



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Escola Tècnica
Superior d'Enginyeria
Informàtica

SOCIAL WEB BEHAVIOUR

PRÁCTICA 1. ANÁLISIS DE HISTORIAL WEB

SOCIAL WEB BEHAVIOUR

CRISTINA I. FONT
DCADHA
crifonju@upv.es



1. Introducción

Navegar por internet es algo muy común entre la sociedad desde hace años. Los números indican que cada día millones de conexiones se realizan en todo el mundo y que estas se encuentran en constante crecimiento.

Cada una de las personas que usan la red, independiente del motivo por el que lo hagan (comunicarse, informarse o realizar transacciones) lo hacen siguiendo sus propios patrones de comportamiento, del mismo modo que sucede en la vida real. Por lo tanto, cada uno de nosotros dejamos una huella digital, clara y definible, en la navegación que realizamos.

Y esta navegación se encuentra recopilada y guardada en los equipos que se usan para realizarla o en los servidores de las empresas que recopilan los datos. El análisis del conjunto de los ficheros asociados (logs, chats, agendas, emails, información de aplicaciones, historial de navegación, archivos caché, localizaciones, etc.) con el uso de internet, junto con la información recogida por terminales móviles, permite entender mejor el comportamiento de los usuarios. Este análisis es de gran utilidad, no solo para luchar contra delitos o crímenes cometidos online, ya que permite identificar el comportamiento base de las actividades normales y específicas de los usuarios, sino que también permite ofrecer contenido más relevante para el usuario, focalizando los esfuerzos de las empresas y ofreciendo un mayor margen de beneficio.

La información de navegación se transmite de diversas formas, una de las principales fuentes de información para los servicios de recopilación de datos es precisamente el sistema de protocolo de comunicación que se utiliza para navegar por la Web. Los navegadores, por ejemplo, ofrecen diferentes datos que pueden ser de utilidad para la segmentación de los usuarios. Existen páginas que nos permiten tener conocimiento de cuánta información tienen los navegadores, por ejemplo, *What every Browser knows about you* (<http://webkay.robinlinus.com/>), en la que se tiene una explicación sobre qué se puede saber y porqué. Además, existen otras como *Clickclickclick* (<https://clickclickclick.click>) que permiten saber la extensión de la información que ofrecemos, únicamente moviendo el ratón en una página.

Esta información, es solo aquella que localiza y sitúa a un usuario, pero no permite conocer exactamente gustos y aficiones. Para eso es necesario recopilar los datos vía otros medios. Un ejemplo claro de empresa que recopila datos sobre los usuarios es Google. La empresa utiliza estos datos para mejorar sus productos y ofrecer mejores servicios personalizados a sus usuarios. Para ello recopila datos relativos a:

- El contenido buscado
- Los sitios web visitados
- Los videos visualizados
- Anuncios visitados
- Ubicación del usuario
- Información del dispositivo
- Aplicaciones utilizadas (dispositivos móviles)
- Dirección IP y las cookies

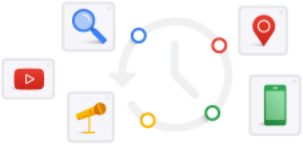
Además, al utilizar una cuenta personal de Google, estos tienen acceso a información sensible como correos electrónicos, agendas, eventos, contactos, fotos y videos personales y documentos almacenados en sus sistemas.













Por lo tanto, Google es capaz de almacenar una gran cantidad de información sobre cada uno de sus usuarios, para después utilizarla individual y colectivamente. Esta información se encuentra tanto en ficheros físicos en los dispositivos desde los que se navega, como en los propios servidores de la empresa. Por lo que puede ser accedida vía online. En la página web de actividad de la cuenta personal (<https://myactivity.google.com/myactivity>) se puede observar la actividad de búsquedas y visitas de la cuenta de usuario mediante diversas visualizaciones.

Por otra parte, desde la cuenta personal de Google, en la pestaña “Datos y personalización” (<https://myaccount.google.com/data-and-personalization>) se puede acceder a toda la información recopilada.

Controles de la actividad de tu cuenta

Guarda tu actividad si quieres disfrutar de una experiencia más personalizada en Google. Puedes activar o pausar estos ajustes cuando quieras.



	Actividad en la Web y en Aplicaciones	 Activado	>
	Historial de ubicaciones	 Activado	>
	Actividad de Voz y Audio	 Activado	>
	Información de tus dispositivos	 Activado	>
	Historial de búsquedas de YouTube	 Detenido	>
	Historial de reproducciones de YouTube	 Detenido	>

[Gestionar los controles de la actividad de tu cuenta](#)

Esta información puede ser modificada en cualquier momento por parte del dueño de la cuenta, del mismo modo, es posible descargar la información que resulte de interés o quiera analizarse. Para descargarla será necesario seleccionar el producto Google del que se quiere la información, indicar la cantidad o el formato de descarga en las opciones en las que exista la posibilidad, y generar el fichero de descarga.

Una vez creado el fichero, el usuario recibirá un correo indicando que el fichero está preparado y puede acceder a descargarlo en un enlace temporal, ya que una vez pasados unos días este será eliminado.

← Descargar tus datos

Tu cuenta, tus datos.
Exporta una copia.

Crea un archivo con tus datos de los productos de Google

[ADMINISTRAR ARCHIVOS](#)



Seleccionar los datos que incluir

Selecciona los productos de Google que quieras incluir en tu archivo y elige la configuración de cada producto. Solo tú podrás acceder a este archivo. [Más información](#)

2. Objetivos de la práctica

- Obtener una visión global de la información relativa a los usuarios existente en internet.
- Analizar y entender la información de usuario accesible en la web por empresas y terceros.
- Conocer y tener una visión crítica de lo que supone el intercambio de información sobre el comportamiento de los usuarios en la web.

3. Desarrollo de la práctica

Para esta práctica únicamente serán necesarios los datos de navegación. Por lo que una vez se accede a la página de Descarga de Datos, se puede pulsar el botón “No seleccionar ninguno” para desmarcar todos los productos. A continuación, se busca Chrome y se selecciona la opción “Seleccionar datos de Chrome” para desmarcar todo y únicamente seleccionar la casilla “BrowserHistory”.



Chrome

Todos los tipos de datos
de Chrome



Datos de Chrome

☐ Incluir todos los datos de Chrome

☒ Seleccionar datos de Chrome

DATOS DE CHROME

Se ha seleccionado 1 tipo.

De este modo, se descargará únicamente el historial de navegación Web del usuario. Para lograr un mejor análisis, lo ideal es descargar la información relativa a las conexiones, ubicación y texto de búsqueda, ya que la combinación de todos esos datos, ofrece una imagen clara del perfil del usuario y su comportamiento en la web.

Una vez descargado el fichero con el historial, este se encontrará en formato JSON. Para trabajar con este de forma sencilla, lo más cómodo es pasarlo a CSV o XLS, para ello existen diversas formas:

- Utilizar un programa online de conversión de datos. Por ejemplo, <https://jsonformatter.org/json-to-csv>. Dependiendo del peso del fichero, la carga del archivo se puede demorar bastante.
- Utilizar un programa para realizar la conversión. Por ejemplo, *OpenRefine* (<http://openrefine.org/>) es un programa de código abierto, rápido y sencillo. (Ver anexo para instrucciones)
- O usar scripts en otros lenguajes. Por ejemplo, en Python:

```
import json
import csv

f = open('data.json')
data = json.load(f)
f.close()

f = open('data.csv')
csv_file = csv.writer(f)
for item in data:
    csv_file.writerow(item)

f.close()
```

Una vez el fichero se encuentra convertido, se puede abrir con herramientas como Excel o herramientas de código libre similares (*OpenOffice*, *LibreOffice*, la propia herramienta de hojas de cálculo de Google, etc.) para ver el contenido del fichero.

En él encontramos las columnas:

- URL: dirección visitada
- Título: título de la pagina
- ID Cliente: identificador del usuario
- Time usec: fecha de la visita
- Page Transition: cómo se llega a la visita
- Favicon URL

Las más interesantes para la práctica son las marcadas en amarillo ya que permiten realizar análisis preliminares.

El siguiente paso para poder trabajar con el fichero es convertir la fecha a un formato legible. Para eso utilizaremos las fórmulas de Excel. En caso de ser necesario incluiremos una columna vacía a la derecha de la columna de **Time_usec**. En esta, pegaremos la fórmula:

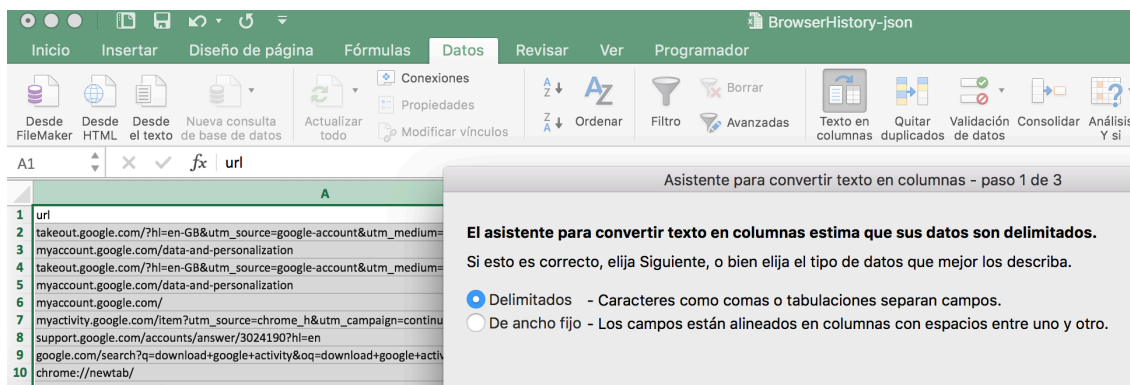
=LEFT(B2;10)/(60*60*24)+"1/1/1970"

Cambiando “LEFT” por IZQUIERDA en caso de que las instrucciones del programa se encuentren en castellano. Y modificando la celda a la que hace referencia.

En caso de que la fecha no aparezca correctamente, deberemos indicar desde el menú superior que la casilla es una fecha y no texto general. Una vez completado se debe arrastrar para que se transforme toda la columna.

El siguiente paso es el de limpiar las URLs para poder contabilizar la cantidad de visitas que se realizan a un mismo portal (y de paso anonimizar la navegación). Para ello, lo más fácil es copiar la columna URL, crear una nueva hoja de Excel y pegar la selección anterior. Una vez pegadas se utiliza la opción “Reemplazar” para eliminar todo aquello que no esté relacionado con el dominio de la parte inicial:

- https://
- http://
- www.

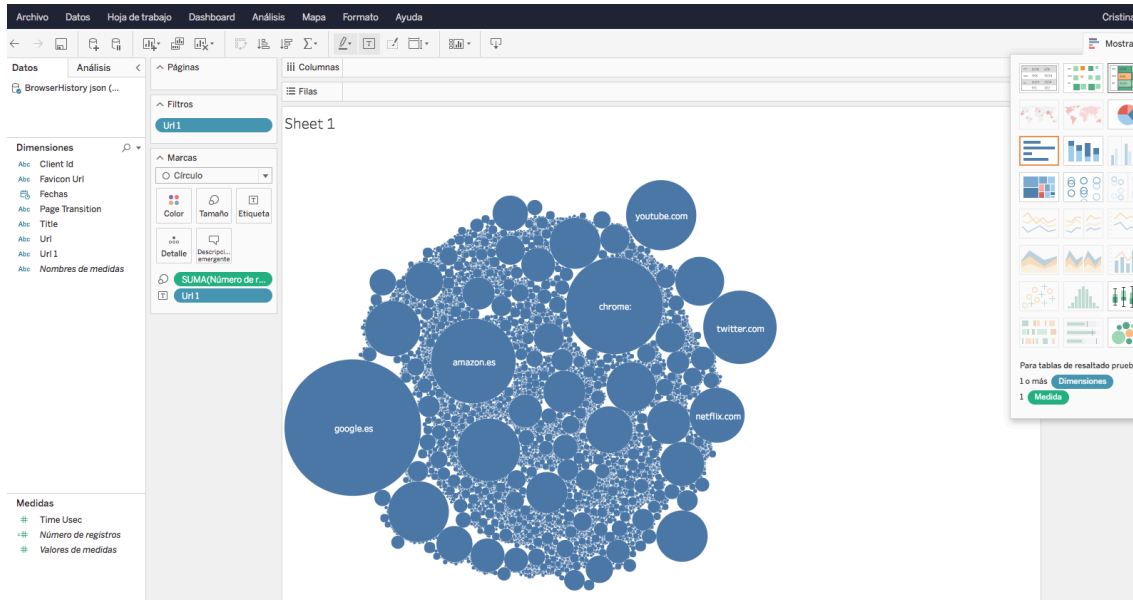


Tras eso, se cambia al menú superior “Datos”, desde el que se utilizará la función “Texto en Columnas”, para indicar que se desea separar los datos “Delimitados”. En el Paso 2, se seleccionará “Otro” y se incluirá el símbolo “/” para dejar en la primera columna únicamente los dominios. Deberemos copiar la columna y pegarla en una columna nueva en la hoja anterior junto con el resto de datos. La nueva hoja creada para limpiar los dominios se puede eliminar.

Por último, ya solo queda guardar el fichero que se utilizará en la herramienta de visualización. En este caso se utilizará la versión online de *Tableau* (<https://online.tableau.com/>). Una vez se accede a la aplicación, se debe crear un “Nuevo libro de trabajo” en el que se subirá el documento de Excel.



Cuando la aplicación cargue completamente el fichero, aparecerá el *Dashboard* de trabajo, desde el que se podrán realizar diversas combinaciones para tratar de analizar los datos, encontrar patrones, intereses, etc.



4. Preguntas y entrega de la práctica

La práctica deberá ser entregada en Tareas de *Poliformat* mediante el envío de **un único documento** con una o más capturas realizadas con la herramienta *Tableau*, con una explicación crítica del contenido de esta, la información que se puede extraer de las mismas, posibles usos, etc.

Es interesante utilizar el resto de columnas contenidas en el Excel o añadiendo información recogida de otras fuentes para extraer más información.

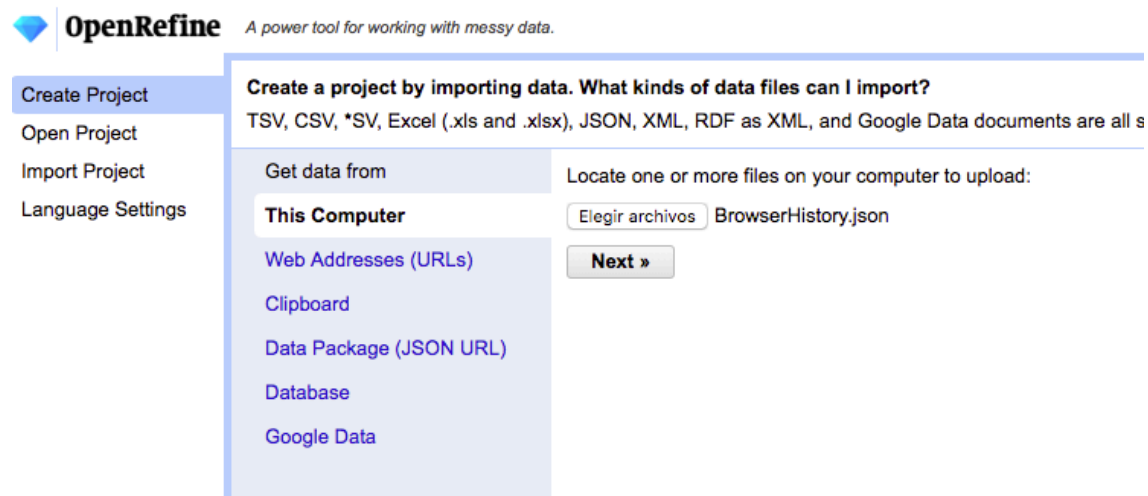
Fecha máxima de entrega: 23 de febrero 16:30h.

Además, se deberá contestar a las siguientes preguntas e incluirlas en el mismo documento:

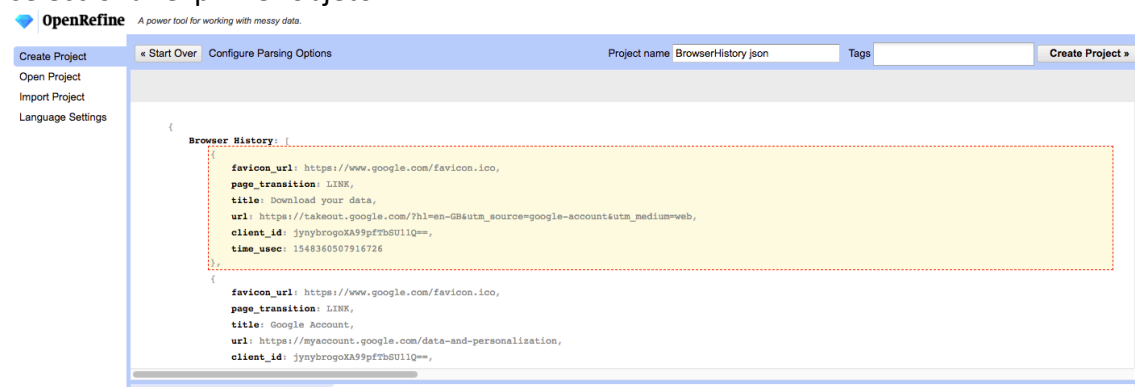
- ¿Pueden considerarse los historiales y los perfiles web únicos para un usuario?
- ¿Son estables los historiales web y los perfiles de interés? En otras palabras, ¿pueden considerarse buenas huellas digitales de comportamiento y pueden los sitios web de seguimiento confiar en estos datos?
- ¿Qué implicaciones para la seguridad y privacidad crees que tiene el acceso a estos datos por parte de empresas o terceros?
- Si los datos fueran accedidos por una empresa, ¿qué datos de interés podrían sacar de tu información y qué valor tendrían?

Anexo I: Herramienta OpenRefine

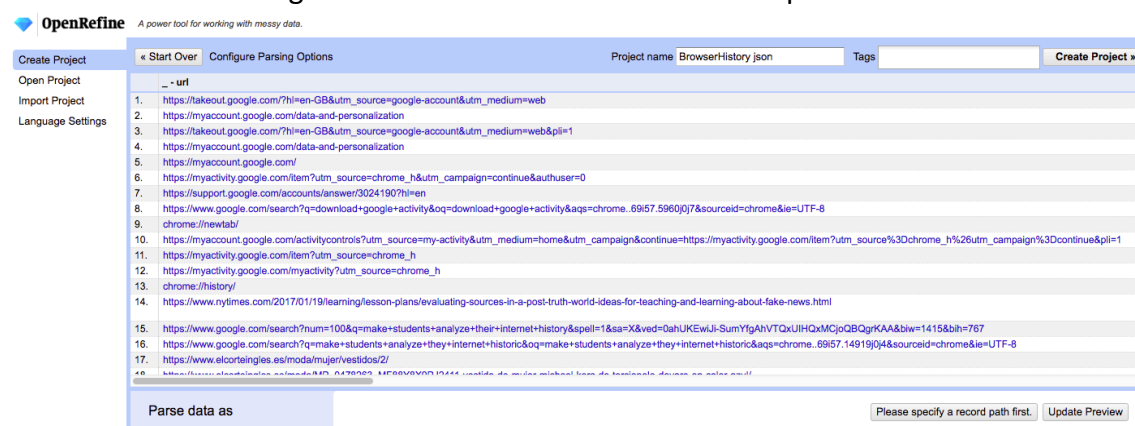
Para usar la herramienta, basta con seleccionar el fichero que se desea leer:



Se selecciona con el selector amarillo la información, normalmente basta con seleccionar el primer objeto:



Automáticamente se generará una visualización del fichero parseado:



Para salvar el documento basta con pulsar “Create Project”, se generará el documento con todas las filas y entonces se podrá pulsar “Export” para guardar en el formato deseado.