TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRA 1 – Web Scraping

Alumno: Luis Manuel Molina Coca (luimoco)

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

Queremos conseguir un dataset con información relevante sobre aspectos relativos a la privacidad del usuario a la hora de instalar aplicaciones en su dispositivo móvil.
En este contexto, la organización sin ánimo de lucro Exodus Privacy, se dedica a ayudar a la gente a comprender los problemas derivados del rastreo de aplicaciones Android mediante la elaboración de informes de seguridad, por parte de voluntarios, en el código público de las aplicaciones para listar rastreadores integrados y los permisos del dispositivo a los que accede tras la instalación en el dispositivo.

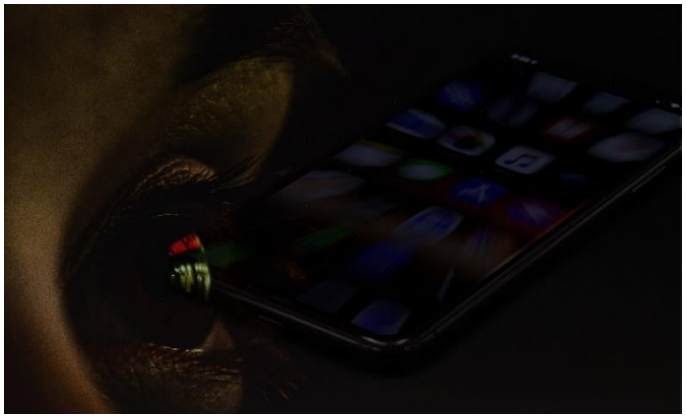
2 ***Definir un título para el dataset****. Elegir un título que sea descriptivo.*

Dataset del tratamiento de la privacidad del usuario por parte de las aplicaciones móviles.

3 ***Descripción del dataset****. Desarrollar una descripción breve del conjunto de datos que se han extraído. (Es necesario que esta descripción tenga sentido con el título elegido)*

El dataset *Dataset del tratamiento de la privacidad del usuario por parte de aplicaciones móviles*, proporciona información de rastreadores que se han incluido en la aplicación y los permisos del dispositivo que ha de aceptar el usuario en el momento de su instalación. Adicionalmente proporciona información propia de características de la aplicación interesantes para el tratamiento analítico del uso general de rastreadores y permisos.

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



COMPOSICIÓN REALIZADA A PARTIR DE IMÁGENES DE STOCK unplash.com DE JOSHUA-RODRIGUEZ Y MAULIK-SUTARIYA

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

El dataset cuenta con los siguientes atributos extraídos de cada página de informe de seguridad de las aplicaciones:

- **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 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.
- **[Icon]**: Opcional: Lista de componentes RGBA del icono de la aplicación en tamaño 32x32. Si se trata de la versión del dataset sin iconos éste se encuentra almacenado en un fichero de nombre Id.png

Cuando se sube al sitio web un nuevo informe de seguridad de aplicación, se genera una nueva web con un identificador secuencial que forma parte de la URL. Desde el momento de creación, el dato permanece en su web particular. Ej. https://reports.exodus-privacy.eu.org/es/reports/153373/
Los datos particulares del informe pueden ser actualizados, situación que se refleja de manera textual como fecha de actualización. El rastreador, recoge en el dataset se recoge la última versión sin actualizar datos en caso de diferir.
El dato, de una misma aplicación, no es sustituido ni eliminado (por ejemplo, con la publicación de una nueva versión de la app). Se genera otro informe diferente para la aplicación/versión con otro código donde se incorporan los nuevos datos de seguridad.

Se han utilizado técnicas de Web Scraping para recorrer el sitio web secuencialmente, y recoger para cada aplicación sus datos a partir de los patrones concretos detectados en el análisis del código html de 100 informes aleatorios.
En el documento de ANÁLISIS EXODUS USER-AGENT.PDF se presentan más detalles del proceso de creación del user-agent.
En el documento DISEÑO EXODUS USER-AGENT.PDF, en concreto en el módulo *Rastrear el html*, se profundiza el modo técnico de recolección de cada dato en particular.

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

Exodus Privacy es una organización francesa sin ánimo de lucro. Está dirigida por "hacktivistas" con el propósito de proteger la privacidad a nivel global.
Los miembros que componen la comunidad de Exodus, pueden aportar informes acerca de nuevas aplicaciones móviles para contribuir al conjunto de datos expuesto.
La organización se sustenta a base de trabajo colaborativo, recibiendo el apoyo de ocho organizaciones y se financia por donaciones de particulares. https://exodus-privacy.eu.org/en/page/who/

He contactado con Exodus para agradecer e interesarme por su política de web scraping puesto que no se encuentra el fichero *robots.txt*. Al no recibir contestación a tiempo, he optado por utilizar una cabecera estándar de navegador en lugar de identificar al user-agent y tiempos de cortesía entre peticiones para poder generar el dataset en plazo debido al gran volumen de información del sitio Web. No obstante, sigo esperando para conocer su parecer acerca de la publicación de la información y poder ofrecerla también al público como colaboración en su propia labor.

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

Como en toda revolución, la industria trata de sacar tajada del recurso que proporcione riqueza en cada momento. En esta era en la que el dato es tan valioso, han proliferado las herramientas para extraer la mayor cantidad de información posible de personas e instituciones mediante el uso de **rastreadores** de empresas analíticas y con un abuso poco justificado de los **permisos** necesarios para el funcionamiento de la aplicación con tal de recopilar toda la información posible.
En ocasiones esta información es utilizada de manera benévola para la mejora del servicio o interés de la ciudadanía, pero en la mayoría de las ocasiones se usa de manera **fraudulenta**, bien para la venta de información sin consentimiento ni beneficio del usuario o directamente secuestrando o dañando el dispositivo con fines de extorsión.
Es muy interesante, y es uno de los puntos de vista de la organización Exodus, que la población, en general gran productora de los datos, tome conciencia de este interés por su información y esté prevenida ante un uso fraudulento que ataque a su propio derecho de privacidad.
Con este dataset, será posible generar conocimiento sobre la evolución de este mercado de los datos que maximiza la extracción de información minusvalorando la privacidad del usuario.

Bajo un punto de vista **operacional**, podría ser utilizada para realizar rastreos en las aplicaciones instaladas por el usuario en un dispositivo propiedad de la empresa para minimizar riesgos de fuga de información, como cortafuegos para la instalación de aplicaciones potencialmente maliciosas, incrementar la funcionalidad en aplicaciones de control parental, etc.
Bajo un punto de vista **analítico**, el dataset tiene una utilidad destacada en tareas de analítica descriptiva orientadas al ámbito de la situación actual y la evolución en la pérdida de privacidad:

- Clústering de aplicaciones benévolas centradas en el servicio o centradas en recopilación de datos.
- Crecimiento y evolución de rastreadores y permisos en distintas versiones de la aplicación o historia a nivel global.
- Reglas de asociación entre permisos solicitados del dispositivo y rastreadores incluidos en la aplicación.
- Clasificación de aplicaciones en base a sus permisos o propósitos comerciales de datos se refiere.

En analítica predictiva, se podría utilizar para:

- Realizar una clasificación de porcentaje de riesgo, utilizando como entrada al algoritmo los permisos o rastreadores que contiene una aplicación para alertar al usuario del riesgo que supone su instalación.
- También se podrían realizar modelos con propósitos más anecdóticos por la incorporación de los iconos de las aplicaciones tales como generadores de nuevos iconos inspirados en aplicaciones relacionadas.

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

Released Under CC BY-NC-SA 4.0 License: Permite crear y compartir adaptaciones del dataset mientras se haga bajo la misma licencia o equivalente y nunca permitiendo usos comerciales del mismo.

Se ha elegido esta licencia puesto que está centrada en contenido y dado el carácter público y no lucrativo de la organización Exodus, no poder permitir el uso comercial con la información extraída de sus informes.

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

El script Python *exodusWS.py* se encuentra en la carpeta *src* repositorio https://github.com/luimoco-uoc/exodusWebScraping.

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

DOI 10.5281/zenodo.4261664 https://zenodo.org/record/4261664#.X6fqv2hKiUk

* Nota: Atendiendo al apartado 4 Almacenamiento y compartición de datos del libro de texto **Subirats L., Calvo, M.** (2019) *Web scraping*. FUOC, he considerado que el archivo de datos más adecuado para este dataset sería **JSON** en lugar de CSV debido a que precisa de objetos lista para poder representar alguno de los atributos importantes del dataset.

Contribuciones	Firma
Investigación previa	**LMC**
Redaccion de las respuestas	**LMC**
Desarrollo código	**LMC**