M2.851 – Tipología y ciclo de vida de los datos: PRA1 - Modelado de un juego de datos

Autores: Joel Bustos e Iván Ruiz

Fecha: 14 de abril 2020

Introducción. Entrega final 14 de abril de 2020

Presentación

Mediante el presente informe, exponemos de manera ordenada y sintética, los retos que hemos identificado, y las propuestas de resolución que hemos encontrado, para abordar la práctica que se nos proponía.

Partíamos de un objetivo global, el de diseñar e implementar una solución de *web scraping* que tenía que abordar y resolver, algunos de los siguientes objetivos:

- Identificación de una fuente de datos con información interesante, que permita realizar algún tipo de estudio o explotación de datos posterior.
- Identificación de las tecnologías, posibles restricciones o particularidades, de la información almacenada en el sitio web.
- Diseño del flujograma de rascado de datos.
- Desarrollo, implementación y depuración de errores.
- Suministrar una herramienta de *web scraping,* con un nivel de documentación suficiente, que permita a cualquier usuario entenderla y utilizarla.
- Funcionalidades complementarias como controles de calidad y generación de registros del proceso.

A lo largo de las distintas secciones del presente informe, mostramos la motivación del proyecto de rascado de datos, así como las particularidades, retos, y conclusiones halladas, en relación a los datos del sitio web https://www.hackmageddon.com/.

Resulta interesante explicar, en esta sección introductoria, que una práctica como la propuesta, nos ha animado a explorar distintas soluciones y librerías Python, como: *BeautifulSoup*, *Requests* y *Selenium*. Además, a partir de debates y pruebas realizadas, hemos sido capaces de identificar fortalezas y debilidades de las distintas alternativas, consiguiendo llegar a un desarrollo final, que combina la funcionalidad con la facilidad de uso mediante la simplificación del código.

Nos gustaría, por último, destacar la importancia de la comunicación y del reparto de tareas, en un proyecto en el que dos personas tienen que colaborar en todas las etapas, tanto en la identificación del reto, como en la propuesta, la implementación y la validación de la solución final.

Respuesta a las cuestiones que se plantean en el guion de la práctica

1. Contexto

En el mundo digital en el que nos movemos es imprescindible, para las empresas, poder garantizar la confidencialidad de los datos creados, procesados y almacenados, para así, asegurar la continuidad de las operaciones de negocio. Por este motivo, las áreas de seguridad de la información, han asumido un papel de vital importancia.

En este sentido, los equipos de seguridad han necesitado incorporar, cada vez más, perfiles técnicos en el área de la ciberseguridad. Estos equipos técnicos, normalmente con un conocimiento muy específico, en ocasiones no disponen de demasiadas herramientas que les permitan ser proactivos y anticiparse a las nuevas tendencias y técnicas de ciberataque. De esta forma, acaban adaptando un comportamiento reactivo, realizando tareas de mantenimiento y de respuesta ante incidentes.

Nos planteábamos, como contexto para la presente práctica, recopilar datos históricos de ciberataques con el objetivo de crear un modelo predictivo que sirviese de soporte al equipo de seguridad de una empresa. Idealmente, estudiando lo que está ocurriendo en relación a delitos cibernéticos, los equipos internos de las distintas entidades, podrían tratar de prepararse mejor contra aquellos riesgos a los que los modelos estadísticos les pudieran sugerir que se encuentran más expuestos.

Partiendo de esta idea, hemos encontrado que existen múltiples recursos en forma de informes sobre amenazas y ataques, tales como [1], [2] o [3]. Estos informes, son muy útiles para los profesionales de seguridad, porque ofrecen una síntesis de los principales hechos ocurridos. Sin embargo, es difícil crear modelos estadísticos sobre este tipo de recursos, que presentan datos semiestructurados o desestructurados.

Por este motivo, al ver que no era del todo fácil encontrar una fuente de datos estructurada, nos hallamos ante la necesidad de crear una base de datos a partir de la cual, un científico de datos, pudiera hacer tareas de *data mining*.

En nuestras averiguaciones, encontramos la página https://www.hackmageddon.com/, en la que no sólo se ofrecen informes sintéticos para distintos periodos, sino que también, se facilitan los datos a partir de los cuales se construyen los informes.

Abordar un proyecto de rascado de datos, o *web scraping*, sobre una fuente de información como esta, podría permitirnos diseñar analíticas específicas, tales como el estudio de tendencias y correlaciones.

En esta primera sección, las principales tareas que hemos llevado a cabo, han sido dirigidas a resolver los dos primeros objetivos parciales, presentados en la introducción del informe. Estos objetivos son:

- **Identificación de la fuente de datos:**

  Tal y como hemos expuesto, nos ha parecido interesante, poder avanzar y profundizar en el estudio estadístico de los principales incidentes de seguridad documentados en *Hackmageddon*.

  Una vez encontrada la temática de interés (datos cuantitativos y estadísticos sobre incidentes de seguridad), ha resultado necesaria la comparación de distintos sitios web e informes relacionados con esta problemática. Finalmente, se ha llegado a la conclusión, de que las brechas de seguridad, son un problema creciente que requiere de tareas de análisis y prevención, para poder proporcionar una protección efectiva [4].

- **Identificación de la tecnología y posibles restricciones**.

  Una vez hemos decidido extraer datos del sitio web *Hackmageddon*, hemos necesitado hacer una evaluación técnica mínima, que nos permitiese definir y decidir, algunos de los mecanismos de rascado de datos.

  En primer lugar, hemos analizado el archivo *robots.txt*, accesible mediante: https://www.hackmageddon.com/robots.txt, y cuyo contenido es el siguiente:

  ```
  User-agent: *
  Disallow: /wp-admin/
  Allow: /wp-admin/admin-ajax.php
  ```

  De la definición anterior, se desprende [5] que las directrices definidas son aplicables a cualquier agente o robot (User-agent: *), restringiendo su acceso al directorio /wp-admin/, pero habilitando específicamente el archivo /wp-admin/Admin-ajax.php.

  Es decir, de las pautas que el propietario del sitio web ha definido, no encontramos indicaciones que prohíban el acceso a los directorios web, sobre los cuales, pretendemos extraer información mediante *scraping*.

  Las URLs principales de consulta son:

  ```
  hackmageddon.com/2017-master-table/
  hackmageddon.com/2018-master-table/
  hackmageddon.com/category/
  hackmageddon.com/2020/
  hackmageddon.com/2019/
  ```

  Adicionalmente, el contenido de la página web no es muy extenso, por lo que la eficiencia, no será un problema importante. De esta forma, se realizarán descargas secuenciales y no concurrentes.

  Por último, en cuanto a las tecnologías identificadas, hemos visto distintas herramientas y técnicas para la creación de la página web. Ante esta heterogeneidad de tecnologías, se ha requerido el uso de distintas técnicas de extracción de datos mediante *web scraping*.

Por este motivo, se ha combinado y homogeneizado, el proceso de rasgado de datos mediante la navegación web con *Selenium*, y el parseo de datos a partir de *Request* y *BeautifulSoup.*

A continuación, se definen los principales procesos que se han realizado durante el proceso de *web scraping.*

- Extracción de la información contenida en tablas mediante código HTML (datos tabulados mediante tags <td>).
- Navegación web a través de las referencias halladas en el código HTML (tags <a href>).
- Descarga de ficheros a partir de Google Apps y *JavaScript*.

Ha resultado muy interesante dedicar un tiempo al estudio de todo lo anterior. Adicionalmente, la distribución de la información a lo largo de la web, y el planteamiento propuesto para llevar a cabo el proceso de extracción de datos, ha permitido segmentar la asignación de tareas de desarrollo, entre los dos miembros del equipo.

2. Título del dataset

Información y referencia de ciberataques documentados por hackmageddon.com en el periodo 2017-2020.

3. Descripción

Hablamos de "información y referencia", porque el *dataset* que presentamos, ofrece datos sobre los cuales aplicar técnicas de minería de datos con el objetivo de obtener conocimiento y, además, referencias de los incidentes en forma de noticias, que permitirán disponer de más detalles sobre el contexto de los ataques.

El set de datos extraído de *hackmageddon.com*, ofrece una **visión temporal de los ataques que se han producido y documentado a gran escala** (de ahí que se trate de incidentes que tienen noticias asociadas), así como de los autores, mecanismos de ataque y principales afectados por dichos ataques.

Dado que pretendemos continuar con la limpieza y explotación de datos en una práctica posterior, el set de datos obtenido tras la tarea de *web scraping,* se encuentra la carpeta *Data/00_raw* de GitHub. En esta carpeta, se incluyen los siguientes ficheros:

- *Master Data 2017 2020-04-11 20.28.20.csv.* Set de datos correspondiente a los datos históricos que el sitio web almacena para el año 2017 completo.
- *Master Data 2018 2020-04-11 20.20.20.csv.* Set de datos correspondiente a los datos históricos que el sitio web almacena para el año 2018 completo.

- *Scrapping 2020-24 20.28.20.csv*. Set de datos correspondiente a datos desde el 1 de enero de 2019 hasta el momento de ejecución del *web scraper*.

Como se aprecia en la propia nomenclatura de los ficheros, hemos incorporado marcas de tiempo, con el objetivo de tener trazabilidad del momento en el que se realiza la descarga de datos. En el ejemplo, se observa que los archivos han sido descargados el día 11 de abril de 2020, a las 20:28:20 horas.

Tal y como se describe en la documentación del código, el *scraper,* acepta como parámetros de entrada un rango de fechas para descargar los datos.

Esta metodología, permite la actualización de los datos haciendo un uso mucho más eficiente del robot, es decir, si planteamos utilizar este *scraper* para la construcción de una base de datos de incidentes, cabe esperar que inicialmente será necesaria una descarga completa de datos y, posteriormente, será necesario actualizar dicha base, mediante descargas incrementales periódicas. De esta forma, gracias al planteamiento que ofrecemos, una vez descargado y validado el set de datos completo inicial, en ejecuciones posteriores, podría solicitarse únicamente la descarga de los datos adicionales requeridos.

Otro aspecto positivo de este planteamiento, es la posibilidad de relanzar procesos fallidos durante periodos de tiempo acotados. Es decir, si por algún motivo, existe información que no se ha descargado correctamente, cabe la posibilidad de relanzar el proceso para un periodo de tiempo determinado, sin tener la necesidad de ejecutar el proceso completo.

Además, como líneas futuras en tareas de limpieza y acondicionamiento, podrían validarse estas descargas incrementales, para verificar que no se estuvieran incorporando casos duplicados.

4. Imagen identificativa

El set de datos contendrá información de los ciberataques realizados a empresas de distintos países. Estos ataques, son producidos por distintos autores desde cualquier parte del mundo.

En la *Figura 1,* se puede observar la información del dataset a través de líneas que conectan a los atacantes con los afectados de los ciberataques. De esta forma, se pueden detectar aquellas zonas con mayor y menor número de ciberataques. Adicionalmente, los colores de las líneas, representarían distintos tipos de ataques cibernéticos.

Por otra parte, también nos ha parecido representativa del tipo de estudios que se pueden llevar a cabo, con datasets como el que vamos a generar, la mostrada en [6]. Este sitio web habla de la importancia de la monitorización del tráfico.

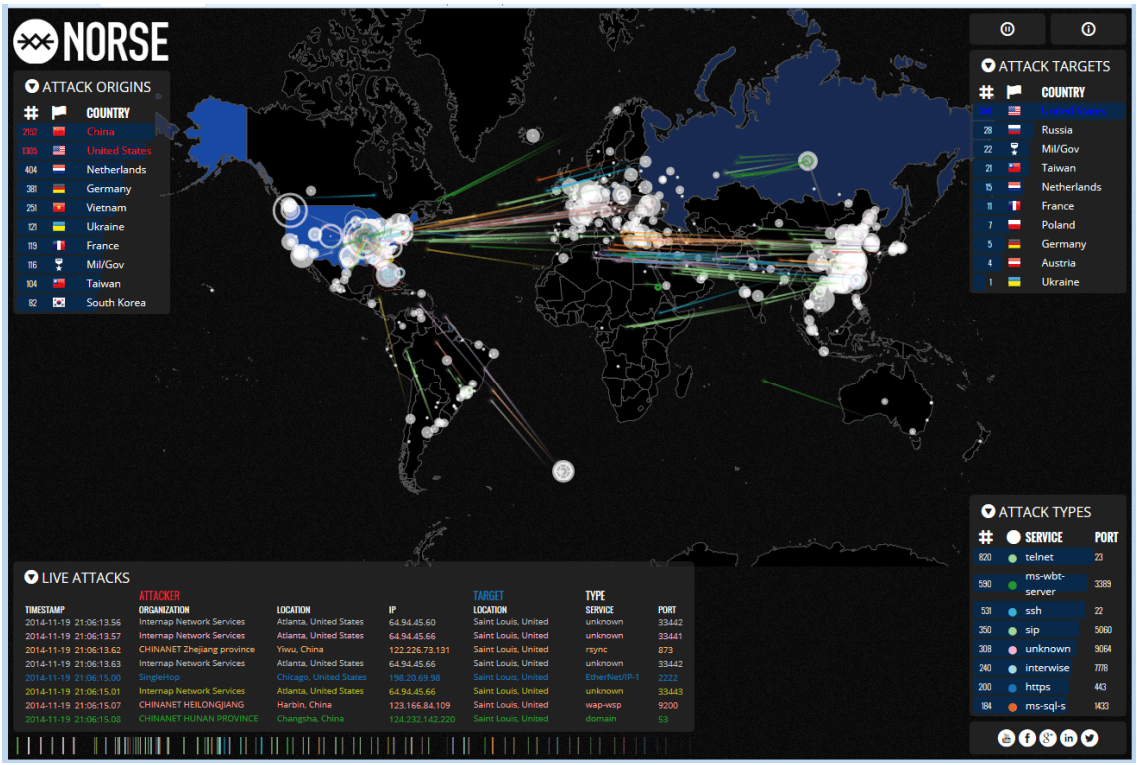

Figura 1: Representación de los ataques cibernéticos que se producen en tiempo real

5. Contenido

Los sets de datos que hemos generado contienen los siguientes campos:

Id (*id*). Secuenciador numérico presente en algunos formatos de informe. Dado que no se trata de un valor único universalmente en el sitio web, será necesario aplicar acondicionar y normalizar el campo en la segunda práctica.

Fecha del rascado (valor incluido en la nomenclatura del fichero). Disponer de este dato nos permitirá llevar a cabo procesos de calidad. Además, dotaremos de historicidad y volatilidad a la base de datos, ofreciendo información acerca de la vigencia de los datos extraídos. Este campo es generado por la propia ejecución del *scraper* y se encuentra incluido en la nomenclatura del documento rasgado. Por este motivo, será necesario procesarlo en la segunda práctica.

Fecha del informe (*date_report*). Una de las aportaciones de nuestro *web scraper* es permitir el rascado periódico de los informes quincenales. De este modo, con este campo, capturamos cuando se publicó el informe con el objetivo de establecer controles de calidad de datos.

Para los años 2018 y 2017, al obtenerse la información a través del rasgado de un único informe, el campo *date_report* se añadirá, a posteriori con los valores “Anual 2018” y “Anual 2017”, respectivamente. Este procedimiento, formará parte de las tareas de normalización y limpieza de datos que acometeremos en una segunda práctica.

Autor del informe (*author_report*). Persona que aparece como autor en cada uno de los informes.

Fecha del ciberataque (*date*). Fecha en la que se produjo el ataque. Este campo será utilizado en prácticas posteriores para tareas de limpieza y validación del set de datos, ya que hemos observado alguna incoherencia entre la fecha del informe y la fecha del ciberataque, que será necesario resolver posteriormente.

Atacante (*author*). Nombre del causante del ataque cibernético. En este campo. se han detectado casos con varios atacantes, y casos, en los que se desconoce la autoría del ataque. Durante las tareas de procesado de datos, que se realizarán en la próxima práctica, se normalizarán estos valores a "Varios" y "No disponible" respectivamente.

Objetivo (*target*). Descripción del objetivo del ataque.

Tipo de objetivo (*target_class*). Clasificación de los distintos objetivos en categorías. Estos, se pueden agrupar en una lista de valores finita.

Según hemos podido identificar, existen un máximo de 24 categorías, de acuerdo a la relación siguiente:

C Manufacturing
D Electricity gas steam and air conditioning supply
E Water supply, sewerage waste management, and remediation activities
G Wholesale and retail trade
H Transportation and storage
I Accommodation and food service activities
J Information and communication
K Financial and insurance activities
L Real estate activities
M Professional scientific and technical activities
N Administrative and support service activities
Not Found
O Public administration and defence, compulsory social security
O Public administration, defence, compulsory social security
P Education
Q Human health and social work activities
R Arts entertainment and recreation
S Other service activities
U Activities of extraterritorial organizations and bodies
V Fintech
X Individual
Y Multiple Industries
Y Multiple targets
Z Unknown

Como mostramos en la tabla anterior, existen casos en los que no resulta posible obtener un valor de categoría (valor "*not found*"). Para estos casos, en la segunda práctica, podremos estudiar cómo informarlos.

Descripción (*description*). Texto suficientemente descriptivo de lo ocurrido en el ataque.

Técnica de ataque (*attack*). Metodología de ataque empleada. Este campo pretende ser un resumen mucho más ejecutivo del que ofrece el campo "descripción", incluyendo información suficientemente detallada del ataque. Existen 305 valores distintos en el set de datos obtenido, por lo que no se incluirá la lista de valores.

Categoría de ataque (*attack_class*). Clasificación de las distintas metodologías de ataque en función de su motivación o finalidad. En este caso, sí que se agrupan los principales tipos de ataques en una lista cerrada de valores, que presentamos a continuación:

CC
>1
CC
CE
CW
CW?
Cyber Crime
Cyber Espionage
Cyber Warfare
H
Hacktivism
N/A
Not Found

Como se puede apreciar, también existen valores no encontrados ("*not found*") que podremos estudiar en la segunda parte de la práctica de la asignatura.

Pais (*country*). País afectado por el ataque. En caso de que haya varios países afectados, normalizaremos al valor a "Varios".

Enlace documento (*link*). URL con información mucho más descriptiva de lo ocurrido y del impacto del ataque. Esta documentación adicional, podría ser el punto de partida para proyectos de *scraping* posteriores, que tuvieran la finalidad de enriquecer los datos generados en el presente proyecto.

Visitas (*views*). En los casos en los que existe esta información, almacenamos también el número de visitas recibidas del informe, en el momento de ejecución del *scraper*.

Tags (*tags*). Conjunto de palabras clave asignadas a un ciberataque, con el objetivo de describirlo a partir de etiquetas. No todos los datos extraídos contienen esta información.

Para conseguir obtener todos los campos anteriores, y siguiendo otro de los objetivos parciales presentados en la sección introductoria, podríamos resumir los principales pasos que sigue el **flujograma del desarrollo** en:

1) **Rascado de datos de las tablas** que componen los informes quincenales. Como por ejemplo: https://www.hackmageddon.com/2020/03/17/16-29-february-2020-cyber-attacks-timeline/

2) **Rascado iterativo de los distintos informes quincenales** que existen publicados en el *timeline*. https://www.hackmageddon.com/category/security/cyber-attacks-timeline/.

   Este proceso nos permitirá disponer de los datos quincenales de los ataques que se han producido entre 2019 y la actualidad (último informe disponible, primera quincena de marzo de 2020)

3) **Rascado especial de las dos tablas históricas** disponibles para los años 2018 y 2017, a través de los enlaces siguientes:

   https://www.hackmageddon.com/2018-master-table/

   https://www.hackmageddon.com/2017-master-table/

4) **Compatibilidad con distintos navegadores y configuraciones**. Nuestro desarrollo , a fecha de elaboración del presente informe, es compatible con los navegadores Firefox y Chrome[1]. Además, ofrece robustez ante las distintas configuraciones presentes en los navegadores web.

5) **Parametrización, por parte del usuario, del periodo temporal sobre el que se debe realizar el rasgado web** con el objetivo de obtener el set de datos deseado.

6) Incorporación **de un logging para el control de errores y control de calidad**. Se generarán logs de ejecución, con el objetivo de hacer un seguimiento del número de registros obtenidos correctamente, e identificar aquellos en los que se ha producido algún error. De esta forma, se pretende detectar posibles modificaciones en el código HTML, para posteriormente, adaptar el script y que siga siendo operativo.

7) **Control del *timming* entre peticiones web.** Se ha establecido un periodo de espera entre peticiones, con el objetivo de no saturar la página web, y dar tiempo a que las distintas acciones de navegación se completen correctamente. Adicionalmente, para evitar ser bloqueados, y simular peticiones realizadas por un 'humano', se ha modificado el *User-Agent* en las cabeceras de las peticiones.

8) Modificación del código para **ejecución *stealth* o silenciosa**, que permita la descarga sin necesidad de que la interfaz de usuario del navegador esté presente (ejecución en segundo plano).

9) **Exportación de datos brutos** a tres ficheros csv que serán objeto de consolidación, normalización y limpieza en la segunda práctica de la asignatura.

6. Agradecimientos

Aunque la motivación y el interés por el mundo de la ciberseguridad tiene muchos autores a los que agradecerles, en este caso en particular, es obligado agradecerle a Paolo Passeri el diseño y el desarrollo de su sitio web Hackmageddon.

Paolo es consultor de seguridad y bloguero en el campo de la seguridad de la información [7]. De hecho, seguramente que la frecuencia de actualización de sus estadísticas cibernéticas, permitirán contribuir a mejorar la detección y prevención de este tipo de actos maliciosos.

De manera adicional al diseño y a la provisión de información sobre ataques, Paolo es director de inteligencia cibernética en Netskope [8], compañía que además de prestar servicios en el ámbito de la seguridad de la información, ofrece multitud de artículos y reflexiones en forma de blog.

7. Inspiración

A priori, sin haber realizado tareas de explotación de datos, se podrían responder preguntas acerca de cuál es el tipo de ciberataque más frecuente, con el objetivo de poder generar más recursos de seguridad y prevención.

Además, gracias a las distintas variables categóricas que incluye el set de datos, podrían diseñarse pruebas de datos dirigidas a identificar el sector de empresas que es más vulnerable (recibe más ciberataques). De la misma forma, observar que país es el más atacado y ver si tiene relación con el tipo de empresas de ese país.

En el ámbito temporal, como también disponemos de información del momento en el que se han producido los ataques, podrían diseñarse modelos para detectar patrones temporales y, de esta forma, poder predecir ciberataques. ¿Es más normal que se produzcan ciberataques los fines de semana o entre semana?. ¿En qué periodo del año se producen más ciberataques? El modelo generado, permitiría dar respuesta a preguntas como las anteriores y se podría, por ejemplo, predecir la volumetría de trabajo y adaptar el número de miembros necesarios para formar un equipo de ciberseguridad.

Adicionalmente, combinando todas las perspectivas (tipología de ataque, de entidad o país atacado, y del momento temporal en que se produce), podría ser muy interesante estudiar la tendencia, y predecir qué sectores, países o momentos futuros del tiempo serán más susceptibles de volver a registrar un ataque de amplia repercusión. De este modo, añadiendo esta capacidad predictiva al proyecto de minería de datos, no sólo podría mejorarse el dimensionamiento y capacidades que requiere un equipo de ciberseguridad que trabaje en un sector o país concreto, sino que, además, se podrían tener marcados momentos de alto riesgo de acuerdo al modelo.

De manera más directa, a continuación, se expone una lista con algunas preguntas que los modelos de análisis de datos podrían tratar de resolver.

- ¿Qué países son más regularmente atacados, y qué previsión hay en el futuro de que estos países sigan recibiendo ataques?

- En función del sector (financiero, sector público, sanidad, etc.), ¿cuáles son los principales riesgos contra los que debo prepararme?

- ¿Existe alguna correlación entre momentos temporales en los que se producen los ataques, y los tipos de ataques?

- Los principales actores maliciosos (atacantes) en el panorama internacional, ¿están cambiando sus patrones de ataque, o cabe esperar que continúen empleando las mismas técnicas?

8. Licencia

De acuerdo a los distintos niveles de permisos de uso y distribución, y compromisos adquiridos, proponemos que el set de datos tenga licencia **CC BY-NC-SA 4.0 License** [9], por los motivos que exponemos a continuación, y que están alineados con la interpretación simplificada que *Creative Commons* ofrece en [10]:

Dado el carácter divulgativo y académico que tiene el proyecto que hemos desarrollado, nos parece enriquecedor para la comunidad, que la información se pueda compartir y adaptar de acuerdo a las siguientes restricciones:

- Atribución. Será necesario hacer referencia y reconocer la autoría, así como proporcionar un enlace a la licencia y hacer referencia a los cambios que hubieran podido realizarse. Elegimos este modo de licencia, precisamente para reconocer en primera instancia el esfuerzo de recopilación y publicación de datos realizado por el sitio web del que extraemos los datos, y en última instancia al autor, Paolo Passeri, a quien ya hemos mencionado en la sección de agradecimientos.

- No comercial. No podrá utilizarse el set de datos para una finalidad comercial, ya que creemos que el mayor esfuerzo de producción de los datos corresponde al sitio web original, y dado que desconocemos el impacto que los fines comerciales podrían tener sobre ese autor, preferimos mantener este material para fines puramente de análisis y prospección de información.

- Compartir Igual. En caso de que alguien haga uso del set de datos, o lo transforme de alguna manera, deberá difundir sus contribuciones bajo la misma licencia que el original

Tal y como se indica en el sitio web de Creative Commons, el uso de este modo de licencia podrá indicarse con dos botones de apariencia como los siguientes:

En caso de publicar el set de datos en algún sitio web, se recomienda hacer uso del siguiente fragmento HTML para hacer referencia al modo de licencia seleccionado:

```
<a rel="license" href="http://creativecommons.org/licenses/by-nc-sa/4.0/"><img alt="Licencia de Creative Commons" style="border-width:0" src="https://i.creativecommons.org/l/by-nc-sa/4.0/88x31.png" /></a><br />Este obra está bajo una <a rel="license" href="http://creativecommons.org/licenses/by-nc-sa/4.0/">licencia de Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional</a>.
```

9. Código

El código utilizado para el desarrollo del proyecto, puede encontrarse en la carpeta *src* del repositorio *Cyberattacks-History: https://github.com/iruiper/Cyberattacks-History*

10. Dataset

El dataset puede encontrarse en la carpeta *data/00_raw* del repositorio Cyberattacks-History: https://github.com/iruiper/Cyberattacks-History

Tal y como hemos mencionado a lo largo del informe (ver sección 3), el raspado de datos proviene de distintas partes del sitio web con formato irregular. Por este motivo, disponemos de tres archivos csv. Su consolidación, limpieza y normalización se realizará en la segunda parte de la práctica.

Contribuciones

Investigación Previa	JBP – IRP
Redacción de las respuestas	JBP – IRP
Desarrollo código	JBP – IRP

Bibliografía/Recursos

[1] Check Point Research – Centro de publicaciones. [en línea] [última visita: 12/Abr/2020] – URL: https://research.checkpoint.com/category/threat-intelligence-reports/

[2] FireEye – Annual Threat Reports. M-Trends 2020. [en línea] [última visita: 12/Abr/2020] – URL: https://www.fireeye.com/current-threats/annual-threat-report.html

[3] Center for Strategic & International Studies (CSIS). Significant Cyber Incidents. [en línea] [última visita: 12/Abr/2020] – URL: https://www.csis.org/programs/technology-policy-program/significant-cyber-incidents

[4] phoenixNAP – Global IT Services. 81 Eye-Opening Data Breach Statistics for 2020. [en línea] [última visita: 12/Abr/2020] – URL: https://phoenixnap.com/blog/data-breach-statistics

[5] Critchlow, Will (2013) – Publicado en MOZ. "Learn About Robots.txt with Interactive Examples". [en línea] [última visita: 12/Abr/2020] – URL: https://moz.com/blog/interactive-guide-to-robots-txt

[6] Firewall.cx – "The importance of monitoring and controlling web traffic in enterprise & SMB networks – Protecting from malicious websites – Part 1". [en línea] [última visita: 12/Abr/2020] – URL: http://www.firewall.cx/general-topics-reviews/security-articles/1064-security-protect-enterprise-smb-network-web-monitoring-part1.html

[7] CyberSecurity news – "Paolo Passeri". [en línea] [última visita: 12/Abr/2020] – URL: https://cybersecuritynews.es/paolo-passeri/

[8] netskope – Página web principal de la compañía. [en línea] [última visita: 12/Abr/2020] – URL: https://www.netskope.com/

[9] Creative Commons – "Attribution-NonCommercial-ShareAlike 4.0 International". [en línea] [última visita: 12/Abr/2020] – URL: https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

[10] Creative Commons – "Attribution-NonCommercial-ShareAlike 4.0 International". Human-readable summary. [en línea] [última visita: 12/Abr/2020] – URL: https://creativecommons.org/licenses/by-nc-sa/4.0/