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

Autores: Joel Bustos e Iván Ruiz

Fecha: 2 de abril 2020

Introducción. Borrador semana 30 de marzo de 2020

Presentación

El presente documento describe el planteamiento que hemos hecho para resolver el reto que se nos plantea en esta primera práctica.

En relación al desarrollo del código, hemos de decir que nos encontramos aún familiarizándonos con todo el entorno que es necesario manejar, y las distintas librerías de Python que emplearemos. Por ahora, se ha planteado el uso de los materiales dispuestos en los recursos, como la librería BeautifullSoup y librerías adicionales como Scrappy o Selenium.

Asimismo, hemos realizado una evaluación inicial del sitio web que pretendemos scrapear. Primero de todo, se ha realizado una búsqueda del archivo *robots.txt,* para saber que directorios están disponibles/permitidos. Adicionalmente, utilizando las funciones de inspección de navegadores web, se ha identificado la estructura del site.

Las secciones resaltadas en amarillo se encuentran en curso, y podrán sufrir cambios significativos tanto como resultado de los comentarios que podamos recibir, como del propio desarrollo de la práctica.

Respuesta preliminar de las cuestiones que se plantean

---

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. Este equipo técnico, normalmente con un conocimiento muy específico, en ocasiones no dispone 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 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 también 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 ésta, podría permitirnos diseñar analíticas específicas, estudio de tendencias y correlaciones, y además podríamos plantearnos una actualización del set de datos con una cierta periodicidad

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 set de datos 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.

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 los atacantes con los afectados de los ciberataques. Se pueden detectar zonas con mayor y menor número de ciberataques. Los colores de las líneas representarían los tipos de ciberataque.

Nos ha parecido representativa del tipo de estudios que se pueden llevar a cabo con datasets como el que vamos a generar, la que se presenta en [4]. Este sitio web habla de la importancia de la monitorización del tráfico. Nuestro proyecto irá dirigido a tratar de analizar los hechos históricos sobre ciberataques.

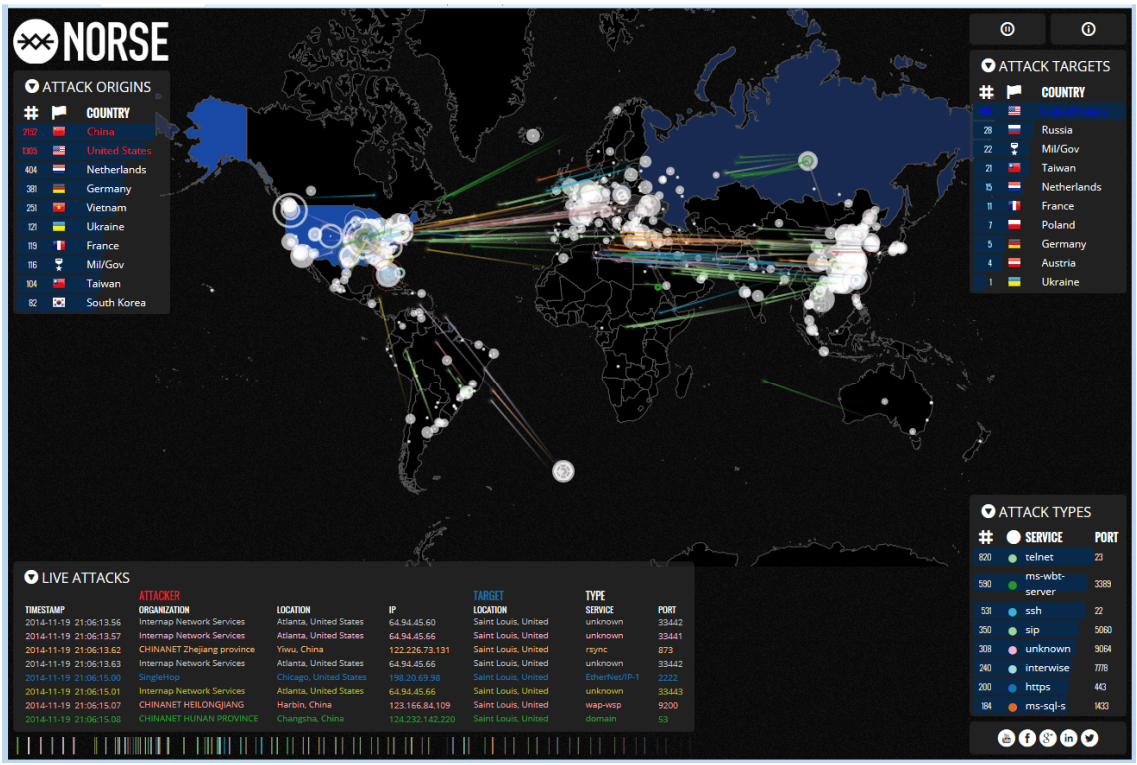

Figura 1: Representación de los ataques cibernéticos que se producen en tiempo real

5. Contenido

Los sets de datos que pretendemos generar contendrían, al menos, los siguientes campos:

Fecha del rascado. Disponer de este dato nos permitirá llevar a cabo procesos de calidad. Además dotaremos de historicidad a la base de datos.

Fecha del informe. Una de las aportaciones de nuestro *web scraper* es permitir el rascado periódico de los informes quincenales que se suelen publicar. De este modo, con este campo capturamos de qué informe concreto se han obtenido los datos.

En el caso de los años 2018 y 2017, existe información detallada para todo el año, por lo que la técnica de *scraping* que aplicaremos será distinta, y el valor que asignaremos en este campo para esos valores históricos será "Anual 2018" y "Anual 2017" respectivamente.

Autor del informe. Persona que aparece como autor en cada uno de los informes.

Fecha del ciberataque. 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. En aquellos casos en los que se conozca la autoría, este campo contendrá al atacante identificado. En todos los casos sin valor, normalizaremos el valor a "No disponible; en los casos en los que haya varios atacantes identificados, normalizaremos al valor "Varios".

Objetivo. Descripción del objetivo del ataque.

Tipo de objetivo. Los distintos objetivos se pueden agrupar en una lista cerrada de posibles objetivos, que se recogen en esta variable cualitativa discreta. **Suministraremos la lista cerrada de posibles valores cuando la hayamos acotado tras el proceso de *scraping***.

Descripción. Texto suficientemente descriptivo de lo ocurrido en el ataque.

Técnica de ataque. Las técnicas de ataque se pueden agrupar en una lista cerrada, que se recogen en esta variable cualitativa discreta. **Suministraremos la lista cerrada de posibles valores cuando la hayamos acotado tras el proceso de *scraping***.

Categoría de ataque. Viene a constituir la motivación del ataque, y se representa con una lista cerrada de posibles categorías de ataque. **Suministraremos la lista cerrada de posibles valores cuando la hayamos acotado tras el proceso de *scraping***

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

Enlace documento. URL con información mucho más descriptiva de lo ocurrido y el impacto del ataque. Esta librería de documentación adicional al dataset podría constituir proyectos de *scraping* adicionales, que vinieran a enriquecer los datos generados en el presente proyecto.

El proyecto de extracción de los datos anteriores se va a abordar de acuerdo a las siguientes etapas:

1) Rascado de datos de las tablas que componen los informes quincenales. Se puede consultar un ejemplo de tabla en 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 site, trazables a partir de 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 2018 y la actualidad (último informe disponible, primera quincena de marzo de 2020)

3) Rascado especial de las dos tablas históricas disponibles en el site 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) Consolidación de los datos rascados a partir de los tres proceso anteriores, y exportación a un único set de datos.

6. Agradecimientos

[A completar a la finalización]

7. Inspiración

[A completar con ideas de cómo explotar y qué tipo de analíticas de datos podrían llevarse a cabo a partir del set de datos generado]

A primera vista a modo de borrador se podrían responder preguntas acerca de cual es el tipo de ciberataque más empleado, para poder crear más recursos que permitan hacer frente a estos.

Identificar el sector de empresas que es más vulnerable (recibe más cyber ataques). De la misma forma, observar que país es el más atacado y ver si tiene relación con el tipo de empresas de ese país.

Detectar patrones de tiempo en los ciberataques. Es más normal que se produzcan ciberataques los fines de semana, entre semana, en determinados periodos del año. De esta forma se podría predecir la volumetría de trabajo para el equipo de ciberseguridad.

8. Licencia

[A completar a la finalización]

Contribuciones

Investigación Previa	JBP – IRP
Redacción de las respuestas	JBP – IRP [En curso]
Desarrollo código	[En curso]

Bibliografía/Recursos

[1] https://research.checkpoint.com/category/threat-intelligence-reports/

[2] https://www.fireeye.com/current-threats/annual-threat-report.html

[3] https://www.csis.org/programs/technology-policy-program/significant-cyber-incidents

[4] http://www.firewall.cx/general-topics-reviews/security-articles/1064-security-protect-enterprise-smb-network-web-monitoring-part1.html