **INSPIRACIÓN.** EXPLICAR POR QUÉ ES INTERESANTE ESTE CONJUNTO DE DATOS Y QUÉ PREGUNTAS SE PRETENDEN RESPONDER. ES NECESARIO COMPARAR CON LOS ANÁLISIS ANTERIORES PRESENTADOS EN EL APARTADO 6.

Como ya hemos comentado anteriormente, la magnitud y por consiguiente potencial de YouTube hace que multitud de individuos busquen o vean en esta plataforma la oportunidad de montar un negocio. Es por esto por lo que es verdaderamente relevante el ser consciente de cuáles y de qué tipo son las tendencias más contemporáneas en un mundo de constante y frenético cambio. Tener conocimiento acerca de las métricas que determinan el éxito de un video en tendencias, así como sus características, posibilita y aumenta la probabilidad de que un usuario sea capaz de entender y desarrollar el contenido que la sociedad demanda.

A lo que pretendemos dar respuesta, es a qué hace que un video sea tendencia (o dicho en otras palabras resulte atractivo o de éxito), y cuáles son sus características, además de su naturaleza.

Consecuencia de esto, permitirá al usuario que realice el análisis configurar su estrategia según se adapte a sus objetivos, tal como mencionamos en el apartado 1.

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

Repasando los objetivos de este proyecto y teniendo en cuenta el origen de la información recopilada, hemos considerado conveniente ajustar la licencia para la publicación de los datos al tipo **CC BY-SA 4.0 LICENSE**. Hemos decidido ajustarnos a esta licencia ya que los derechos y restricciones nos han parecido los más adecuados:

DERECHOS:

- ☐ Copiar y compartir el material en cualquier formato
- ☐ Adaptar el material para cualquier propósito, incluso comercial.

RESTRICCIONES:

- ☐ Se debe otorgar la debida atribución al autor y conservar el aviso de licencia.

9. **CÓDIGO.** ADJUNTAR EL CÓDIGO CON EL QUE SE HA GENERADO EL DATASET, PREFERIBLEMENTE EN PYTHON O, ALTERNATIVAMENTE, EN R.

El código se ha realizado con Python. Para la realización del Web Scraping se ha optado por utilizar la librería Selenium. A continuación, facilitamos el enlace al proyecto dentro de Git, donde se puede encontrar el código a ejecutar "WebScraping_YouTube.py":

- ☐ Enlace al Git del proyecto: https://github.com/ncalderonm/Proyecto_PRA1

Para ejecutar el programa, es necesario seguir los siguientes pasos:

- ☐ \$ git clone https://github.com/ncalderonm/Proyecto_PRA1.git

- \$ cd Proyecto PRA1
- \$ pip install -r requirements.txt
- \$ python WebScraping_Youtube.py

Los resultados del WebScraping se almacenarán en un csv dentro de la carpeta “Dataset”, que también podemos observar en el enlace a Git.

Además, dejamos a continuación algunos de los resultados obtenidos en el análisis sobre la evaluación inicial de la web.

FASE PREVIA / EVALUACIÓN INICIAL

Antes de realizar cualquier tipo de Scraping, partiremos de un análisis inicial que nos ayudará a contextualizar el proyecto:

- I. **Archivo robots.txt:** fichero en el que podremos encontrar las restricciones del sitio web, muy útil para evitar, o al menos reducir, la probabilidad de ser bloqueados. Vemos por un lado a qué robots da permisos “Mediapartners-Google*” así como en la parte inferior a qué directorios excluye:

```
# robots.txt file for YouTube
# Created in the distant future (the year 2000) after
# the robotic uprising of the mid 90's which wiped out all humans.

User-agent: Mediapartners-Google*
Disallow:

User-agent: *
Disallow: /channel/*/community
Disallow: /comment
Disallow: /get_video
Disallow: /get_video_info
Disallow: /get_midroll_info
Disallow: /live_chat
Disallow: /login
Disallow: /results
Disallow: /signup
Disallow: /t/terms
Disallow: /timedtext_video
Disallow: /user/*/community
Disallow: /verify_age
Disallow: /watch_ajax
Disallow: /watch_fragments_ajax
Disallow: /watch_popup
Disallow: /watch_queue_ajax

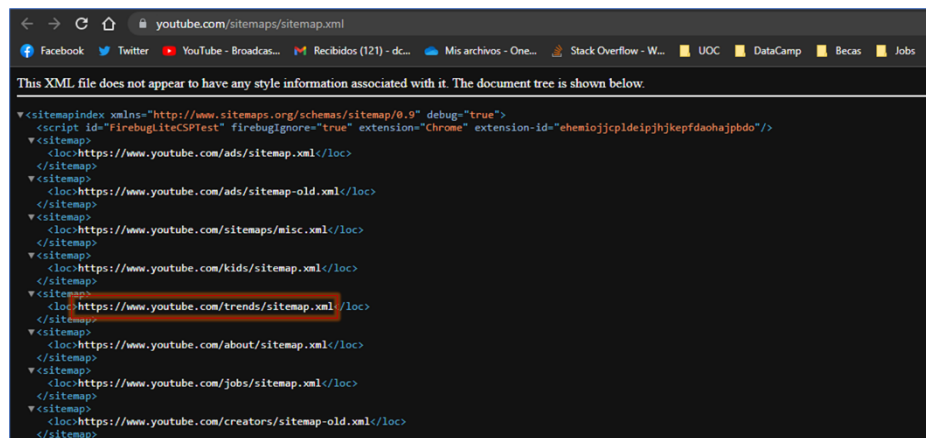
Sitemap: https://www.youtube.com/sitemaps/sitemap.xml
Sitemap: https://www.youtube.com/product/sitemap.xml
```

- II. **Sitemap (Mapa del sitio web):** presenta el mapa del sitio, es decir, la estructura del mismo pudiéndose observar a través de diferentes métodos (hemos descargado pero no usaremos “Firebug Lite”). Facilitan el rastreo web y permiten a los usuarios navegar de manera más cómoda por el sitio.

Como podemos ver, el fichero robots.txt nos ha mostrado las direcciones del sitemap:

```
Sitemap: https://www.youtube.com/sitemaps/sitemap.xml
Sitemap: https://www.youtube.com/product/sitemap.xml
```

Si accedemos al primer enlace, <https://www.youtube.com/trends/sitemap.xml> podemos observar el **sitemap** de Tendencias que se encuentra en formato **XML**:



De cara a inspeccionar este **sitemap**, lo haremos a través del siguiente código:

```
sitemap = "https://www.youtube.com/trends/sitemap.xml"

response = requests.get(URL)

with open('./sitemap.xml', 'wb') as file:

    file.write(response.content)

# Opción 1:

with open('./sitemap.xml', 'rb') as file:

    xml = file.read()

    xml = BeautifulSoup(xml, "html.parser")

    xml_p = xml.prettify()

    print(xml_p)

# Opción 2:

xml = BeautifulSoup(open(xml_sm, encoding="utf8"), "lxml")

xml_content = xml.prettify()

print(xml_content)
```

Que nos dará un **output** como el siguiente observando dicha estructura:

```
<!DOCTYPE html>

<html lang="es-ES" style="font-size: 10px;font-family: Roboto, Arial, sans-serif;" system-icons="" typography="" typography-spacing="">

<head>

<meta content="IE=edge" http-equiv="X-UA-Compatible"/>

<script nonce="iLqQ3XR/pOCDAC4WSe2WeQ">

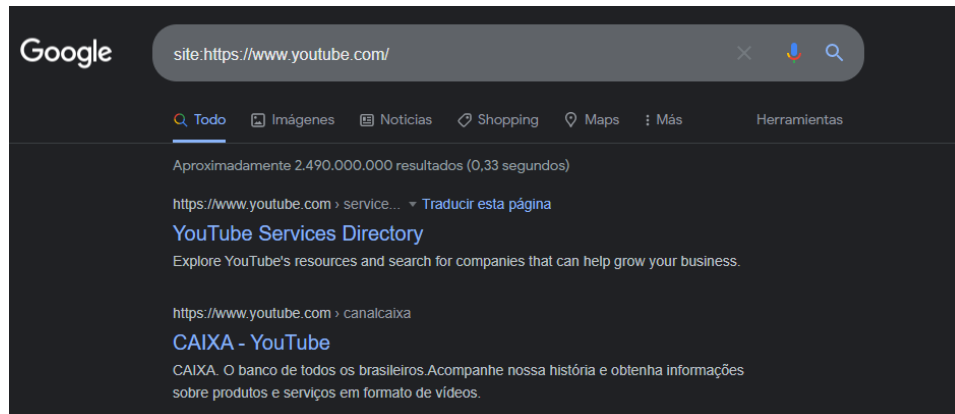
    var ytcfg={d:function(){return window.yt&&yt.config_||ytcfg.data_||(ytcfg.data_={})},get:function(k,o){return k in ytcfg.d()?ytcfg.d()[k]:o},set:function(){var a=arguments;if(a.length>1)ytcfg.d()[a[0]]=a[1];else for(var k in a[0])ytcfg.d()[k]=a[0][k]}};

window.ytcfg.set('EMERGENCY_BASE_URL', '\\error_204?t\x3djserver\x26level\x3dERROR\x26client.name\x3d1\x26client.version\x3d2.20220406.09.00');
```

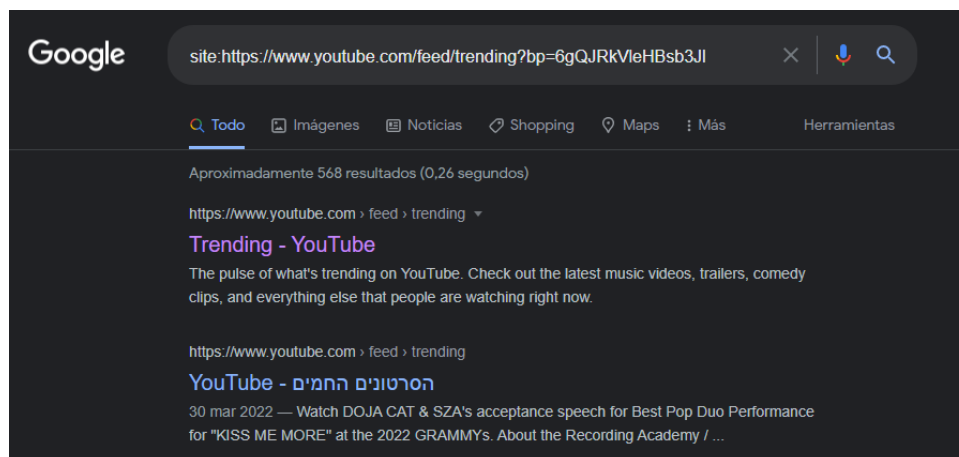
```
</script>
```

```
<script nonce="iLqQ3XR/pOC
```

- III. **Tamaño:** debemos de tener en cuenta el número de páginas a la hora de realizar las descargas. A través del comando `site` en el buscador de Google, vemos que para Youtube hay casi 2,5kM de resultados, un número muy considerable:



Por otro lado, si nos centrásemos en las tendencias, vemos que el número se reduce drásticamente, siendo mucho más manejable:



- IV. **Tecnología:** según el tipo de tecnología usada, se usará un tipo de WebScraping u otro. Utilizaremos el comando `"builtwith.builtwith("URL")` para extraer dicha información (indistintamente de URL de URL_principal):

```
{'font-scripts': ['Google Font API'], 'javascript-frameworks': ['RequireJS']}
```

Vemos que el **javascript-framework** es **"RequireJS"**, que al mismo tiempo es compatible con la versión de Chrome que manejamos:

RequireJS is a JavaScript file and module loader. It is optimized for in-browser use, but it can be used in other JavaScript environments, like Rhino and Node. Using a modular script loader like RequireJS will improve the speed and quality of your code.

IE 6+ compatible ✓
Firefox 2+ compatible ✓
Safari 3.2+ compatible ✓
Chrome 3+ compatible ✓
Opera 10+ compatible ✓

<https://requirejs.org/>

v. **Propietario:** vamos a conocer al propietario de la página, que puede resultarnos de interés. Usaremos:

```
import whois  
  
print(whois.whois('URL'))
```

Obtenemos (indistintamente de URL de URL_principal):

```
{  
  "domain_name": [  
    "YOUTUBE.COM",  
    "youtube.com"  
  ],  
  "registrar": "MarkMonitor, Inc.",  
  "whois_server": "whois.markmonitor.com",  
  "referral_url": null,  
  "updated_date": "2022-01-14 09:38:42",  
  "creation_date": "2005-02-15 05:13:12",  
  "expiration_date": [  
    "2023-02-15 05:13:12",  
    "2023-02-15 00:00:00"  
  ],  
  "name_servers": [  
    "NS1.GOOGLE.COM",  
    "NS2.GOOGLE.COM",  
    "NS3.GOOGLE.COM",  
    "NS4.GOOGLE.COM",  
    "ns2.google.com",  
    "ns3.google.com",  
    "ns1.google.com",  
    "ns4.google.com"  
  ],  
  "status": [  
    "clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited",  
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",  
    "clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited",  
    "serverDeleteProhibited https://icann.org/epp#serverDeleteProhibited",  
    "serverTransferProhibited https://icann.org/epp#serverTransferProhibited",  
    "serverUpdateProhibited https://icann.org/epp#serverUpdateProhibited",  
    "clientUpdateProhibited (https://www.icann.org/epp#clientUpdateProhibited)",  
    "clientTransferProhibited (https://www.icann.org/epp#clientTransferProhibited)",  
    "clientDeleteProhibited (https://www.icann.org/epp#clientDeleteProhibited)",  
    "serverUpdateProhibited (https://www.icann.org/epp#serverUpdateProhibited)",  
    "serverTransferProhibited (https://www.icann.org/epp#serverTransferProhibited)",  
    "serverDeleteProhibited (https://www.icann.org/epp#serverDeleteProhibited)"  
  ],  
  "emails": [  
    "abusecomplaints@markmonitor.com",  
    "whoisrequest@markmonitor.com"  
  ],  
  "dnssec": "unsigned",  
  "name": null,  
  "org": "Google LLC",  
  "address": null,  
  "city": null,  
  "state": "CA",  
  "zipcode": null,  
  "country": "US"  
}
```

Como se puede observar, el principal y único propietario de la página es la empresa YouTube.

TABLA DE CONTRIBUCIONES AL TRABAJO:

Contribuciones	Firma
Investigación previa	DCM, NCM
Redacción de las respuestas	DCM, NCM
Desarrollo del código	DCM, NCM