TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRA 1 – Web Scraping

Alumno: Luis Manuel Molina Coca (luimoco)

Análisis Exodus User-Agent

Análisis del archivo robots.txt

No se ha detectado la presencia del archivo robots.txt en el sitio web. Tampoco hay una política concreta sobre las restricciones de utilización de user-agents sobre la página.

No obstante, el sitio se encuentra indexado por el user-agent de google puesto que aparecen páginas concretas mediante la búsqueda avanzada con la condición *site:exodus-privacy.eu.org*.

Examinar el mapa del sitio web

El sitio web no cuenta con un recurso *sitemap* oficial donde observar la estructura. En la página principal no se encuentran etiquetas *<urlset>*
Tampoco existe una página dedicada a explicar la estructura del sitio por lo que realizamos una exploración y un mapa a nivel particular.

- Página principal https://exodus-privacy.eu.org/en/:
  Logotipo, descripción y accesos a qué hacen y al negocio concreto, analizar una app.
  - What https://exodus-privacy.eu.org/en/page/what/
    Explicación del negocio de Exodus
    Ayuda sobre la interpretación de los informes de las aplicaciones
    Publicidad de su propia aplicación Android
    Más detalles y redes sociales.
  - Who https://exodus-privacy.eu.org/en/page/who/
    Componentes de Exodus, creadores y colaboradores, código de conducta y contacto.
  - Contribute https://exodus-privacy.eu.org/en/page/contribute/
    Métodos de colaboración y financiación de Exodus
  - Press https://exodus-privacy.eu.org/en/page/press/
    Notas de prensa en inglés y francés.
  - Events https://exodus-privacy.eu.org/en/page/events/
    Calendario de eventos actuales y pasados.
  - Blog https://exodus-privacy.eu.org/en/post/
    Publicaciones del equipo de exodus.

- FAW https://exodus-privacy.eu.org/en/page/faq/
  Sección de preguntas frecuentes de seguridad, privacidad y de Exodus.

- Informes https://reports.exodus-privacy.eu.org/es/reports/######/
  - Más de 150.000 páginas con informes de aplicaciones móviles del ecosistema Android.
  - Cada nuevo informe tiene un identificador secuencial y se encuentra en la ruta anterior sustituyendo ###### por este identificador.
    Por ejemplo, https://reports.exodus-privacy.eu.org/es/reports/152144/ es el informe 152.144 que se dedica a la aplicación Smartha en su versión 1.44.28
  - Pueden existir varios informes de la misma aplicación, cada uno con diferente identificador, correspondientes a distintas versiones de la misma.
  - Un informe de aplicación lo compone:
    - Icono de la aplicación.
    - Nombre de la aplicación.
    - Enlace a la información de la aplicación en un repositorio de aplicaciones. (Google Play, F-droid)
    - Información agregada con la cuenta del número total de rastreadores que contiene y los permisos que solicita para su instalación.
    - Versión de la aplicación y enlace a un listado con los distintos informes realizados a versiones recientes o anteriores de la aplicación.
    - Información detallada de rastreadores.
      - Nombre y lista de propósitos.
    - Información detallada de permisos.
      - Nombre, descripción y advertencia de peligrosidad.
    - Resolución de dudas del informe.
    - Datos técnicos de la aplicación:
      - Huella digital.
      - Emisor:
        - Nombre.
        - Organización.
        - Localidad.
        - Estado ó provincia.
        - País.
      - Asunto, con los mismos campos que Emisor.
      - Número de serie.
      - Huella digital del instalable:
        - Hash de comprobación.
        - URI del Gestor.
        - Nombre de la aplicación.
        - UAID.
        - Versión.
        - Código de la versión.
        - Hash del icono.

Tamaño del sitio web

Una búsqueda avanzada en Google con la condición *site:exodus-privacy.eu.org*, arroja un total de 13.300 páginas indexadas.
No obstante, el número de páginas es mucho mayor puesto que el sitio web genera una página por cada informe con lo que se eleva a más de 150.000 páginas disponibles.

Tecnología

El sitio web está realizado utilizando la tecnología *Hugo* y su código fuente está publicado en GitHub. https://github.com/Exodus-Privacy/website
No obstante, los datos no se encuentran en este repositorio.
Tecnología empleada en la sección de informes de la web obtenida con la librería builtwith de Python:

```
{'web-servers': ['Nginx'],
 'web-frameworks': ['Django', 'Twitter Bootstrap'],
 'programming-languages': ['Python'],
 'javascript-frameworks': ['Handlebars', 'jQuery']}
```

Propietario

No ha sido posible encontrar información del propietario a través del dominio ni con la librería *python-whois* ni mediante un servicio web.
No obstante, existe información en la web en su sección Who: https://exodus-privacy.eu.org/en/page/who/

Atributos del dataset

- **Id**: Identificador de la aplicación/versión en el dataset.
- **Name**: Nombre de la aplicación.
- **Tracker_Count**: Agregado con el total de rastreadores.
- **Permissions_Count**: Agregado con el total de permisos.
- **Version**: Versión de la aplicación.
- **Downloads**: Cantidad aproximada de descargas de la aplicación.
- **Analysis_Date**: Fecha de inspección de la información de privacidad de la aplicación.
- **Trackers**: Lista de nombres de rastreadores detectados en la aplicación y sublista de sus propósitos.
- **Permissions**: Lista de permisos utilizados por la aplicación.
- **Permissions_Warning_Count**: Agregado con el total de permisos sensibles de entre los que solicita la aplicación.
- **Country**: Código del país del creador de la aplicación.
- **Developer**: Nombre del desarrollador de la aplicación.

- **Icono**: Se tratará de un atributo opcional, puesto que su inclusión en el fichero lo hace crecer mucho de tamaño. Se podrá rastrear incluyendo la lista de componentes RGBA de la imagen o extrayendo la imagen a fichero con un nombre que la permite ser asociada a los registros del dataset.

Formato del dataset

A partir de la estructura de la información disponible, se elige el fomato Json como formato de almacenamiento del dataset debido a que, tanto la información de rastreadores como la información de permisos son muy variables por aplicación/versión y una representación en listados de clave, valor sería la más adecuada. Para el resto de atributos, la información clave, valor también puede ser muy útil.

No obstante, para el tratamiento del dataset por modelos analíticos, sí sería necesario por parte del científico de datos, un tratamiento previo según el propósito para disponer de los distintos rastreadores y propósitos o el listado de permisos como variables “dummy”.

También es posible que aparezca información incompleta según el repositorio de aplicaciones del que se toma la aplicación para análisis. Incluso también hemos detectado la posibilidad de tener valores erróneos en los atributos de emisor al ser cumplimentados por personas físicas, jurídicas o una división territorial distinta en su lugar de residencia.

```
{
        “Id”: id,
        “Name”: “nombre”,
        “Tracker_count”: tracker_count,
        “Permissions_count”: permissions_count,
        “Version”: “version”,
        “Downloads”: “downloads”,
        “Analysis_date”: “analysis_date”,
        “Trackers”: [
                        {
                          “name” : “name”,
                           “Purpose”: [“purpose”]
                         }
                      ]
        “Permissions”: [“permissions”]
        “Permissions_warning_count: permissions_warning_count
        “Country”: “country”
        “Developer”: “developer”
        “Icon”: [ [RGBA] ]
}
```

Prevención del Web Scraping

- Con tal de evitar el bloqueo de IP:
    - Debido al gran volumen de informes existentes +150.000, se tratará de realizar una primera intervención con hilos de ejecución paralela de subconjuntos del total de datos existentes desde varias máquinas con distintas conexiones a Internet.
    - Cada hilo de ejecución va a disponer de mecanismos de control de tiempo para realizar retrasos o pausas proporcionales según el estado actual del rastreo: retrasos proporcionales y gestión de excepciones.
    - Tras la extracción inicial, el rastreador podrá seguir analizando nuevos informes con un espaciado de tiempo mucho mayor.
    - En las labores de desarrollo y pruebas, se extraerá inicialmente un conjunto de 100 y 1.500 páginas html sin procesar respectivamente para poder desarrollar y afinar el módulo de extracción de atributos.
- Es posible caer en trampas de araña puesto que la generación de informes es dinámica y el rastreador no puede determinar cuál es el último informe. Hay que establecer algún criterio de parada cuando la petición web devuelva errores 404 sucesivamente.
- No se ha detectado la presencia de Captchas
- No se ha detectado el uso de CSS Sprites o HTMLCSS.

Resolución de obstáculos

- Utilización de una cabecera ficticia que emule la intervención de un navegador web al no disponer públicamente de la política de utilización de user-agents.
- No es necesario estar logueado ni manejar cookies de sesión para el rastreo de esta web.
- No hemos podido encontrar el robots.txt para conocer y respetar las condiciones.
- Por respeto al tratarse de una organización autofinanciada sin ánimo de lucro de carácter público y abierto, para no saturar el servidor y prevenir bloqueos de IP, utilizaremos métodos para garantizar el espaciado de peticiones http.
- Se realizará el tratamiento oportuno de timeouts y excepciones para asegurar una correcta extracción y el espaciado de peticiones durante la intervención.
- El contenido se genera dinámicamente y no es posible disponer del dato del número de informes que contiene en tiempo de ejecución. El sitio se genera dinámicamente cada vez que se publica un informe, y se podría caer en una trampa de araña. Se aprovechará el hecho de que si el informe no existe se arroja una excepción 404, para marcar el final del proceso.