BENEMÉRITA UNIVERSIDAD
AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

DOCTORADO EN INGENIERÍA DEL LENGUAJE Y DEL CONOCIMIENTO

TESIS: **METODOLOGÍA PARA LA IDENTIFICACIÓN Y CLASIFICACIÓN DE DELITOS CIBERNÉTICOS EN MÉXICO UTILIZANDO LA RED TOR**

Tesis que para obtener el título de
DOCTOR EN INGENIERÍA DEL LENGUAJE Y DEL CONOCIMIENTO

PRESENTA:

JULIO JESÚS SALAS CONDE

DIRECTOR DE TESIS:
DR. MANUEL ISIDRO MARTÍN ORTÍZ
BUAP

COASESOR EXTERNO
DR. VICTOR CARNEIRO DÍAZ
UNIVERSIDAD DE LA CORUÑA ESPAÑA

Noviembre 2022

Contenido

Resumen

Hoy en día, los delitos cibernéticos son un tema de interés mundial debido a su complejidad e impacto, este trabajo expone una revisión a los delitos cibernéticos en la Dark Web que impactan en México, con el objetivo de desarrollar una metodología la cual descubra redes onion usando el navegador Tor, extraiga información contenida en redes onion, se depure la información para obtener información útil, utilice un algoritmo de clasificación que permita identificar patrones de similitud entre las páginas identificadas con delitos cibernéticos con respecto a una ontología de términos dentro de estas redes onion descubiertas previamente, y finalmente presentar estos resultados en páginas web a través de la visualización de información en tablas de datos y gráficas de análisis de estos mismos, esto con el fin de ayudar en la prevención y el combate de los delitos cibernéticos en instituciones de seguridad del país y que permita al investigador de éstas obtener información relevante la cual no obtiene de la Web tradicional, de tal forma que se impacte directamente en la sociedad que ha sufrido este tipo de delitos y requiere el estudio de nuevas tecnologías que aporten información fresca que ayude a la resolución de los casos de investigación en delitos cibernéticos.

Abstract

Today, cybercrimes are a topic of global interest due to their complexity and impact, this work presents a review of cybercrimes on the Dark Web that impact Mexico, with the aim of developing a methodology that which discovers onion networks using the Tor browser, extracts information contained in onion networks, depurate the information to obtain useful information, uses a classification algorithm that allows identifying patterns of similarity between the pages identified with cybercrimes with respect to an ontology of terms within these previously discovered onion networks, and finally present these results on web pages through the visualization of information in data tables and analysis graphs of these, this in order to help in the prevention and the fight against cybercrimes in security institutions of the country and that allows the investigator of these to obtain relevant information the which it does not obtain from the traditional Web, in such a way that it has a direct impact on the society that has suffered this type of crime and requires the study of new technologies that provide fresh information that helps to resolve cybercrime investigation cases.

1. Introducción

1.1 Descripción General

Hoy en día, los delitos cibernéticos son un tema de interés mundial debido a su complejidad e impacto. Las agencias gubernamentales de investigación de este tipo de conductas en México, necesitan de metodologías y herramientas que les permitan estudiar y comprender el fenómeno del crimen cibernético, esta disertación expone una revisión de este fenómeno en la Dark Web, con el objetivo de desarrollar una metodología la cual descubra redes privadas virtuales con TOR, incluyendo sitios web onion como foros, salas de chat, blogs, redes sociales, páginas web de compra y venta de armas, estupefacientes y pornografía infantíl, entre otros, posteriormente se descarguen estos sitios onion, se depuren para obtener la información relevante, y a través de identificar patrones en páginas con conductas delictivas basadas en una ontología (previamente definida) apoyada en la Ley Federal Contra la Delincuencia Organizada de México (publicada en el Diario Oficial de la Federación el 7 de noviembre de 1996 con última reforma publicada el 20 de mayo de 2021), además de generar un algoritmo de clasificación que permita identificar delitos cibernéticos dentro de las redes onion descubiertas.

Una vez recabada esta información de las redes onion con conductas delictivas basadas en la Ley Federal Contra la Delincuencia Organizada de México, se publica esta información en páginas web a través de tablas de datos y gráficas conteniendo información de conductas delictivas en Tor, con el fin de ayudar en la prevención y el combate de los delitos cibernéticos por las instituciones de seguridad del país y que permita a los investigadores obtener información relevante, la cual no se obtiene de la Surface Web, impactando directamente en la sociedad que ha sufrido este tipo de delitos y lo cual requiere del estudio de nuevas tecnologías que

aporten información que contribuya a la resolución de los casos de investigación.

Además, se presentan elementos que ayuden a la reflexión de lo que conlleva la investigación de delitos cibernéticos en la Dark Web.

1.2 Problemática

En México se cuenta con una estrategia de ciberseguridad a nivel federal y estatal, la cual atiende delitos cibernéticos en atención a denuncias de los ciudadanos a través del ministerio público o vía telefónica atendidos en la línea del “911” o a través de servicio web o aplicaciones móviles.

Los investigadores de los delitos cibernéticos tienen las atribuciones de investigar en el ciberespacio conductas delictivas que se cometan a través de medios electrónicos como computadoras, tabletas y celulares entre otros, e interactuar con redes de comunicaciones como internet. Además reciben instrucción por parte de ministerios públicos y otras agencias de seguridad solicitándoles buscar datos como: nombres de personas, domicilios, placas de carros, pseudónimos, etc. en internet y otras fuentes de información.

Comúnmente estas investigaciones se llevan a cabo mediante buscadores (como Google y Bing), y redes sociales (Facebook, Twitter, Instagram, etc.), sin embargo, estas redes de información, aunque amplias, son limitadas en cuanto al total de información que se encuentra en todo el contenido de internet, y los delitos cibernéticos se cometen en mayor medida dentro de las llamadas redes privadas virtuales dentro de la Deep Web siendo TOR la más usada.

En la actualidad es de vital importancia el estudio de las redes privadas virtuales en Internet, ya que es conocido el uso de las mismas para la comisión de delitos en los ámbitos informáticos, por ejemplo, se puede citar que existen delitos cibernéticos en redes TOR, los cuales requieren de un análisis para ser identificados y proporcionar información relevante a los investigadores.

Además, es importante destacar que no solamente esto sucede en la red TOR, existen otras redes como Ares, Freenet, I2P, en la Deep Web, sin embargo, esta es la más usada.

Estos delitos cibernéticos en la red TOR, requieren de un análisis para ser identificados y proporcionar información relevante a investigaciones llevadas a cabo por investigadores de los mismos, ya sea para el monitoreo de este tipo de delitos o para la búsqueda de información relevante en el esclarecimiento de un hecho.

El no investigar los delitos cibernéticos en las redes privadas de la Dark Web, es un problema que afecta a la ciudadanía directamente, esto debido a que se puede encontrar información relevante la cual no se encuentre en la Surface Web, que pueda ayudar a las instituciones de seguridad pública en el esclarecimiento de un delito y llevar a la resolución del mismo.

Además, se convierte en un problema para las instituciones de seguridad pública el no realizar investigaciones en la Dark Web, debido a que deben mantenerse a la vanguardia en los avances tecnológicos cibernéticos como se hace en países de primer mundo.

Debido a que los delincuentes cibernéticos utilizan hoy en día redes privadas como TOR para cometer estos delitos, es indispensable que las

autoridades en este campo conozcan y tengan el conocimiento suficiente para prevenir y combatir estos mismos con estas tecnologías.

Por lo tanto, es necesario investigar los delitos cibernéticos en la Dark Web para mejorar el proceso de investigación en un caso y dar más y mejores datos a las autoridades correspondientes, así como prevenir estos delitos al monitorear estas conductas en estas redes privadas.

También es importante tener conocimiento del marco legal aplicable a este tipo de delitos y los medios por los que se pueden investigar. Así como investigar cuáles son las medidas necesarias para acceder a este tipo de redes privadas como TOR.

Es necesario realizar una investigación a fondo del entorno que encierra el uso de redes privadas virtuales como TOR para enfrentar y resolver los retos y problemáticas que conlleva hacer uso de estas tecnologías, así como un estudio sobre lo que es permitido en el aspecto legal y los límites que conlleva.

Otra problemática importante a considerar es la capacitación que se le debe proporcionar al investigador de delitos cibernéticos en estas redes privadas virtuales, ya que se necesitan conocimientos especializados para cubrir la identidad y la seguridad al usar estas tecnologías.

1.3 Objetivos

Objetivo General

Desarrollar una metodología de identificación, recuperación, clasificación y presentación de los datos que ofrece la Dark Web a partir de la red TOR para descubrir datos relevantes de delitos cibernéticos que ayuden

en el proceso de investigación que conllevan las autoridades competentes.

Objetivos específicos:

- Estudiar e implementar mecanismos para identificar redes onion.
- Investigar, diseñar e implementar modelos para extraer la información contenida en las redes onion identificadas.
- Elaborar una Ontología de términos que ayude a identificar las Conductas Delictivas de las redes onion extraídas.
- Modelar los datos obtenidos e impementar un algorítmo de clasificación de delitos cibernéticos con la información extraída de las redes onion.
- Presentar los resultados obtenidos al aplicar la metodología con el fin de descubrir las redes *onion* con conductas delictivas e información relevante para determinar el aporte dentro de un conjunto de casos previamente establecidos.

1.4 Límites de la tesis

- El alcance de esta metodología se centrará en la red privada TOR, dejando fuera otras redes privadas dentro del espacio de la Deep Web como i2p, Freenet, Ares, Emule, entre muchas otras, sin que esto signifique que los conceptos de esta metodología no puedan aplicarse a estas otras redes.
- El alcance de la información recopilada y obtenida puede ser aplicable para las agencias de seguridad de investigación de delitos cibernéticos en México.
- El marco legal en el que se estudiará la información recopilada de estos delitos cibernéticos será la legislación en México.

- Se tienen limitaciones en cuanto a uso de métodos de autenticación, captchas y adquisición de criptomonedas para poder acceder a ciertos sitios e información dentro de la red privada virtual de TOR.
- Se tienen limitaciones en cuanto a los diversos idiomas que se utilizan en TOR, se priorizará la información recopilada en inglés y español.

1.5 Organización de la tesis

El capítulo 1 lo conforma el protocolo del trabajo a desarrollar como son sus objetivos, problemáticas, y límites en el desarrollo del mismo.
El Capítulo 2 sirve como soporte teórico a algunos conceptos utilizados a lo largo del documento.
El Capítulo 3 es un resumen de los trabajos que pertenecen al estado del arte de los Delitos Cibernéticos, enfoques computacionales y técnicas desarrolladas y validadas en la investigación de la Dark Web, recolección automática de texto, y clasificación de datos en Deep Web.
El Capítulo 4 presenta la metodología de identificación y clasificación de Delitos Cibernéticos en México en la Red TOR, empleando el Cálculo de similitud de Jaccard, así como procedimientos y herramientas empleadas para obtener los resultados deseados.
El Capítulo 5 presenta los resultados obtenidos con respecto a la metodología, esta metodología se enseñó en los meses de marzo y abril de 2019 en la materia de Inteligencia de Fuentes Abiertas de la Especialidad de Inteligencia Policial de la entonces Policía Federal de México, resultando como trabajo final una página web con acceso restringido, la cual contiene la clasificación de conductas delictivas y no delictivas identificadas en la red privada de TOR, esto a través del estudio de Ontología de Delitos de Delincuencia Organizada en México, posteriormente está se implementó en el área de Prevención de Delitos Electrónicos de la Dirección General Científica de la Guardia Nacional para consulta y gestión en investigaciones y ciberpatrullaje. Por último se muestran dos páginas web, la primera contiene el resultado de lo recopilado de páginas en TOR a principios de 2022 y permite realizar filtros por Tipo y Subtipo, Idioma y por fechas de

alta de la página en el sistema y última consulta de la misma, así como los datos procesados de la página en cuestión, la segunda página web muestra estadísticas con el número total de redes onion, la frecuencia con que se repiten las redes onion, el número de redes por idioma, el estatus de código de estado de las páginas, y la frecuencia de los títulos de las redes onion, de tal forma que aporten nuevo conocimiento a sus investigaciones en casos reales.

Las conclusiones son expuestas en el Capítulo 6.

2. Marco Teórico

En este capítulo se describen conceptos relacionados a los delitos cibernéticos, clasificación de Internet, clasificación automática de textos, ontologías y redes privadas virtuales. Esto con el fin de entender el desarrollo de la metodología propuesta en el capítulo 4.

2.1 Delitos Cibernéticos

El Centro de Investigación de Delitos Informáticos define el ***delito cibernético*** como "la comisión del delito utilizando la tecnología electrónica". Puede ser un robo de activos, una destrucción de activos o un medio para convertir un activo en una amenaza (por ejemplo, el malware denominado Ransomware que impide a los usuarios acceder a su sistema o a sus archivos personales y que exige el pago de un rescate para poder acceder de nuevo a ellos). El delito cibernético también puede permitir el robo de identidad (por ejemplo, datos de funcionarios públicos), acecho e intimidación. El Departamento de Seguridad Nacional de los EE.UU. también ha identificado amenazas de seguridad cibernética a los intereses nacionales y comerciales (Rechtman, 2017).

Según (Cassou, 2009), el delito informático, se entiende como toda aquella conducta ilícita susceptible de ser sancionada por el derecho penal, consistente en el uso indebido de cualquier medio informático. Agencias internacionales como la Organización para la Cooperación y el Desarrollo Económicos (OCDE), lo define como cualquier conducta, no ética o no autorizada, que involucra el procesamiento automático de datos y/o la transmisión de datos.

Conforme el artículo 21 de la Constitución Política de los Estados Unidos Mexicanos (Cámara de Diputados, 2012), la investigación de los deli-

tos corresponde al Ministerio Público y a las policías, las cuales actuarán bajo la conducción y mando de aquél en el ejercicio de esta función.

Debido a que una de las principales tareas en la investigación de una conducta delictiva es la búsqueda de información en Internet, y además al contar cada vez con una cantidad mayor de datos en este medio, se necesita de metodologías y herramientas que separen la información relevante de la no relevante en el marco de la investigación delictiva.

2.2 Internet

Según (BrightPlanet, 2013), el Internet se construye alrededor de páginas web que hacen referencia a otras páginas, sí se tiene una página web de destino que no tiene enlaces entrantes, se ha ocultado esa página y no puede ser encontrada por usuarios o motores de búsqueda (no está publicada con un link). Un ejemplo de esto sería una publicación de un blog que aún no se ha indexado. La publicación del blog puede existir en el Internet público, pero a menos que se conozca la URL exacta, nunca se encontrará.

Deep Web, Surface Web y Dark Web

Partiendo de las acepciones de (BrightPlanet, 2013), podemos entender la Deep Web, Surface Web y Dark Web de la siguiente forma:

La ***Deep Web*** es una parte de Internet no accesible a los motores de búsqueda de rastreo de enlaces como Google, Yahoo y Bing. Una forma típica en que un usuario puede acceder a esta parte de Internet es escribiendo una consulta dirigida en un formulario de búsqueda web, recuperando así el contenido de una base de datos que no está enlazada a los buscadores convencionales. En términos sencillos, la forma de acceder a la Deep Web es realizando una búsqueda pormenorizada dentro de un sitio web en particular.

La ***Surface Web*** es otra parte de Internet que se puede hallar a través de técnicas de rastreo de enlaces, conocida como Link-crawling, que significa que los datos enlazados se pueden encontrar a través de un hipervínculo desde la página principal de un dominio y el buscador puede extraer estos datos.

La ***Dark Web*** es una parte de la World Wide Web que necesita un tipo especial de software para acceder y se refiere específicamente a una colección de sitios Web que existe en una red cifrada a la que no se puede acceder mediante los motores de búsqueda tradicionales o incluso que visitan los navegadores Web tradicionales. Una vez que esté dentro de la Dark Web, se puede acceder a los sitios Web y otros servicios a través de un navegador de la misma manera que en una Web tradicional. Sin embargo, hay algunos sitios que están ocultos de manera efectiva, lo que significa que tradicionalmente no han sido indexados por un motor de búsqueda y, por lo tanto, solo se puede acceder a dichos sitios si conoce la dirección del sitio (Rafiuddin, 2017).

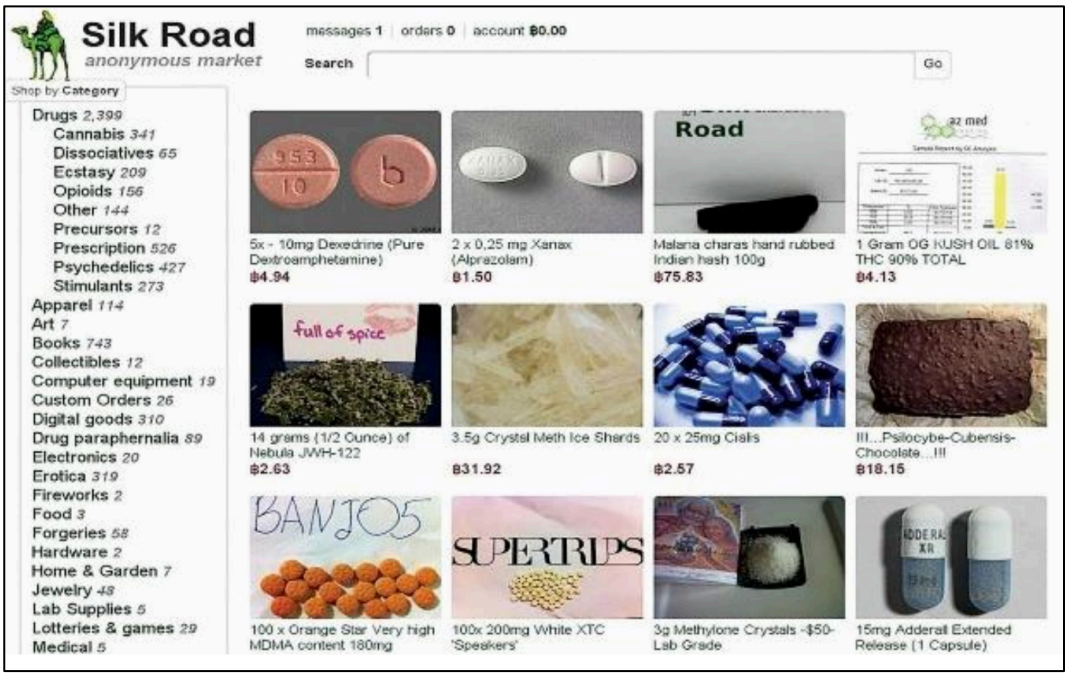


Figura 2.1. Sitio oculto de la Dark Web (Rafiuddin, 2017).

(Balduzzi M.) Define lo siguiente:

Deep Web es cualquier contenido de Internet que, por diversas razones, no puede ser o no está indexado por los motores de búsqueda como Google. Esta definición incluye páginas web dinámicas, sitios bloqueados (como aquellos en los que necesita responder a un CAPTCHA para acceder), sitios desvinculados, sitios privados (como aquellos que requieren credenciales de inicio de sesión), contenido no HTML / contextual / Acceso a las redes.

Las redes de acceso limitado cubren sitios con nombres de dominio que han sido registrados en dominioos de alta descentralización del Sistema de Dominio de Nombres (DNS) que no son administrados por la Corporación de Internet para Nombres y Números Asignados (ICANN), como dominios *.BIT*, sitios que se ejecutan en DNS estándar pero tienen dominios de nivel superior no estándar, y finalmente en darknets (páginas web y servicios a los que no se puede acceder a través de los motores de búsqueda tradicionales). Las darknets son sitios alojados en la infraestructura que requiere software específico como TOR antes de que se pueda acceder. Gran parte del interés público en la Red Profunda radica en las actividades que ocurren dentro de las Darknets.

Una persona inteligente que compra medicamentos de drogas recreativos en línea no querrá escribir palabras clave en un navegador normal. Él / ella requerirá navegar en internet de forma anónima, utilizando una infraestructura que nunca llevará a las partes interesadas a su dirección IP o ubicación física. Los vendedores de drogas también no quieren instalarse en ubicaciones en línea donde la policía pueda determinar fácilmente, por ejemplo, quién registró ese dominio o dónde está la ubicación de la dirección IP del sitio en el mundo real.

Hay muchas otras razones aparte de comprar drogas en las que la gente quisiera permanecer anónima, o para fijar los sitios que no podían ser

remontados a una localización o una entidad física. La gente que quiere proteger sus comunicaciones de la vigilancia del gobierno puede requerir la cobertura de darknets. Los denunciantes pueden querer compartir una gran cantidad de información privilegiada a los periodistas, pero no quieren dejar rastro en papel. Los disidentes en regímenes restrictivos pueden necesitar el anonimato para permitir que el mundo sepa lo que está sucediendo en su país.

Pero en el otro lado de la moneda, la gente que quiere tramar un asesinato contra un objetivo de alto perfil utilizando Internet, querrá hacer uso de un método que garantice que no se pueda rastrear la ubicación. Otros servicios ilegales como la venta de documentos como pasaportes y tarjetas de crédito también requerirán una infraestructura que garantice el anonimato. Lo mismo podría decirse de las personas que tienen información personal de otras personas como direcciones y datos de contacto.

Cuando se habla de Deep Web, es inevitable que la frase "Clear Web" o “Surface Web” aparezca. Es exactamente lo opuesto a la Web profunda: la parte de Internet que puede ser indexada por los motores de búsqueda convencionales y accesibles a través de navegadores web estándar sin necesidad de software y configuraciones especiales.

Hay mucha confusión entre los dos espacios. Sin embargo, la Dark Web no es la Deep Web, la primera es sólo una parte de la Deep Web. La Dark Web se basa en Darknets, redes en las que se realizan conexiones entre pares de confianza. Algunos ejemplos de los sistemas Dark Web incluyen TOR y el Invisible Internet Project (I2P).

2.3 Redes privadas virtuales (TOR)

Las redes privadas virtuales son otro aspecto de la *Deep Web,* que existe dentro del Internet público, y a menudo requieren software adicional para accederlas. El Proyecto “**T**he **O**nion **R**outer” (TOR), su objetivo es

tener una forma de usar Internet con la mayor privacidad posible, y la idea era enrutar el tráfico a través de múltiples servidores y cifrarlo en cada paso del camino. Oculto dentro de la red pública está esta red privada de contenido diferente y a la que sólo se puede acceder mediante el navegador de TOR.

Mientras que la libertad personal y la privacidad son objetivos admirables de la red TOR, la capacidad de moverse en Internet con anonimato completo nutre una plataforma madura para lo que se consideran actividades ilegales en algunos países, incluyendo:

- Mercados de sustancias controladas
- Armerías vendiendo diferentes tipos de armas
- Pornografía infantil
- Fugas no autorizadas de información confidencial
- Lavado de dinero
- Infracción de copyright
- Fraude de tarjetas de crédito y robo de identidad

En 2001 se estimó que la *Deep Web* contenía aproximadamente 3 millones de dominios existentes y era de 400 a 500 veces el tamaño de la *Surface Web.*

Las compañías de motores de búsqueda desarrollaron sistemas capaces de indexar rápidamente millones de páginas web en un corto período de tiempo, permitiendo así a los usuarios buscar con precisión el índice asimilado. Los motores de búsqueda no encuentran o almacenan todo el contenido en una página web, simplemente llevan a la ubicación de un contenido. Esta falta de retención de datos permite a los motores de búsqueda obtener la información mínima relevante sobre cada página web individual.

Normalmente, los motores de búsqueda almacenan las palabras más frecuentemente mencionadas, las ubicaciones de esas palabras y cualquier metadato (título de la página web, URL de la página web, palabras clave, etc.) al indexar páginas web. La cantidad de datos almacenados de cada página es una diferencia crucial entre los motores de búsqueda y los recolectores (harvesters).

Los recolectores extraen cada palabra cada vez que acceden a una página web, teniendo así ventajas de capacidades analíticas y almacenar versiones de páginas web.

2.4 Mecanismos para descubrir redes onion

Existen varias alternativas con las cuales podemos obtener nombres de redes con dominio onion, desde las más tradicionales como escribir la palabra “onion” en un buscador tradicional como Google, hasta usar servicios dedicados a la obtención de estos sitios, sin embargo existen ventajas y desventajas en estas alternativas, uno de los principales problemas de utilizar los buscadores tradicionales se debe a que la información que se recupera en la mayoría de las veces ya no es actual y es obsoleta, esto es, que ya no responden estos sitios onion dentro de la red Tor, mientras que el uso de servicios requiere de instalación y configuración de herramientas por lo general en un ambiente Linux, pero eso al final da mejor resultado. Algunas de estas alternativas son las siguientes:

Ahmia, es un sitio Web tradicional (https://ahmia.fi/address/) que busca servicios ocultos en la red Tor.

OnionScan, es una herramienta gratuita y de código abierto para investigar redes onion dentro de la Dark Web. Ayuda a los investigadores a supervisar y rastrear sitios web obscuros.

En la figura 2.2 inciso a) se muestra el laboratorio de correlación OnionScan que ofrece una forma de descubrir las relaciones entre diferentes sitios onion. La búsqueda da como resultado una página que muestra todo tipo de correlaciones detectadas por OnionScan. Este laboratorio permite etiquetar los resultados de búsqueda incluidos los resultados de otras etiquetas y buscar todas las páginas etiquetadas, Figura 2.2 inciso b).

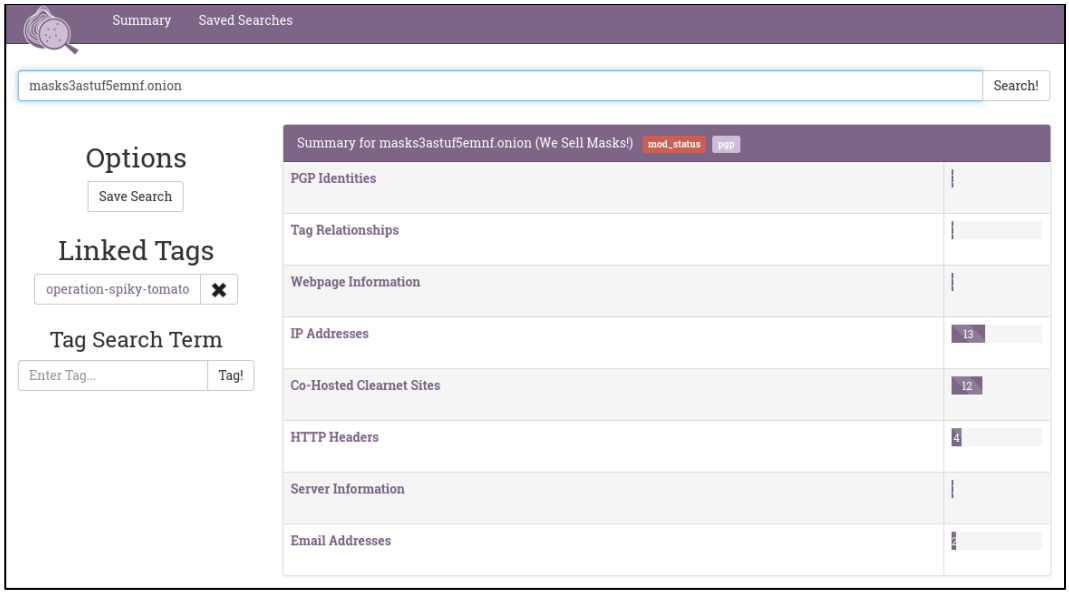


Figura 2.2 a) Búsqueda de correlación de una red onion

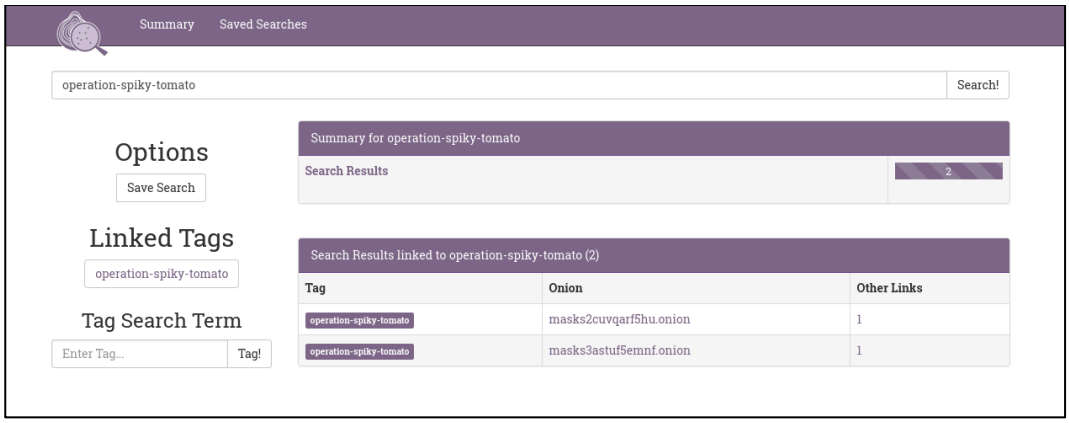


Figura 2.2 b) Resultado de redes onion correlacionadas al etiquetado.

Hunchly, Es una empresa en Internet de servicios web, que ofrece herramientas para investigación (Figura 2.3).

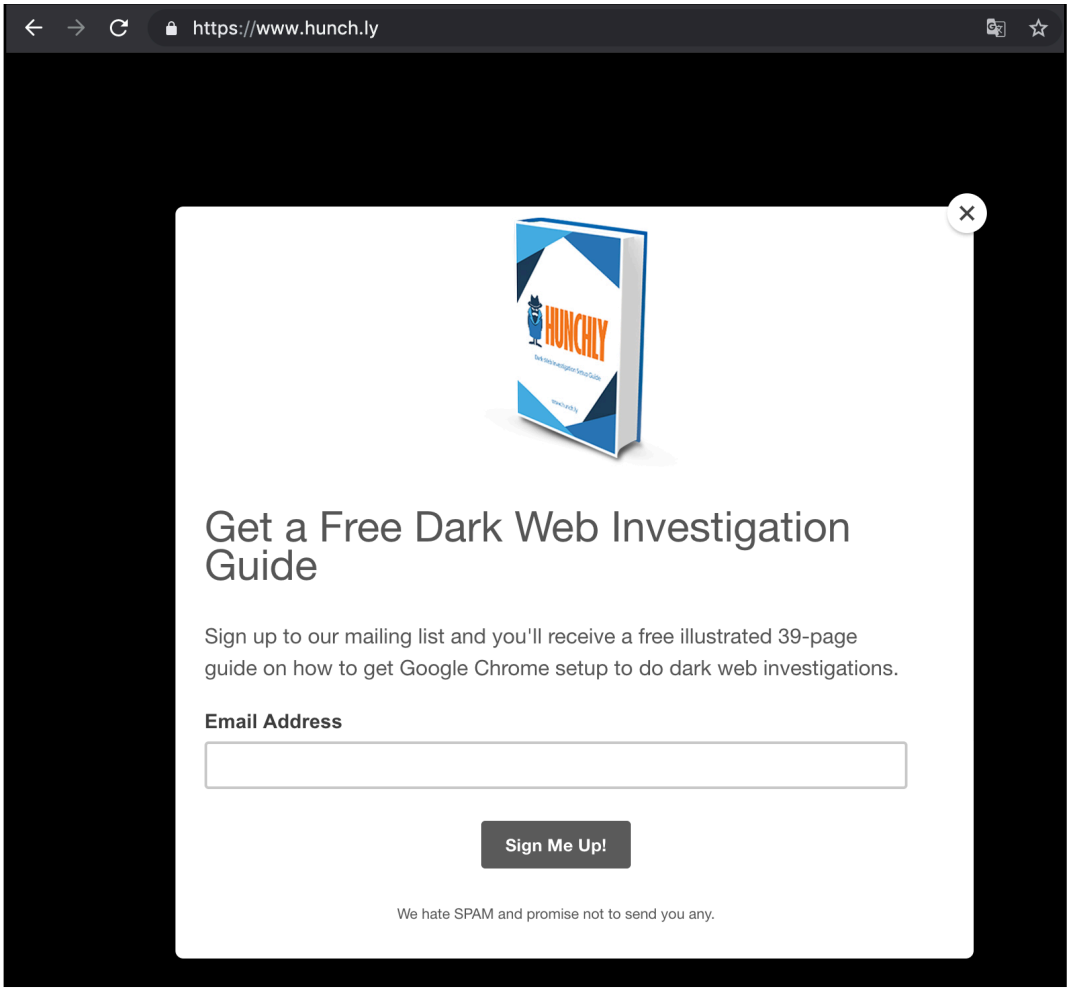


Figura 2.3 Guía de investigación en la Dark Web.

2.5 Recuperación de archivos en Internet

La recuperación de páginas o sitios Web en Internet se puede realizar de diferentes formas, desde dar click derecho sobre la página y elegir Guardar como… hasta utilizar programas dedicados a ello, esto de forma automática, e incluso semi-automatizada, con programas que requieren licencia de uso o gratuitos, a esto también se le llama “Crawler”. Algunos de ellos son:

GNU wget, Es un paquete de software gratuito para recuperar archivos mediante HTTP, HTTPS, FTP, FTPS (https://www.gnu.org/software/wget/).

Tor-browser-selenium, es un crawler que permite almacenar el tráfico de TOR, basado en Linux y Python

Los Crawlers se definen como "programas de software que atraviesan el espacio de información de la World Wide Web siguiendo los enlaces de hipertexto y recuperando los documentos web mediante el protocolo HTTP estándar" (Cheong, 1996). Son programas que pueden crear una colección local o índice de grandes volúmenes de páginas web (Cho y García-Molina 2000). Los Crawlers se pueden utilizar para los motores de búsqueda de uso general o para la construcción de la colección específica del dominio. Estos últimos se denominan rastreadores enfocados o temáticos ((Chakrabarti, 1999), (Pant, 2002), (Pant, 2002)).

Existe la necesidad de un Crawler enfocado, que pueda recopilar los foros de Dark Web. Muchos Crawlers de este tipo, se han centrado en la recopilación de páginas web estáticas en inglés desde la "superficie web". Un Crawler orientado al foro de Dark Web se enfrenta a varios desafíos de diseño. Una preocupación importante es la *accesibilidad*. Los foros web son dinámicos y a menudo requieren membresías. Son parte de la "hidden web" (Florescu et al., 1998, Raghavan y García-Molina 2001) la cual no es fácilmente accesible a través de la navegación web normal o del rastreo estándar. También hay consideraciones de minería web multilingüe. Más del 30% de la web está en idiomas diferentes al “inglés”(Chen y Chau 2003). Estos foros contienen archivos de texto estáticos y dinámicos, archivos de registro y varias formas de multimedia (por ejemplo, imágenes, archivos de audio y video). La recopilación de diversos tipos de contenido presenta muchos desafíos únicos que no se encuentran con el spidering estándar de archivos indexables (basados en texto).

2.6 Ontología

“Una ontología define los términos básicos y las relaciones que com-

prenden el vocabulario de un área o tópico, así como las reglas para combinar los términos y las relaciones que definen las extensiones al vocabulario". (R. Neches, 1991).

Se trabajó en la creación de una ontología usando la herramienta de Protégé con la finalidad de comprender las clases de las conductas delictivas que permitan definir las entidades a buscar dentro de la clasificación de delitos una vez recuperadas las redes onion.

Protégé

Protégé es un editor y marco de ontologías de código abierto y gratuito para construir sistemas inteligentes, fue desarrollado por el Centro de Investigación de Informática Bio médica de Stanford en la Escuela de Medicina de la Universidad de Stanford. que está respaldado por la subvención GM10331601 del Instituto Nacional de Ciencias Médicas Generales de los Institutos Nacionales de Salud de los Estados Unidos, (Musen, 2015).

2.7 Herramientas utilizadas

Algunas de las herramientas o paquetes necesarios para hacer posible la realización de ese trabajo se explican a continuación y se hará referencia a ellos posteriormente.

Privoxy

Privoxy es un proxy web sin almacenamiento en caché con capacidades de filtrado avanzadas para mejorar la privacidad, modificar los datos de la página web y los encabezados HTTP, controlar el acceso y eliminar anuncios y otra basura desagradable de Internet.

Privoxy tiene una configuración flexible y puede personalizarse para adaptarse a las necesidades y particularidades en investigaciones a realizar. Tiene aplicación tanto para sistemas autónomos como para redes

multiusuario. Se puede descargar en https://www.privoxy.org/.

Wget

GNU Wget es un paquete de software gratuito para recuperar archivos mediante HTTP, HTTPS, FTP y FTPS, los protocolos de Internet más utilizados. Es una herramienta de línea de comandos no interactiva, por lo que puede llamarse fácilmente desde scripts, terminales sin soporte de X-Windows, etc. Se puede descargar en https://www.gnu.org/software/wget/

Selektor

SelekTOR para Linux es un frontend de Interface Gráfica de Usuario GUI basado en Java, y de código abierto para Tor que se ejecuta en modo cliente, permite la conexión a Tor y poder elegir el nodo de salida para los navegadores que admiten el proxy del sistema mediante archivos PAC como es el caso del Browser Firefox, está licenciado bajo la GPL2. Se puede descargar en https://www.dazzleships.net/selektor-for-linux/

html2text

Paquete de linux que convierte archivos html a texto. Se puede encontrar cómo instalarlo en https://www.howtoinstall.co/es/ubuntu/xenial/html2text

dos2unix

dos2unix incluye utilidades para convertir archivos de texto con finales de línea de DOS o Mac a finales de línea de Unix y viceversa.

awk

La utilidad awk interpreta un lenguaje de programación de propósito especial que facilita el manejo de trabajos simples de reformateo de da-

tos, diseñado para procesar datos basados en texto, ya sean archivos o flujos de datos. Se puede consultar su manual en el sitio https://www.gnu.org/software/gawk/manual/gawk.html

Índice Jaccard

El índice de Jaccard, o el coeficiente de similitud de Jaccard, es una medida de similitud o diversidad en un conjunto. En los gráficos, se puede usar para encontrar qué vértices están cerca uno del otro, en función de sus vecinos comunes. Aunque fue inventado para la botánica, sus aplicaciones prácticas van desde el análisis de las comunidades en las redes sociales, hasta los algoritmos de aprendizaje automático y la ciberseguridad (Krawezik, G. P. et al, 2018)

$$\forall (u,v) \in \mathcal{G} : \mathcal{J}_{uv} = \frac{|U \cap V|}{|U \cup V|} = \frac{|U \cap V|}{|U| + |V| - |U \cap V|}$$

Figura 2.4. Fórmula del Índice de Jaccard.

Similitud Coseno

La similitud de coseno es una técnica popular para evaluar la similitud de datos de alta dimensión desde el ángulo de los vectores. Se explota ampliamente en el procesamiento de lenguaje natural (NLP) para calcular la similitud entre dos vectores de texto de alta dimensión. La idea básica es considerar dos objetos de datos como dos vectores en el espacio de datos “**m”** dimensional, y la similitud entre estos dos vectores se evalúa calculando el coseno del ángulo entre ellos. La similitud de coseno ha atraído una gran atención en la comunidad de investigación y se ha utilizado ampliamente en la minería de datos, como la clasificación y agrupación, la agrupación de fases, la atribución de documentos de patente, reconocimiento de patrones y diagnóstico médico (Gao, X., & Wu, S., 2018).

2.8 Aprendizaje Computacional

Cuando el ser humano adquiere conocimientos, habilidades, actitudes o valores a través del estudio, de la experiencia o la enseñanza, decimos que aprende. Este proceso es fácil para el humano, sin embargo, lograr que una máquina aprenda como lo hace el ser humano es una interrogante que existe desde los inicios de las computadoras. Actualmente no existe una máquina capaz de aprender de la misma manera que lo hace el hombre, sin embargo, se han creado algoritmos eficaces para algunas tareas de aprendizaje.

En términos muy generales, podemos decir que un programa aprende si el desempeño obtenido para realizar alguna tarea, mejora con la experiencia.

De manera formal. Se dice que un programa de computadora aprende de la experiencia "**E"** con respecto a una clase de tareas "**T"** y una medida de desempeño "**P"**, si su desempeño en las tareas "**T"**, medido con "**P"**, mejora con la experiencia "**E".**

Podemos decir entonces que el Aprendizaje Computacional estudia los procesos computacionales que hay detrás del aprendizaje en humanos y en las máquinas. Esta disciplina juega un papel importante en muchas áreas de la ciencia (Araujo, 2009).

2.9 Clasificación Automática de Textos.

La clasificación de textos surge de la necesidad de separar documentos de un tema o clasificación específica de un conjunto de documentos de diferentes temas. Al lograr clasificar los documentos por temas, la búsqueda de información se puede realizar de manera más sencilla.

Debido al elevado número de documentos que pueden pertenecer a una colección de documentos, principalmente en formato electrónico, realizar la clasificación en forma manual, provoca que la tarea sea complica-

da, costosa y que requiera mucho tiempo, por lo que surge la idea de hacerlo automáticamente.

Así es como surge el área de Clasificación Automática de Textos, en la cual se han utilizado diferentes métodos estadísticos y más recientemente técnicas de Aprendizaje Computacional.

El primer paso para realizar la tarea de Clasificación Automática de Textos utilizando técnicas de Aprendizaje Computacional, consiste en obtener los atributos que describan el texto a clasificar, así como transformarlos a una representación adecuada para ser utilizados por los algoritmos de Aprendizaje Computacional. A este paso previo se le llama extracción de características. En la siguiente sección se explica con mayor detalle cómo se realiza la extracción de características en la Clasificación Automática de Textos. Posteriormente se presentan los algoritmos más utilizados en el área de Clasificación Automática de Textos.

La extracción de características generalmente consiste en tres etapas: Pre-procesamiento, Indexado y Reducción de dimensionalidad.

El pre-procesamiento consiste fundamentalmente en eliminar aquellos elementos que generalmente no contienen información para la tarea de la clasificación. Consta de tres posibles fases básicas:

- *Eliminación de etiquetas*. Si los documentos utilizados contienen algún tipo de etiquetas o cabeceras (ej. etiquetas de html o xml), estás podrán ser removidas, debido a que en algunos casos no proporcionan información útil para la clasificación.

- *Eliminación de palabras vacías*. Las palabras vacías son palabras que son muy frecuentes y que por lo general no contienen información, por ejemplo: pronombres, preposiciones, conjunciones, artículos, etc.

- *Lematización de palabras*. Por lematización nos referimos al proceso de remover los sufijos para reducir una palabra a su lema o raíz. Por ejemplo, comprender, comprenderlo y comprendió tienen la raíz *comprend*.

(Araujo, 2009).

3. Trabajos Relacionados

Se realizó una revisión de investigaciones, artículos científicos y herramientas, relacionados con los temas de Delitos Cibernéticos, Dark Web, y Detección automática de texto en los últimos años, a continuación, se muestran algunos de ellos.

3.1 Estado del arte

(Chen H. , 2011) Presenta diez capítulos sobre enfoques computacionales y técnicas desarrolladas y validadas en la investigación de la Dark Web. Este proyecto de Dark Web de la Universidad de Arizona es un programa de investigación científica a largo plazo que tiene como objetivo estudiar y entender el fenómeno del terrorismo internacional (jihadista) a través de un enfoque computacional centrado en los datos. Su objetivo es recopilar "TODO" el contenido web generado por grupos terroristas internacionales, incluyendo sitios web, foros, salas de chat, blogs, sitios de redes sociales, videos, mundo virtual, etc. Desarrollaron minería de datos multilingües, minería de texto y web. Técnicas de minería para realizar análisis de enlaces, análisis de contenido, análisis de métricas web (sofisticación técnica), análisis de sentimientos, análisis de autoría y análisis de video. Los enfoques y métodos desarrollados en este proyecto contribuyen a avanzar en el campo de la Informática de Inteligencia y Seguridad (ISI). Estos avances ayudarán a las partes interesadas a realizar investigaciones sobre el terrorismo y a facilitar la seguridad y la paz internacionales.

Según (Yan Wang, Crawling ranked deep Web data sources, 2016):

“En la era de los grandes datos, la gran mayoría de los datos no provienen de la Web superficial, la Web que está interconectada por hipervínculos e indexada por la mayoría de los motores de búsqueda de propósito general.

En su lugar, el tesoro de datos valiosos a menudo reside en la Web profunda, la Web que se oculta detrás de las interfaces de consulta".

Dado que numerosas aplicaciones, como la integración de datos y los portales verticales, requieren datos Web profundos, se desarrollaron varios métodos de rastreo para recopilar exhaustivamente una fuente de datos Web profunda con el coste mínimo (o casi mínimo).

La mayoría de los métodos de rastreo existentes suponen que se devuelven todos los documentos coincidentes con las consultas. En la práctica, las fuentes de datos suelen devolver las primeras **k** coincidencias. Esto hace difícil la recolección exhaustiva de datos: los documentos altamente clasificados se devolverán varias veces, mientras que los documentos con baja clasificación tienen pocas posibilidades de ser devueltos.

En este artículo, descompusieron este problema en dos sub-problemas ortogonales, es decir, problemas de consulta y clasificación parcial, y propusieron un método de rastreo basado en frecuencia de documentos para superar el problema de clasificación parcial.

Lo racional del método es utilizar las consultas cuyas frecuencias de documentos están dentro del rango especificado para evitar el efecto del límite de retorno en el ranking de búsqueda positivo y reducir significativamente la dificultad de rastrear la fuente de datos clasificados. El método está ampliamente probado en una variedad de conjuntos de datos y se compara con dos métodos existentes. El resultado experimental demuestra que este método supera a los dos algoritmos en un 58% y 90% en promedio.

El Deep Web Analyzer (DeWA) ha sido diseñado con el objetivo de apoyar investigaciones en el rastreo de actores maliciosos, explorar nuevas amenazas y extraer datos significativos de Deep Web, por ejemplo:

nuevas campañas de malware.

DeWA consta de los 6 módulos siguientes (Figura 3.1):

1. Un módulo de recolección de datos, responsable de encontrar y almacenar nuevas URL de múltiples fuentes
2. Un Gateway Universal, que permite acceder a los recursos ocultos en darknets como TOR e I2P, y para resolver direcciones DNS personalizadas
3. Un módulo de *Scouting* de Páginas, responsable de rastrear las nuevas URLs recopiladas
4. Un módulo de Enriquecimiento de Datos que se encarga de integrar la información explorada con otras fuentes
5. Un módulo de almacenamiento e indexación, que pone a disposición los datos para su posterior análisis
6. Visualización y herramientas analíticas

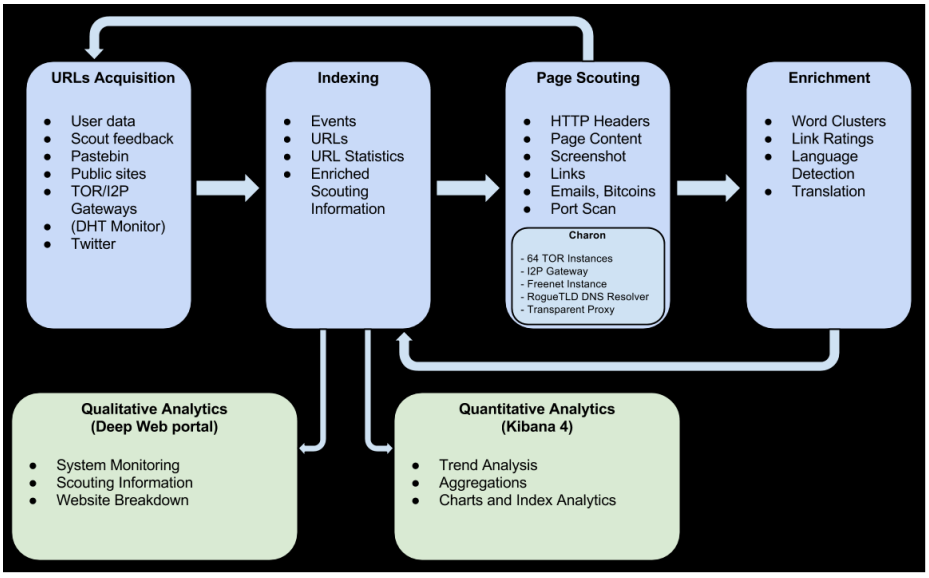


Figura 3.1. Fuente: Cybercrime in the Deep Web. Black Hat, EU, Amsterdam 2015. Balduzzi M., Ciancaglini V. (Trend Micro).

En el siguiente ejemplo, se ha agrupado 2 años de datos de acuerdo con el esquema de las URL (por ejemplo, http, https, ftp, ...). De todos los

dominios recogidos, casi 22.000 son (previsiblemente) asociados al protocolo http (s), siendo los datos que alojan la actividad principal. Pero si filtramos esos dominios, el resto muestra algunos datos interesantes, como se muestra en la figura 3.2:

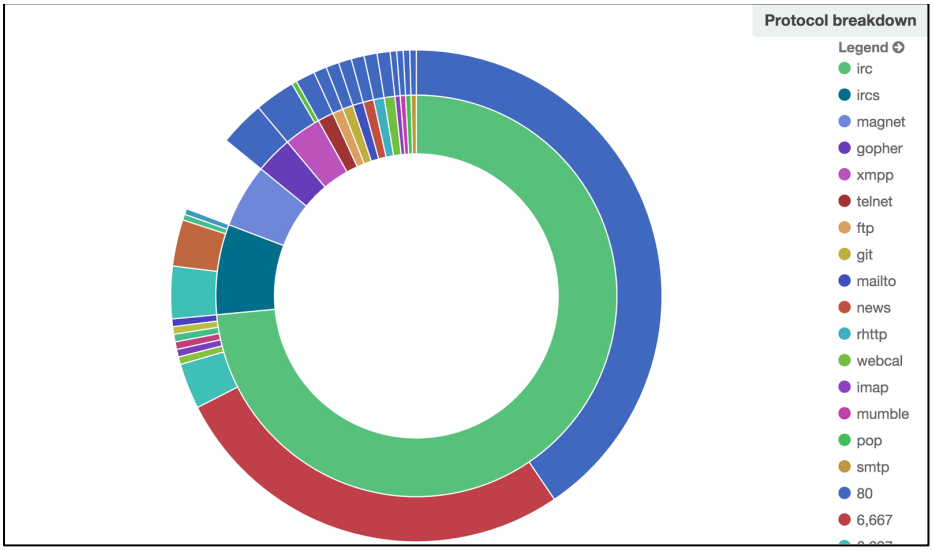


Figura 3.2. Fuente: Cybercrime in the Deep Web. Black Hat EU, Amsterdam 2015. Balduzzi M., Ciancaglini V. (Trend Micro).

Más de 100 dominios están de hecho alojando IRC (S): éstos son normalmente servidores de chat que se pueden utilizar como punto de encuentro para los agentes malévolos para negociar mercancías, o como canal de la comunicación para botnets (es un término que hace referencia a un conjunto o red de robots informáticos o bots, que se ejecutan de manera autónoma y automática. El artífice de la botnet puede controlar todos los ordenadores/servidores infectados de forma remota). El mismo concepto se aplica a los 7 dominios XMPP (es decir, IMs de tipo Jabber), que representan otro protocolo para servidores de chat que se ejecutan en TOR.

A continuación se muestran varios ejemplos de actividades maliciosas en la Deep Web:

En la figura 3.3 se muestra la venta de Pasaportes falsos en un sitio de la red onion:

Figura 3.3. Ciudadanía de los EEUU en venta por debajo de 6000 USD

USD http://xfnwyig7olypdq5r.onion/

La figura 3.4 muestra la venta de cuentas robadas

Figura 3.4. Cuentas robadas. No verificadas vendidas a granel - 80% válido o reemplazo ofrecido, http://3dbr5t4pygahedms.onion/

La figura 3.5 muestra la venta de réplicas de tarjetas de crédito:

Figura 3.5. Clonación de tarjetas de crédito

http://ccccrckysxxm6avu.onion/

Las figura 3.6 muestran la venta de servicios de alquiler de asesinos:

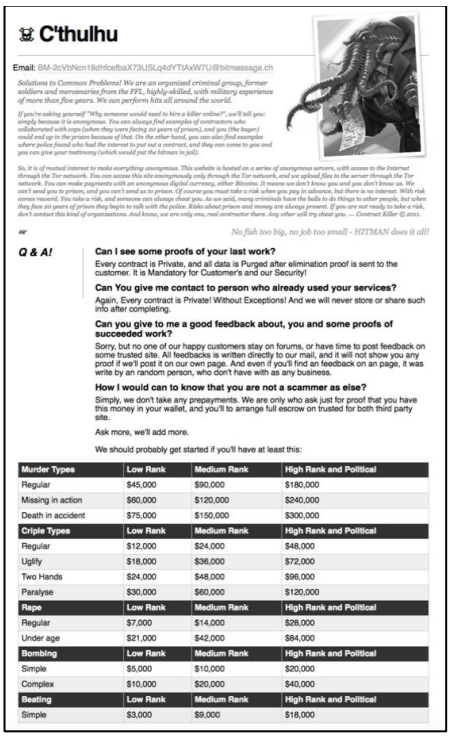


Figura 3.6. Cthulhu Currículum - Servicios de asesinato en alquiler

http://cthulhuuap7ch47k.onion

La figura 3.7 muestra un servicio de minado de bitcoins:

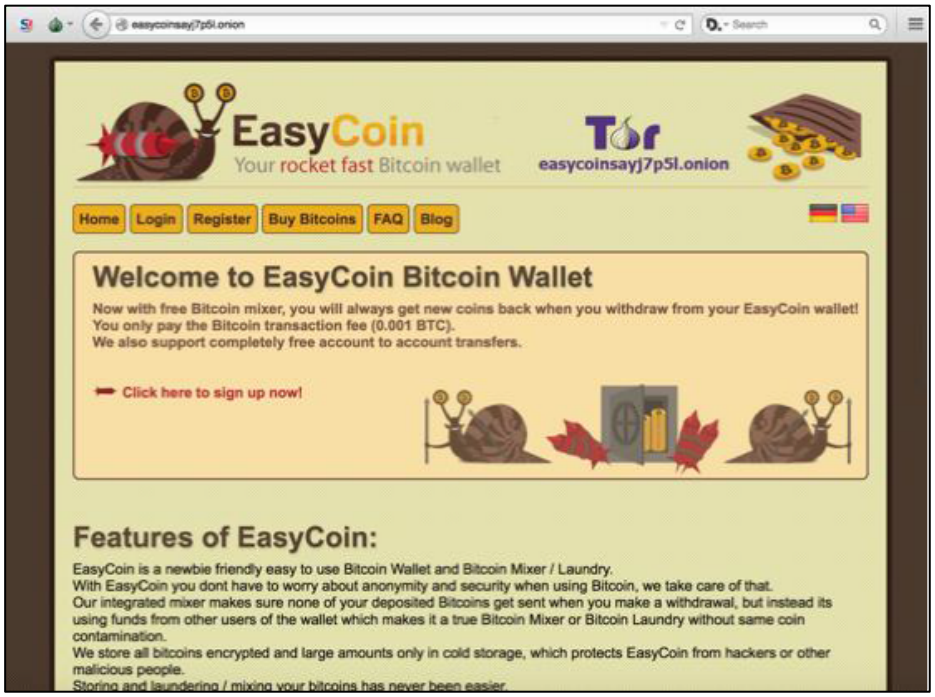


Figura 3.7. Servicio de minado EasyCoin - Bitcoin
http://easycoinsayj7p5l.onion

La figura 3.8 muestra la compra y venta de Bitcoins como moneda virtual:

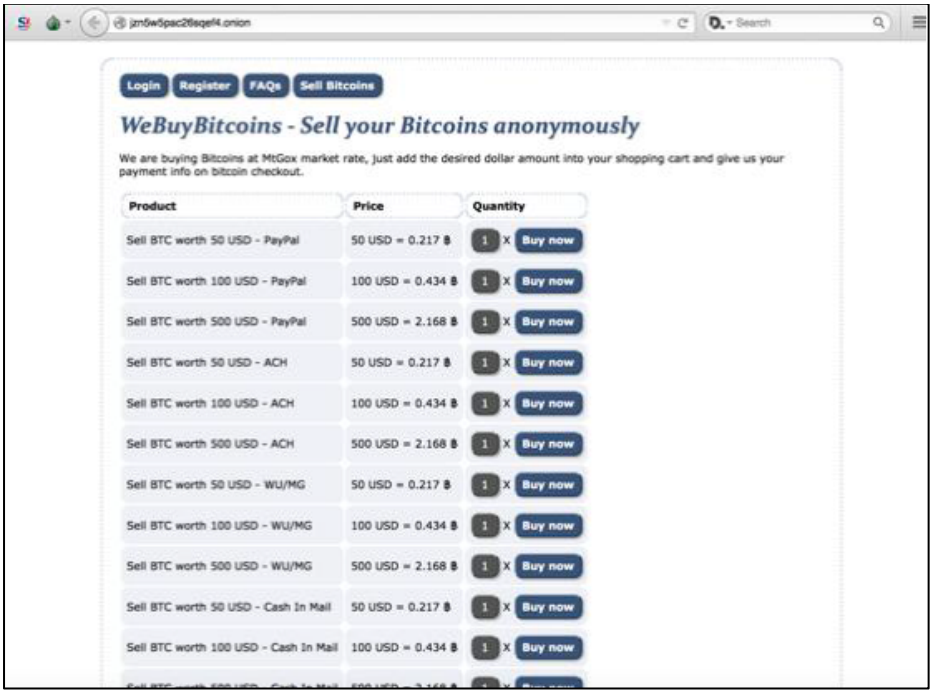


Figura 3.8. WeBuyBitcoins - Intercambio de Bitcoin para pagos en efectivo o electrónicos
http://jzn5w5pac26sqef4.onion

La figura 3.9 muestra la venta de billetes de dólar falsificados:

Figura 3.9. Comprar falsificación 20 USD por aproximadamente la mitad del precio del valor nominal

http://usjudr3c6ez6tesi.onion

La figura 3.10 muestra la venta de información confidencial de personas:

```
Barack Hussein Obama
AGE: 50
DOB: 08/04/1961 (August 4th 1961)
Born In: Honolulu, Hawaii
Married to Michelle Obama (Robinson)
Obama's Yahoo Email Address
bobama@yahoo.com - IP Used to sign in 71.191.175.122 - Arrlington, VA - Verizon
Internet.
Baracks Personal IP (IP of the Whitehouse?) 66.36.206.59 - Washington DC IP that was
signed into both emails.
Obama's AOL (Protected by AOL Security)
qdjdoo23@aol.com
Barack IP used to sign into that E-mail when he was in Rhode Island. 68.14.135.217 -
Cox Communications.
Court Records
Barack H Obama
Defendant
```

Figura 3.10 Aparente cuenta de correo electrónico personal de Barack Obama (no verificado)

http://cloudninetve7kme.onion

La figura 3.11 muestra la venta de drogas:

Figura 3.11. The Peoples Drug Store - venta de heroína, cocaína, éxtasis y más
http://newpdsuslmzqazvr.onion

La figura 3.12 muestra un buscador dentro de la Dark Web:

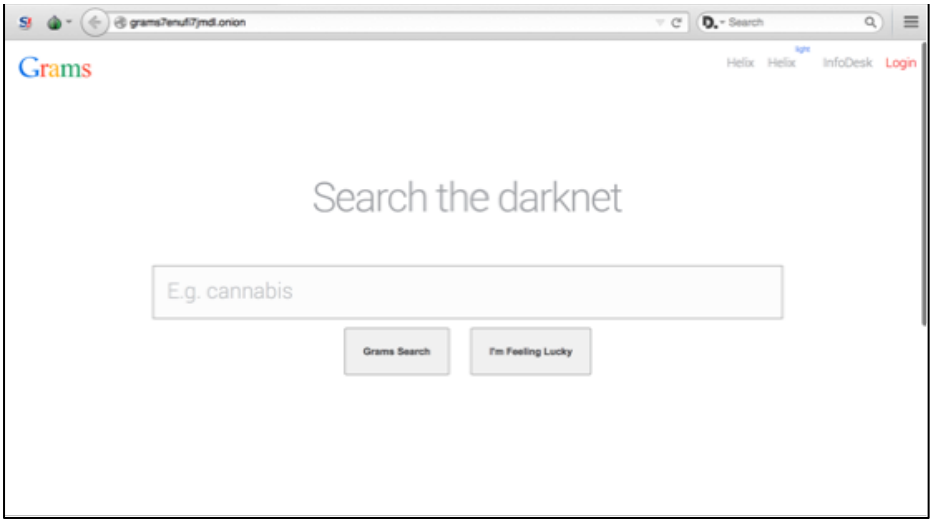


Figura 3.12. Grams - el motor de búsqueda Deep Web para la droga
http://grams7enufi7jmdl.onion

La figura 3.13 muestra programas de Comando y Control gestionados desde TOR:

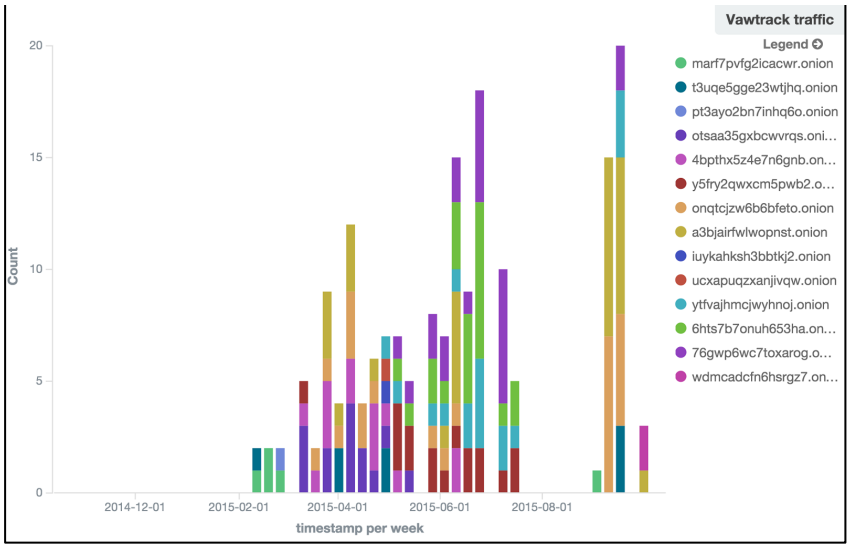


Figura 3.13. C & Cs identificados basados en TOR

La figura 3.14 muestra el virus Cryptolocker:

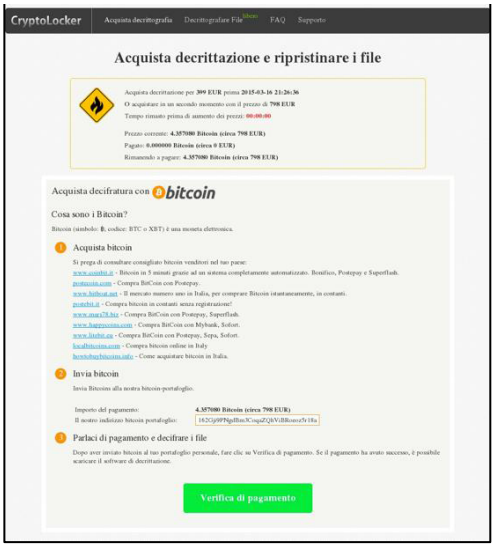


Figura 3.14. Cryptolocker C & C formateado automáticamente para una víctima en Taiwán e Italia. http://ndvgtf27xkhdvezr.onion

(Investigation) En el sitio de internet del FBI (Federal Bureau of Investigation) sobre el Centro de Quejas de Crimen en Internet se describen las siguientes conductas delictivas para la prevención del crimen del Internet:

- Fraude en las subastas
- Comprobar la falsificación de cajero
- Fraude de tarjeta de crédito
- Eliminación de deuda
- DHL / UPS
- Oportunidades de empleo / negocio
- Fraude de Servicios de depósito de garantía
- El robo de identidad
- La extorsión de Internet
- El fraude de inversión
- Loterías
- Carta de Nigeria o "419"
- Phishing / Spoofing
- Ponzi / Pirámide
- Reenvío
- Correo no deseado
- Receptor tercera Parte de los fondos

En 2013, el FBI comprometió una serie de servidores utilizados por los servicios de TOR hidden y los utilizó para entregar un *exploit* para anonimizar a los usuarios de la red TOR. Cuando el usuario visitó una de las páginas entrampadas con el Explorador TOR, el *exploit* abusó de una vulnerabilidad *use-after-free* de Firefox (CVE-2013-1690) con el fin de permitir la ejecución arbitraria de código. La vulnerabilidad fue parchada y lanzada por Mozilla a finales de junio de 2013. El objetivo del *payload* del *exploit* fue obtener la dirección MAC y el nombre de host del dispositivo víctima y enviar los datos a un servidor web controlado por el atacante, pasando por TOR. Ese mensaje también incluyó una identificación única proporcionada por la página entrampada para correlacionar a un usuario específico con una visita específica. El atacante entonces conocía la dirección IP pública, la dirección MAC y el nombre de host de cada usuario

que visitó esta página (Conti, M., Crane, S., Frassetto, T., Homescu, A., Koppen, G., Larsen, P., ... & Sadeghi, A. R, 2016).

(Christos Iliou, George Kalpakis, Theodora Tsikrika, Stefanos Vrochidis and Ioannis Kompatsiaris, 2017) Propone un framework de crawler enfocado para descubrimiento de recursos de un tópico dado dentro de la Surface o la Dark Web. Su propósito es navegar a través de la Surface Web y varias darknets presentes en la Dark Web (como Tor, I2P y Freenet), durante un solo rastreo investiga 11 métodos de selección de hipervínculos, incluyendo una estrategia de combinación lineal dinámica de un enlace base y un clasificador de páginas Web padre. Es aplicado para descubrir procedimientos que produzcan explosivos caseros. Los experimentos de evaluación indican la efectividad del rastreador centrado propuesto tanto para Surface como para Dark Web.

(Kirkpatrick, 2017) Explora el uso de criptomonedas en la Dark Web, que permite transacciones ilegales o inmorales. Se discute cuántas de las actividades en la web oscura incluyen la venta de drogas, la prostitución, la pornografía ilegal, la trata de personas, la investigación de Silk Road y la financiación de ataques cibernéticos.

(Denic, 2017) Analizó los servicios web oscuros que explotan las organizaciones terroristas. El enfoque del estudio fue en la infraestructura de red Tor y los servicios ocultos. Se presentó un estudio, explicación y ejemplos de actividades terroristas. Se presentaron los logros exitosos del gobierno en la anonimización de administradores y clientes de servicios y la técnica de cómo se llevaron a cabo esas operaciones. Para todas esas operaciones exitosas, los delincuentes han usado elementos de red en varios países. La operación más exitosa en la Dark Web fue Operación Onymous. En esta operación, los objetivos eran sitios ilícitos y administradores del sitio. El actor clave de esta operación fue Estados Unidos junto con 16 países europeos. Esta coalición cerró más de 400 servicios

ocultos. La Operación “Pacifier” ha demostrado que a veces los servicios ocultos no anónimos se presentan como una oportunidad para atrapar a delincuentes u objetivos terroristas de alto valor. La acción cibernética bien diseñada con la utilización adecuada de herramientas cibernéticas como NIT ha dado lugar a cargos contra los delincuentes. La incautación de dinero de sitios ilícitos ha demostrado que, sin los ingresos, los infractores no podían tener éxito en sus actividades. En resumen, el gobierno a tono con los socios de la coalición puede detectar, disuadir e interrumpir las amenazas que surgen de la red oscura. Al realizar actividades cibernéticas ofensivas en un área geográfica amplia, se puede reducir la financiación de la red oscura y se puede degradar la infraestructura junto con los servicios ocultos.

Dark Web. La democracia tiene más que temer del propio gobierno y la industria de vigilancia global que de Silk Road o Tor. De hecho, es la web visible, no su equivalente invisible, la que produce distorsiones sociales tales como noticias falsas, hechos alterados, pos-verdades, mimetismos, engaño público, distorsión de mensajes y propagación de rumores. El punto a tener en cuenta es que el estado profundo, para preservar su propia invisibilidad y proteger su base de poder, es necesariamente parcial e inconstante, como lo descubrió recientemente el asesor de seguridad nacional Mr. Michael Flynn.

El caso de Silk Road es un claro recordatorio del esfuerzo continuo del gobierno estadounidense para subvertir los servicios de anonimato. Hace casi cuatro años, el servicio de correo cifrado Lavabit se vio obligado a dejar de operar después de que las autoridades exigieran que revelara las claves SSL del cliente. El anonimato amenaza el autoritarismo y su ejercicio de control. Cuando los grandes y poderosos tipos de gobierno hablan de la red oscura, enfatizan la criminalidad; cuando tecnólogos y libertarios civiles hablan de eso, el énfasis está en la libre expresión. La diferencia puede explicarse ideológicamente.

(Hurlburt, 2017) La raíz de muchas, si no la mayoría, de las amenazas a la ciberseguridad no está en el borde de Internet, sino en su interior, en la Dark Web. Sin embargo, la Dark Web es cada vez más difícil de descifrar a medida que las técnicas de privacidad y encriptación se vuelven más sofisticadas. Los sitios serán mucho menos visibles y solo se podrá acceder mediante invitación. Una vez que la cripto moneda elegida en Dark Web (Bitcoin), está siendo rápidamente reemplazada por Monero, que ofrece mecanismos sigilosos que evitan el rastreo indirecto de aquellos que realizan transacciones, una vulnerabilidad que ha afectado al bitcoin. Las mismas herramientas de código abierto promocionadas por los defensores de la privacidad para proteger los datos personales y eludir la censura y la vigilancia del gobierno también están alimentando la actividad delictiva generalizada.

Dada la sofisticada infraestructura de Dark Web y las capacidades técnicas superiores de muchos de sus habitantes, es poco probable que las técnicas forenses tradicionales tengan un efecto sustancial o duradero. Sin embargo, las nuevas herramientas de aprendizaje automático, minería de datos y análisis podrán convertirse en herramientas en la lucha contra el delito cibernético.

Debido a que Internet es una gran red de redes con billones de nodos interconectados, se pueden descubrir patrones indicativos de actividad potencialmente dañina o ilegal (por ejemplo, botnets, distribución de malware e intercambio de archivos punto a punto) a través de algoritmos avanzados y software de visualización. Estas herramientas no solo pueden ayudar a identificar y deshabilitar varios sitios de Dark Net, sino también a proporcionar evidencia legal contra delincuentes identificados. Los organismos encargados de hacer cumplir la ley a menudo emplean técnicas secretas y controvertidas para derribar sitios ilegales y arrestar a sus operadores, en algunos casos, previenen el enjuiciamiento para evitar

revelar la tecnología que utilizaron. Pero la inteligencia artificial y los análisis de big data ofrecen un método alternativo para descubrir a los autores de estas conductas ilícitas, así como a los defensores de las libertades civiles.

(Ghappour, 2017) Ghappour comienza a escribir sobre el uso de herramientas de hacking por parte de las fuerzas del orden público para perseguir potenciales cibercriminales que han anonimizado sus comunicaciones en la Dark Web, esto presenta un inminente punto álgido entre el procedimiento penal y el derecho internacional. Los actores cibercriminales que usan la Dark Web (por ejemplo, para cometer conductas ilícitas o evadir a las autoridades) oscurecen las huellas digitales dejadas atrás con terceros, volviendo obsoletos los métodos de vigilancia existentes. En respuesta, las fuerzas del orden público han implementado técnicas de hackeo que implementan software de vigilancia a través de Internet para acceder y controlar directamente los dispositivos de los delincuentes. La realidad práctica de las tecnologías subyacentes hace que sea inevitable que las computadoras ubicadas en el extranjero estén sujetas a "búsquedas" e "incautaciones" remotas. El resultado puede ser la mayor expansión extraterritorial de la jurisdicción de ejecución en la historia de la aplicación de la ley de los EE. UU. (lo cual por las leyes internacionales resulta cuestionable).

Este artículo examina cómo el uso del gobierno de las herramientas de hackeo en la Dark Web interrumpe profundamente la arquitectura legal en la que descansan las investigaciones criminales transfronterizas. Estas ciber-operaciones en el exterior plantean preguntas cada vez más difíciles sobre quién puede autorizar estas actividades, dónde pueden ser desplegadas y contra quién pueden ser legalmente ejecutadas. Las reglas del procedimiento penal no regulan el hackeo de las fuerzas del orden público, porque permiten que estas decisiones críticas las tomen los funcionarios del rango correspondiente a pesar de las implicaciones potencialmen-

te perturbadoras en las relaciones exteriores. Este artículo describe un marco regulatorio que reasigna la toma de decisiones a los actores institucionales que mejor se adaptan para determinar la política exterior estadounidense y evita sacrificar la capacidad de las fuerzas del orden público para identificar y localizar probbles cibercriminales que se han refugiado en la Dark Web.

(WIECZNER, 2017) Sean Everett en marzo de 2017, vendió todas sus acciones, incluidas Apple y Amazon, y utilizó una parte de las ganancias para comprar Bitcoin y Ethereum en un sitio llamado Coinbase. La decisión hizo que Everett, el CEO de la empresa de inteligencia artificial Prome, le generó que casi instantáneamente fuera más rico, ya que el valor de las monedas basadas en blockchain se incrementó exponencialmente durante las siguientes semanas. Pero luego, mientras él estaba paseando al perro después de las 10 p.m. el miércoles, 17 de mayo, Everett recibió una llamada. Era T-Mobile, llamándolo para confirmar que estaba cambiando su número de teléfono a un dispositivo diferente. Fue un movimiento sospechoso que Everett ciertamente no había solicitado. Pero incluso mientras le suplicaba al agente que bloqueara el movimiento, ya era demasiado tarde. Menos de cinco minutos después, el servicio celular de Everett se cortó abruptamente, y mientras corría hacia su computadora, se vio robado en tiempo real. Una gran cantidad de notificaciones por correo electrónico confirmaron que alguien había tomado el control de su cuenta principal de Gmail, y luego ingresó en su "billetera" de Coinbase. Habían llegado con la ayuda de su número de teléfono cambiado: la cuenta de Everett requería que iniciara sesión con un código de autenticación de dos factores enviado por mensaje de texto, como una segunda salvaguarda, y ahora el texto había ido directamente al ladrón. Le tomó solo dos minutos al atacante limpiar a Everett de lo que entonces era una cantidad de monedas digitales de unos miles de dólares. Desde la perspectiva de Everett, el atraco aún más doloroso fue lo que vino después: el precio de Ethereum se cuadruplicó en las siguientes tres semanas. Había

alcanzado su máximo histórico de $ 400 dólares. Bitcoin, mientras tanto, había aumentado $ 3,000 dólares por primera vez un día antes.

Por otra parte, la mayor sorpresa para Everett, y resultó ser, para muchos otros entusiastas de Bitcoin, fue que el robo ocurrió en Coinbase. La Coinbase de San Francisco, el mayor mercado de cripto monedas del mundo, es una de las pocas empresas similares cuyos cofres nunca han sido pirateados, una distinción que tiene un peso adicional en el ámbito de la cadena de bloques, donde varias infracciones costosas han sido noticia mundial. Casi cualquier inversionista temprano con quien hablas perdió dinero, en Mt. Gox, un intercambio colapsó en 2014 después de que piratas informáticos saquearon casi $500 millones en Bitcoin. El verano pasado, los ladrones obtuvieron $72 millones del intercambio de cifrado de Hong Kong Bitfinex de un solo golpe.

Pero los hackers nunca han violado la fortaleza virtual de Coinbase, y esa impenetrabilidad le ha ganado una reputación como el lugar más seguro para comprar Bitcoin, ayudándolo a atraer a más de 9 millones de clientes que almacenan al menos 3 mil millones en cripto monedas allí, y que han intercambiado $25 mil millones hasta la fecha en su corretaje minorista, así como su intercambio institucional. Coinbase, en cinco años, recaudó 100 millones en nuevos fondos, valorando a la compañía en $1,600 millones, convirtiéndose en el primer "unicornio" de la industria de la cadena de bloques.

En cada caso, llega la misma realización ciega, poniendo en foco la paradoja inherente de blockchain. La fuerza por excelencia que distingue a la cripto moneda del dinero tradicional -que las transacciones son instantáneas e irreversibles- es también su defecto fatal. "Uno de los motivos [de Bitcoin] para la existencia es que es resistente a la censura", dice Tom Robinson, cofundador y director de datos de Elliptic, una firma de inteligencia de blockchain con sede en Londres. Eso significa que nadie, ni

siquiera un gobierno o un banco central, puede evitar que se produzca una transacción de moneda digital. Y, por lo tanto, las protecciones contra el fraude que los depositantes bancarios tradicionales confían en su mayoría no están disponibles. "Cualquier tipo de carga y reversibilidad sería la antítesis de lo que Bitcoin fue creado para lograr", dice Robinson.

Esa es una razón por la que, cuando los delincuentes quieren hacer un atraco, cada vez más eligen cripto monedas que dólares reales. En 2016, se reportaron $28 millones de dólares en pérdidas por delitos que involucran monedas virtuales al Centro de Quejas contra el Crimen por Internet del FBI, más del triple del total que en el año 2015. Y esa cifra se basa en gran medida en informes voluntarios de víctimas individuales. No incluye robos a gran escala de intercambios como el truco de Bitfinex, por lo que es probable que esten subestimados los verdaderos daños en muchos órdenes de magnitud.

(Rechtman, 2017) La sociedad digital moderna ha dado lugar a innumerables formas de cibercrimen, especialmente en los últimos años. Para contrarrestar esto, las compañías de seguros han comenzado a ofrecer seguros para proteger específicamente contra la amenaza de ataques digitales. Las pólizas disponibles y las primas y cobertura relacionadas aún están en desarrollo. El autor detalla los conceptos e inquietudes relacionados con el delito cibernético y recomienda pasos para que las empresas consideren qué seguro comprar.

En general, la mitigación de riesgos se divide en cuatro categorías: aceptar, compartir, reducir o evitar. El seguro comparte el riesgo con la aseguradora; sin embargo, debido a que este es un cálculo de probabilidad donde la frecuencia y el impacto son total o parcialmente desconocidos, los suscriptores -cuya responsabilidad es evaluar los riesgos que se asumen- son propensos a adoptar un enfoque conservador y suponer que

la frecuencia y el impacto son altos. Hacer lo contrario podría exponer a la compañía de seguros a una alta tasa de grandes reclamos.

(Altayar, 2017) Con el uso creciente de las TIC en diferentes áreas de la vida contemporánea, el acceso generalizado a Internet y el uso de las redes sociales, los individuos y las organizaciones se enfrentan a constantes amenazas y desafíos derivados de las malas consecuencias de los ciberdelitos. La investigación muestra que la incidencia del cibercrimen está aumentando. En respuesta a estas amenazas y desafíos, las agencias gubernamentales y policiales de diferentes países del mundo han promulgado leyes contra el delito cibernético. El objetivo de este documento es comparar las leyes existentes contra el delito cibernético en los países del Consejo de Cooperación del Golfo (CCG). Adopta un enfoque de investigación comparativa. Según los resultados, aunque estos países comparten tradiciones comunes y valores islámicos, legales, culturales y sociales, existen algunas similitudes y diferencias en sus leyes contra el delito cibernético. Además, existe una variación en el alcance de abordar los delitos cibernéticos. El documento resalta algunos problemas relacionados con estas leyes y brinda sugerencias para mejorar e investigar en el futuro.

(Aditya K Sood, Sherali Zeadally, and Rohit Bansal, 2017) Los ciberdelincuentes despliegan botnets para realizar operaciones nefastas en Internet. Las redes de bots se gestionan a gran escala y aprovechan la potencia de las máquinas comprometidas, que se controlan a través de portales centralizados conocidos como paneles C & C. Los paneles C & C se consideran el entorno operativo principal de los atacantes a través del cual los robots se controlan y actualizan a intervalos regulares de tiempo. Los paneles de C & C también almacenan información robada de las máquinas comprometidas como parte de la actividad de ex filtración de datos. En este estudio empírico, fueron analizadas más de 9000 URLs de

C & C para comprender mejor la implementación y las características operativas de las botnets basadas en HTTP.

(Mittal, 2017) Las características únicas del ciberespacio como el anonimato en el espacio y el tiempo, la ausencia de fronteras geográficas, la capacidad de lanzar acciones sorpresa con rapidez y el potencial para comprometer activos en el mundo virtual y real, ha atraído la atención de personas para cometer crímenes en el ciberespacio. La ley de crímenes en el mundo físico enfrenta desafíos en su aplicación a los crímenes en el ciberespacio debido a cuestiones de soberanía, jurisdicción, investigación transnacional y evidencia extraterritorial. En este trabajo se ha intentado aplicar la teoría de la actividad de rutina (RAT) del crimen en el mundo físico al ciberespacio de la escena del crimen. Se ha desarrollado un modelo para la delincuencia en el ciberespacio y se ha argumentado que la ley penal de la delincuencia en el mundo físico es inadecuada en su aplicación a los crímenes en el mundo virtual. Para manejar la delincuencia en el ciberespacio, es necesario abordar los problemas de las leyes aplicables y las jurisdicciones en conflicto mediante la regulación de la arquitectura de Internet mediante leyes especiales del ciberespacio. Se ha presentado un caso para tener una Convención Internacional de Ciberdelincuencia con el Convenio del Consejo de Europa sobre Delito Cibernético como armas de fuego.

(Wadhwa & Arora, 2017) El crimen es una palabra común que siempre escuchamos en esta era de la globalización. Los crímenes se refieren a cualquier violación de la ley o la comisión de un acto prohibido por la ley. En las últimas dos décadas, el cibercrimen se ha convertido en un tema cada vez más ampliamente debatido en muchos ámbitos de la vida. Está claro que el rápido crecimiento de Internet ha creado nuevas oportunidades sin precedentes para delinquir. Se define como delitos cometidos en Internet utilizando la computadora como herramienta o como víctima específica. Este documento presenta los tipos de actividades de

cibercrimen, cuestiones importantes sobre la seguridad, la prevención y la detección de delitos cibernéticos

(Graeber, 2016) Para septiembre de 2010, White se dirigía a una base operativa avanzada fuera de la sede de Kabul, como parte de una célula secreta de inteligencia para ayudar a confrontar a los talibanes y al-Qaida, a bloquear y reducir su flujo de dinero cifrado en línea y conquistar a la población de Afganistán.

Es difícil imaginar que solo unas pocas semanas antes, White había sido simplemente otro posdoctorado de Harvard increíblemente joven en chanclas esperando un verano en Cambridge. Helicópteros de combate y zonas de guerra no estaban en el radar; había café con leche en la plaza y escalada en roca, y en el otro lado del campus, una beca de prestigio en la Escuela de Ingeniería y Ciencias Aplicadas, donde estaba trabajando en la intersección de grandes datos, estadísticas y aprendizaje automático. Se había ganado la posición académica y tenía la expectativa de que continuaría así para siempre: convirtiéndose en profesor, construyendo un laboratorio y tirando papeles blancos desde una torre de marfil.

Pero luego su mentor le pidió que asistiera a una conferencia de fin de semana en DARPA. White sabía que eran los proyectos de los expertos que explicaba la Agencia de Proyectos de Investigación Avanzada de Defensa, el departamento de innovación científica del Pentágono, las personas que le trajeron exoesqueletos biónicos, visión nocturna, el M16, agente naranja, GPS, tecnología sigilosa, satélites meteorológicos y el Internet. Los proyectos DARPA combinan gente inteligente, grandes ideas y alto financiamiento en dólares del gobierno. Su objetivo era ayudar a la nación a evitar sorpresas tecnológicas, y cada cinco a 10 años, lanzar tecnología cambiante con una ventaja estratégica.

(Yan Wang, Crawling ranked deep Web data sources, 2016) En la era de los macro datos, la gran mayoría de los datos no provienen de la Web

de superficie, la Web que está interconectada por hipervínculos e indexada por los motores de búsqueda más generales. En cambio, los datos valiosos a menudo residen en la Web profunda, la Web que se oculta detrás de las interfaces de consulta. Dado que numerosas aplicaciones, como integración de datos y portales verticales, requieren datos web profundos, se desarrollaron varios métodos de rastreo para recolectar exhaustivamente una fuente de datos Web profunda con el costo mínimo (o casi mínimo). La mayoría de los métodos de rastreo existentes suponen que se devuelven todos los documentos que coinciden con las consultas. En la práctica, las fuentes de datos a menudo devuelven la parte superior. Esto dificulta la recopilación exhaustiva de datos: los documentos con una clasificación alta se devolverán varias veces, mientras que los documentos clasificados como bajos tienen pocas probabilidades de ser devueltos. En este documento, descomponen este problema en dos subproblemas ortogonales, es decir, problemas de sesgo de consulta y clasificación, y proponen un método de rastreo basado en la frecuencia de los documentos para superar el problema del sesgo de clasificación. La lógica de este método es utilizar las consultas cuyas frecuencias de documento se encuentran dentro del rango especificado para evitar el efecto de la clasificación de búsqueda más el límite de retorno y reducir significativamente la dificultad de rastrear la fuente de datos clasificada. El método se probó ampliamente en una variedad de conjuntos de datos y se comparó con dos métodos existentes. El resultado experimental demuestra que este método supera a los dos algoritmos en un 58% y un 90% en promedio, respectivamente.

(Andres Baravalle, 2016) En los últimos años, los organismos gubernamentales han tratado inútilmente de luchar contra los mercados web oscuros. Poco después del cierre de "The Silk Road" por el FBI y Europol en 2013, se han establecido nuevos sucesores. Mediante la combinación de cripto-monedas, ~~y~~ herramientas y protocolos de comunicación no estándar, los agentes pueden comerciar anónimamente en un mercado de ar-

tículos ilegales sin dejar algún registro. Este artículo presenta una investigación llevada a cabo para obtener información sobre los productos y servicios vendidos en uno de los mercados más grandes de drogas, identificaciones falsas y armas en Internet, Agora. Nuestro trabajo arroja luz sobre la naturaleza del mercado, existe una clara preponderancia de medicamentos, que representa casi el 80% del total de artículos en venta. La disponibilidad inmediata de documentos falsificados, mientras que compensa un porcentaje mucho menor del mercado, aumenta la preocupación. Finalmente, se discute y presenta el rol del crimen organizado dentro de Ágora.

(Weimann, 2016) Los términos Deep Web, Deep Net, Invisible Web o Dark Web se refieren al contenido en la World Wide Web que no está indexado por los motores de búsqueda estándar. Uno puede describir Internet como compuesto de capas: la capa "superior", o la Web de superficie, se puede acceder fácilmente mediante búsquedas regulares. Sin embargo, las capas "más profundas", el contenido de la Web profunda, no han sido indexadas por los motores de búsqueda tradicionales como Google. Michael K. Bergman, que escribió el artículo seminal sobre Deep Web, comparó la búsqueda de Internet con el arrastre de una red a través de la superficie del océano: es posible que se atrape mucho en la red, pero hay una gran cantidad de información que es más profunda y por lo tanto perdida. De hecho, la mayor parte de la información de la Web está oculta en sitios profundos, y los motores de búsqueda estándar no pueden acceder a ella.

(Ahmed T. Zulkarnine, Richard Frank, Bryan Monk, Julianna Mitchell, Garth Davies , 2016) La Red Tor, una parte oculta de Internet, se está convirtiendo en un lugar ideal para actividades y servicios ilegales, incluidos los grandes mercados de drogas, fraudes financieros, espionaje y abuso sexual infantil. Los investigadores y las fuerzas del orden confían en las investigaciones manuales, que consumen mucho tiempo y, en últi-

ma instancia, son ineficientes. La primera parte de este documento explora el contenido ilícito y criminal identificado por investigadores prominentes en la web oscura. Anteriormente se desarrolló un rastreador web que buscaba automáticamente sitios web en Internet en base a palabras clave predefinidas y seguía los hipervínculos para crear un mapa de la red. Este rastreador ha demostrado un éxito previo en la localización y extracción de datos sobre imágenes, videos, palabras clave y vínculos de explotación infantil en internet público. Sin embargo, como Tor funciona de manera diferente en el nivel de TCP, y utiliza conexiones de socket, otros desafíos técnicos se enfrentan al rastrear Tor. Algunos de los otros desafíos inherentes para el rastreo avanzado de Tor incluyen la escalabilidad, las compensaciones de selección de contenido y las obligaciones sociales. Se discuten estos desafíos y las medidas tomadas para cumplirlos. El rastreador web modificado para Tor, denominado "Dark Crawler", ha podido acceder a Tor al mismo tiempo que accede a Internet pública. Se presentan los hallazgos iniciales sobre qué contenidos extremistas y terroristas están presentes en Tor y cómo este contenido está conectado entre sí en una red mapeada que facilita los crímenes web oscuros. Nuestros resultados hasta ahora indican que los sitios web más populares en la web oscura están actuando como catalizadores para la expansión web oscura al proporcionar la base de conocimientos necesaria, soporte y servicios para construir servicios ocultos Tor.

(Andrew J. Park, Brian Beck, Darrick Fletche, Patrick Lam, and Herbert H. Tsang, 2016) Los grupos extremistas han recurrido a los sitios de Internet y las redes sociales como un medio para compartir información entre ellos. Este estudio de investigación analiza los mensajes del foro y encuentra personas que muestran tendencias radicales a través del uso del procesamiento del lenguaje natural y el análisis del sentimiento. Los datos del foro que se utilizan provienen de seis foros islámicos en la Dark Web los cuales están disponibles para la investigación de seguridad. Este proyecto de investigación utiliza un etiquetador POS para aislar palabras cla-

ve y sustantivos que pueden utilizarse con el programa de análisis de sentimientos. Luego, el programa de análisis de sentimientos determina la polaridad de la publicación. La publicación se califica como positiva o negativa. Estos puntajes se dividen en puntajes radicales mensuales para cada usuario. Una vez que estos clústeres de tiempo se asignan, el cambio en las opiniones de los usuarios a lo largo del tiempo puede interpretarse como un aumento o una disminución del nivel de radicalismo. Luego se compara a cada usuario en una línea de tiempo con otros usuarios radicales y eventos para determinar posibles conexiones o relaciones. La capacidad de analizar un foro para un cambio global de actitud puede ser un indicador de inquietud y posibles acciones radicales o terrorismo.

(Xuefeng Xian, Pengpeng Zhao, Victor S. Sheng, Ligang Fang, Caidong Gu, Yuanfeng Yang, and Zhiming Cui , 2016) Para muchas aplicaciones, encontrar instancias excepcionales o valores atípicos puede ser más interesante que encontrar patrones comunes. El trabajo existente en la detección de valores atípicos nunca considera el contexto de la web profunda. En este documento, se argumenta que, para muchos escenarios, es más significativo detectar valores atípicos en la web profunda. En el contexto de la Web profunda, los usuarios deben enviar consultas a través de una interfaz de consulta para recuperar los datos correspondientes. Por lo tanto, los métodos tradicionales de minería de datos no se pueden aplicar directamente. La principal contribución de este documento es desarrollar un nuevo método de minería de datos para la detección de valores atípicos en redes profundas. En ese enfoque, el espacio de consulta de una fuente de datos web profunda se estratifica en función de una muestra piloto. El muestreo de vecindario y el muestreo de incertidumbre se desarrollan en este documento con el objetivo de mejorar el recuerdo y la precisión en función de la estratificación. Finalmente, una cuidadosa evaluación del rendimiento del algoritmo confirma que el enfoque puede detectar de manera efectiva valores atípicos en la web profunda.

(Somayyeh Aghababaei, Masoud Makrehchi , 2016) Las redes sociales brindan oportunidades cada vez mayores para que los usuarios compartan voluntariamente sus pensamientos y preocupaciones en un gran volumen de datos. Si bien los datos generados por el usuario de cada individuo pueden no proporcionar información considerable, cuando se combinan, incluyen variables ocultas, que pueden transmitir eventos significativos. En este artículo, se plantea la pregunta de si el contexto de los medios sociales puede proporcionar "señales" socio-conductuales para la predicción del delito. La hipótesis es que la multitud de datos disponibles públicamente en las redes sociales, en particular Twitter, puede incluir variables predictivas, que pueden indicar los cambios en las tasas de criminalidad. Se desarrolla un modelo para la predicción de la tendencia del delito donde el objetivo es emplear contenido de Twitter para identificar si las tasas de criminalidad han disminuido o aumentado para el marco de tiempo prospectivo. También se presenta un modelo de muestreo de Twitter para recopilar datos históricos para evitar la pérdida de datos a lo largo del tiempo. El modelo de predicción fue evaluado para diferentes ciudades en los Estados Unidos. Los experimentos revelaron la correlación entre las características extraídas del contenido y las direcciones de la tasa de criminalidad. En general, el estudio proporciona información sobre la correlación del contenido social y las tendencias delictivas, así como sobre el impacto de los datos sociales en la provisión de indicadores predictivos.

(Regner Sabillon, Jeimy Cano, Victor Cavaller, Jordi Serra-Ruiz, 2016) Hoy en día, el cibercrimen está creciendo rápidamente en todo el mundo a medida que surgen nuevas tecnologías, aplicaciones y redes. Además, Deep Web ha contribuido al crecimiento de actividades ilegales en el ciberespacio. Como resultado, los cibercriminales están aprovechando las vulnerabilidades del sistema para su propio beneficio. Este artículo presenta la historia y la conceptualización del delito cibernético, explora diferentes categorizaciones de ciberdelincuentes y ataques cibernéticos, y

expone una taxonomía o tipología exhaustiva de ataques cibernéticos. Las categorías comunes incluyen a la computadora como el objetivo para cometer el delito, donde la computadora se utiliza como una herramienta para perpetrar el delito grave, o cuando un dispositivo digital es una condición incidental para la ejecución de un delito.

(KhalidAl-Rowailya Muhammad Abula ishb Nur Al-Hasan Haldarc Majed Al-Rubaian, 2015) En este artículo, se presenta el desarrollo de un Léxico de análisis de sentimiento bilingüe (BiSAL) para el dominio de seguridad cibernética, que consiste en un léxico de sentimiento para inglés (SentiLEN) y un léxico de sentimiento para árabe (SentiLAR) que se puede utilizar para desarrollar la minería de opinión, y sistemas de análisis de sentimientos para datos textuales bilingües de los foros de Dark Web. Para SentiLEN, se identifica una lista de 279 palabras en inglés relacionadas con amenazas cibernéticas, radicalismo y conflictos, y se diseña un proceso unificador para unificar sus puntajes de opinión obtenidos a partir de cuatro conjuntos de datos de sentimientos diferentes. Mientras que, para SentiLAR, los sentimientos que tienen palabras en árabe se identifican a partir de una colección de 2000 mensajes del foro web de Alokab, que contiene contenidos radicales. El SentiLAR proporciona una lista de 1019 sentimientos con palabras árabes relacionadas con amenazas cibernéticas, radicalismo y conflictos junto con sus variantes morfológicas y polaridad de sentimiento. Para la determinación de polaridad, se realiza un proceso de análisis semi-automatizado por tres expertos en idioma árabe y sus calificaciones se agregan utilizando algunas funciones agregadas. Se desarrolla una interfaz web para acceder a los léxicos (SentiLEN y SentiLAR) del conjunto de datos BiSAL en línea.

(Martijn Spitters, Femke Klaver, Gijs Koot, Mark van Staalduinen, 2015) Las redes de anonimato como Tor albergan muchos mercados clandestinos y foros de discusión dedicados al comercio de bienes y servicios ilegales. A medida que van ganando popularidad, el análisis de su contenido

y de sus usuarios se vuelve cada vez más urgente para muchas partes diferentes, desde organismos encargados de hacer cumplir la ley y de seguridad hasta instituciones financieras. Un problema importante en ciber-forense es que las técnicas de anonimización como el enrutamiento de cebolla de Tor han hecho muy difícil rastrear las identidades de los sospechosos. En este trabajo proponemos configuraciones de clasificación para dos tareas relacionadas con la identificación del usuario, a saber, la clasificación de alias y la atribución de autoría. Aplicamos nuestras técnicas a los datos de un foro de discusión de Tor dedicado principalmente al tráfico de drogas, y demostramos que para ambas tareas logramos una alta precisión usando una combinación de n-gramas del nivel de personaje.

(Matti Nasi, Atte Oksanenb, Teo Keipia and Pekka Rasanen , 2015) Este estudio examina la victimización por delito cibernético, cuáles son algunas de las características comunes de tales crímenes y algunos de los predictores generales de la victimización por cibercrimen entre adolescentes y adultos jóvenes. Una muestra combinada de cuatro países (Finlandia, EE. UU., Alemania y el Reino Unido) se construye a partir de participantes de entre 15 y 30 años de edad. Según los hallazgos, la victimización por crímenes en línea es relativamente poco común (el 6,5% de los participantes fueron víctimas). La difamación y la amenaza de violencia eran las formas más comunes de victimización y acoso sexual, la menos común. El sexo masculino, la menor edad, el origen de los inmigrantes, la residencia urbana, el hecho de no vivir con sus padres, el desempleo y una vida social fuera de línea menos activa fueron predictores significativos para la victimización del delito cibernético.

(Sonali Gupta, Komal Kumar Bhatia, 2014) Una gran cantidad de datos en la WWW sigue siendo inaccesible para los rastreadores de los motores de búsqueda web, ya que solo puede exponerse a demanda a medida que los usuarios completan y envían formularios. La web oculta se refiere

a la recopilación de datos web a los que solo puede acceder el rastreador mediante una interacción con el formulario de búsqueda basado en web y no simplemente atravesando hipervínculos. La investigación en Hidden Web surgió hace casi una década y la línea principal es explorar formas de acceder al contenido en bases de datos en línea que generalmente están ocultas detrás de los formularios de búsqueda. Los esfuerzos en el área se centran principalmente en el diseño de rastreadores Web ocultos que se centran en los formularios de aprendizaje y los llenan de valores significativos. El documento da una idea de los diversos rastreadores de la Red Oculta desarrollados con el propósito de mencionar las ventajas y desventajas de las técnicas empleadas en cada uno.

(Helena Piccinini, Marco A. Casanova, Luiz André P. P. Leme, Antonio L. Furtado , 2014) Este artículo presenta un proceso de diseño, llamado W-RayS, para describir datos geográficos de Deep Web y publicar las descripciones tanto en la Web of Data como en Surface Web. El artículo también describe un conjunto de herramientas que admite el proceso y analiza un experimento en el que se utilizó el kit de herramientas para publicar datos almacenados en un servidor de mapas grande. En resumen, para describir los datos geográficos en formato vectorial, el diseñador debe especificar primero las vistas sobre la base de datos geográfica subyacente que capturan las características básicas de los objetos geográficos y sus relaciones topológicas representadas en los datos vectoriales.

(Zhou Li, Sumayah Alrwais Yinglian Xie, Fang Yu, XiaoFeng Wang, 2013) Las actividades web maliciosas continúan siendo una gran amenaza para la seguridad de los usuarios web en línea. A pesar de la gran cantidad de formas de ataques y la diversidad de sus canales de distribución, en la parte de atrás, todos están organizados a través de infraestructuras web maliciosas, que permiten a los malhechores hacer negocios entre sí y utilizar los recursos de los demás. Identificar los ejes de las infraestructu-

ras oscuras y distinguir aquellos que son valiosos para los adversarios de aquellos desechables son críticos para ganar ventaja en la batalla en contra de ellos. En este documento, utilizando casi 4 millones de rutas URL maliciosas rastreadas desde diferentes canales de ataque, llevamos a cabo un estudio a gran escala sobre las relaciones topológicas entre los hosts en la infraestructura web maliciosa. Nuestro estudio revela la existencia de un conjunto de hosts maliciosos dedicados topológicamente que desempeñan papeles de orquestación en actividades maliciosas. Están bien conectados a otros hosts maliciosos y no reciben tráfico de sitios legítimos. Motivados por sus características distintivas en topología, desarrollamos un enfoque basado en gráficos que se basa en un pequeño conjunto de hosts maliciosos conocidos como semillas para detectar la dedicación de hosts maliciosos a gran escala. El método es general a través del uso de diferentes tipos de datos de semillas, y da como resultado una tasa de expansión de más de 12 veces en la detección con una baja tasa de detección falsa del 2%. Muchos de los hosts detectados operan como redirectores, en particular Traffic Distribution Systems (TDS), que son de larga duración y reciben tráfico de nuevas campañas de ataques a lo largo del tiempo. Estos TDS juegan roles críticos en la administración de flujos de tráfico maliciosos

(Chen H. , 2011) El proyecto Dark Web del Laboratorio de Inteligencia Artificial de la Universidad de Arizona (AI Lab) es un programa de investigación científica a largo plazo que tiene como objetivo estudiar y comprender los fenómenos del terrorismo internacional (yihadista) a través de un enfoque computacional centrado en datos. Nuestro objetivo es recopilar "TODOS" contenidos web generados por grupos terroristas internacionales, incluidos sitios web, foros, salas de chat, blogs, sitios de redes sociales, videos, mundo virtual, etc. El laboratorio mencionado ha desarrollado varios métodos de minería de datos multilingüe, minería de textos y técnicas web de minería para realizar análisis de enlaces, análisis de contenido, análisis de métricas web, análisis de sentimientos, análisis de au-

toría y análisis de video en su investigación. Los enfoques y métodos desarrollados en este proyecto contribuyen a avanzar en el campo de Inteligencia y Seguridad Informática (ISI).

Esta monografía tiene como objetivo proporcionar una visión general del entorno e interior de la Dark Web, sugerir un enfoque computacional sistemático para comprender los problemas e ilustrar con técnicas seleccionadas, métodos y estudios de casos desarrollados por los miembros del equipo de AI Lab Dark Web de la Universidad de Arizona. Este trabajo tiene como objetivo proporcionar una monografía interdisciplinaria y comprensible sobre la investigación Dark Web en tres dimensiones: cuestiones metodológicas en la investigación Dark Web; base de datos y técnicas computacionales para apoyar la recopilación de información y la minería de datos; y desafíos y enfoques legales, sociales, de privacidad y de confidencialidad de datos. Brindará conocimiento útil a científicos, profesionales de la seguridad, expertos en antiterrorismo y responsables políticos. La monografía también puede servir como material de referencia o libro de texto en cursos de posgrado relacionados con la seguridad de la información,

4. Diseño Metodológico

El procedimiento metodológico que se propone para cumplir el objetivo general y los objetivos particulares, se describe en la Figura 4.1, en la cual se refieren los siguientes pasos:

1. Descubrimiento de redes onion mediante mecanismos como “Hunchly”.
2. Recuperar información contenida en las redes onion descubiertas.
3. Diseñar una Ontología de términos que ayude a comprender las conductas delictivas que se quieren clasificar.
4. Modelar los datos obtenidos para su posterior clasificación.
5. Generar un algoritmo de clasificación que permita identificar delitos cibernéticos dentro de las redes onion descubiertas.
6. Utilizar un mecanismo de presentación de los datos a través de estadísticas.

Se considera principalmente a la red privada de TOR como población para el análisis de información en la búsqueda de delitos cibernéticos.

4.1 Descubrimiento de redes onion

Con respecto a lo investigado en el estado del arte, se encontraron dos alternativas funcionales para el descubrimiento de redes onion que nos permitirán contar con una base de datos, en la cual probaremos nuestro algoritmo de clasificación posteriormente, con el objetivo de identificar y clasificar delitos cibernéticos finalmente.

1. Herramienta para investigaciones en línea “Hunchly”

En primer lugar se cuenta con un servicio denominado “Hunchly” (Hunchly, 2018), el cual es una herramienta diseñada para investigaciones en

línea, al suscribirse envía vía correo electrónico diariamente un archivo en Excel clasificado en redes onion activas, inactivas y las descubiertas al día en cuestión. En la figura 4.2 se muestra un ejemplo de ello.

La cantidad de redes activas a Mayo de 2018 es de 6,602 redes onion, e inactivas tenemos 25,532 redes onion, por lo que trabajaremos con las activas en la siguiente etapa de la tesis.

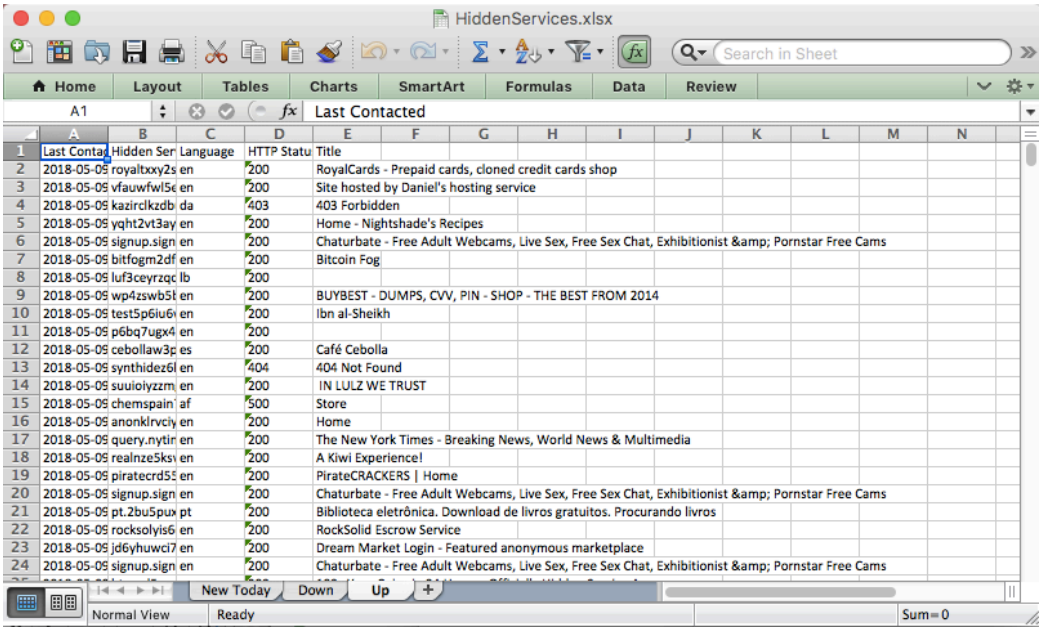


Figura 4.2 Archivo en Excel con redes onion del servicio.

Se tiene un seguimiento de este servicio desde el 17 de Abril de 2017 a la fecha, con el objetivo de identificar un patrón de comportamiento de las redes activas, inactivas y la cantidad de redes que cada día se descubren basada en esta muestra. En la tabla 4.1 podemos ver parte de este comportamiento y podemos consultar el Apéndice 3 para ampliar este comportamiento.

Fecha	Nuevas	Inactivas	Activas
02/06/19	3	1006	4513
01/06/19	1	1181	4584

Fecha	Nuevas	Inactivas	Activas
31/05/19	0	1043	4569
30/05/19	1	1	4636
29/05/19	0	968	4611
28/05/19	1	953	4668
27/05/19	1	961	4566
26/05/19	0	958	4577
25/05/19	0	1023	4528

Tabla 4.1. Clases de Delitos de Delincuencia Organizada.

2. Herramienta OnionScan

OnionScan (Lewis) es una herramienta gratuita y de código abierto para investigar la Dark Web, tiene dos objetivos principales:

- Ayudar a solucionar problemas de seguridad operacional a los operadores de servicios ocultos de redes onion, como configuraciones erróneas.
- Ayudar a los investigadores de delitos a supervisar y rastrear los sitios de Dark Web.

En el Anexo 1, se presentan los pasos para implementar esta herramienta.

Por lo que podemos señalar que en la etapa de descubrimiento de redes onion se obtuvo un total de 8,386 redes, de las cuales se encuentran activas 1,143 redes onion con el proyecto OnionScan, las cuales sirven como Base de Datos para el Algoritmo de clasificación de la siguiente etapa de la tesis.

4.2 Medidas de Seguridad para navegar en TOR

Se investigaron las medidas de seguridad publicadas en artículos de investigación internacionales que trabajan en el tema de la Dark Web y Tor, de las cuales se compilan las siguientes:

a) Uso de Tails.

Tails es un sistema operativo, de arranque desde una USB o DVD, de tal forma que no se utilice el disco duro de nuestro equipo y se preserve la privacidad y el anonimato, ya que las conexiones a Internet son forzadas a ir a través de la red TOR, no deja traza y usa herramientas de cifrado en los archivos, correos electrónicos y mensajería instantánea. Se puede descargar e implementar desde la página:
https://tails.boum.org/index.es.html

b) Uso de máquinas virtuales para conexión a TOR y para navegación.

Para ello se utilizan dos máquinas virtuales Linux, la primera tendrá la máquina virtual “Whonix Gateway” que pasará TODO el tráfico a la red TOR y la segunda máquina virtual servirá para navegar mediante algún “browser” que se conectará a la red de la primera máquina virtual, esto servirá para en caso de ser comprometida la red Tor no se encuentre trabajando en ella.

Aunado a esto se puede agregar al browser de la segunda máquina virtual un “componente” llamado Hunchly que permite realizar investigación de forma segura, de tal forma que se guarden las páginas onion seleccionadas, así como fotografías e información adicional.

c) Al navegar con el navegador de Tor, no permitir el uso de Scripts en las páginas que se visitan, además de no ejecutar programas descargados de esta red, si es necesario descargar y ejecutar alguno de estos programas es importante revisarlo primero con un antivirus o en un servicio gratuito como Virus Total.

d) Se recomienda eliminar los controladores del micrófono y la cámara web cuando se cuente con ellas y si es posible tapar con una cinta la cámara del equipo.

e) Es recomendable implementar una propia Red Privada Virtual (VPN) aun cuando Tor maneja su propio cifrado y VPN.

f) No ingresar datos personales, correos electrónicos de servicios de la Surface Web, nombres de usuarios utilizados anteriormente, ni la misma contraseña para los sitios en los que se registre.

g) Evitar el uso de tarjetas de crédito personales, si es necesario utilice tarjetas preparadas de un solo uso y verifique que el sitio Web esté seguro al verificar la dirección Web comenzando con “https://“y no con “http://“ ya que la “s” significa que los datos enviados y recibidos viajan cifrados.

4.3 Conformación de una Ontología de términos

Para comprender las diferentes clases de delitos cibernéticos que se intentan encontrar dentro de la información de las redes onion identificadas, se trabajó en una ontología de términos basada en Ley Federal Contra la Delincuencia Organizada de México.

No todas las redes onion con las que contamos contienen información de conductas en delitos cibernéticos, por lo que se procedió a clasificar cuáles de ellas podrán servir como muestra para alimentar a nuestro sistema, para ello se conformó una Ontología de términos que nos sirva para entender el objeto a estudiar.

Las ontologías se emplean en todo tipo de aplicaciones informáticas, algunas de ellas se construyen con el único objetivo de alcanzar una comprensión de un dominio de conocimiento específico y otras con pro-

pósito general. Otro tipo de ontologías son diseñadas como catálogos o taxonomías que son incorporadas en otras ontologías o sistemas de información para su reutilización.

Es necesario proporcionar un marco conceptual que represente el conocimiento de Delitos en México para que se tenga una referencia del tipo de delitos sobre el que se trabaja, su definición y soporte legal en México. De tal forma que sirva como base para el desarrollo de esta metodología de identificación y clasificación de delitos cibernéticos sobre la red TOR.

Esta ontología en Delitos está basada en la "Clasificación de los Delitos de Delincuencia Organizada" estipulado por la Ley Federal Contra la Delincuencia Organizada, y las demás leyes federales y códigos de los Estados con motivo de la reforma de junio de 2016, como cumplimiento al mandato constitucional de la implementación del nuevo Sistema Penal Acusatorio (González, 2018). Tiene la siguiente estructura:

A) DELITOS CONTRA LA SEGURIDAD PUBLICA

a) **Contra la Salud:** Acopio y Tráfico de Armas de Fuego de Uso Exclusivo del Ejército Armada y Fuerza Área. Contra la Salud, Narcomenudeo y su equiparable, Desvío de Precursores Químicos, Tráfico de Órganos e Investigaciones Biomédicas;

b) **Contra la Libertad de Tránsito:** Trata de Personas, Tráfico de Menores, Secuestro y Tráfico de Personas;

c) **Contra la Libertad Sexual:** Pornografía, Corrupción de Menores, Lenocinio, Turismo Sexual;

d) **Contra él Patrimonio:** Asalto y Robo de Vehículos.

e) **Contra la Propiedad Intelectual:** En Materia de Derechos de Autor.

B) DELITOS CONTRA EL ESTADO:

a) **Contra la Seguridad Nacional:** Delincuencia Organizada, Terrorismo, Encubrimiento y amenaza de Terrorismo, Terrorismo Internacional, Financiamiento al Terrorismo;

b) **Contra la Seguridad Financiera:** Falsificación, Uso y alteración de Moneda, Operaciones con Recursos de Procedencia Ilícita, Contrabando y su equiparable; y

c) **Contra los Recursos Naturales:** Contra el ambiente y en Materia de Hidrocarburos.

De acuerdo con (González, 2018), las conductas relacionadas con las tecnologías de la información que derivan en delitos cibernéticos cometidos por miembros de la Delincuencia Organizada, sea para defraudar o extorsionar, estos delitos ni siquiera figuran en los relacionados con la Delincuencia Organizada. Se propone que se legisle lo referente al robo de identidad cometido por miembros de la Delincuencia Organizada o individuos no identificados, los dos delitos mencionados ni siquiera están contemplados en los delitos del Fuero Federal, por lo que deben incluirse en el catálogo de delitos comprendidos dentro de la Ley Federal Contra la Delincuencia Organizada y agregarlo a los delitos graves del Código Nacional de Procedimientos Penales.

Debido al objetivo de esta tesis agregaremos a nuestra ontología las conductas delictiva derivadas de las tecnologías de la información:

C) DELITOS CONTRA LAS TECNOLOGÍAS DE LA INFORMACIÓN:

a) **Delitos Cibernéticos:** Acceso Ilícito a Sistemas y Equipos de Informática.

b) **Conductas Delictivas Cibernéticas** (Utilizando Medios Electrónicos)**:** Extorsión, Fraude, Robo de Identidad, Ataques de fuerza bruta, Ciberterrorismo, Código Malicioso, Denegación de Servicio, Defacement, Divulgación no autorizada de Información, Intercepción o Modificación no Autorizada de Comunicaciones, Phishing, Spam.

Ya teniendo claro el dominio, se creó una ontología en el sistema de código abierto “Protégé”, dónde se añadieron los delitos mencionados anteriormente como clases, y se muestran en la Figura 4.3:

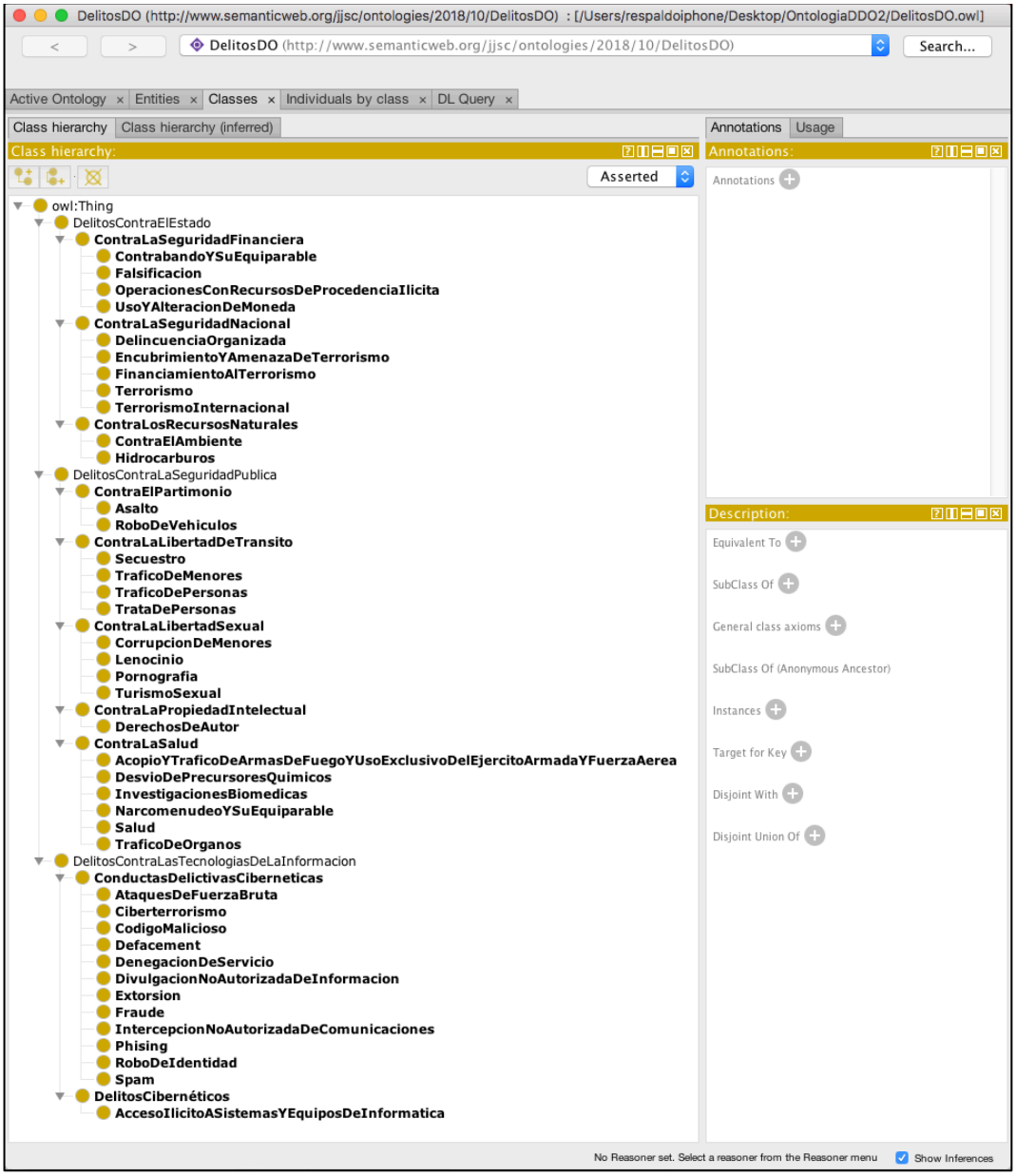


Figura 4.3. Clases de Delitos de Delincuencia Organizada.

4.4 Selección de páginas onion de conductas delictivas

Conforme la identificación de redes onion descubiertas, se identificaron páginas que presentan conductas delictivas, a criterio de un investigador de delitos electrónicos, esto con el fin de que a partir de estas se realicen técnicas para búsqueda de información y Clasificación en base a los Delitos de Delincuencia Organizada en México vista anteriormente, y posteriormente se integren estas palabras a la ontología diseñada.

Conducta delictiva	Número de redes onion
Delitos contra el libre Desarrollo de la Personalidad	74; Pornografía infantíl, Corrupción de Menores, Lenocinio
Delitos contra la Libertad y el Normal Desarrollo Psicosexual	22; Pornografía, Turismo Sexual, Escorts
Delitos Contra la Salud	61; venta de drogas, esteroides, narcóticos
Delitos Contra la Seguridad de la Nación	11; Ciber Guerrilla, Grupos anarquistas y terroristas
Delitos Contra la Seguridad Pública	17; Venta de armas y municiones
Delitos Contra la Vida y la Integridad Corporal	11; Asesinatos y violencia
Delitos en Contra de las Personas en su Patrimonio	208; Transacciones de bitcoins, clonación de tarjetas, estafas
Delitos en Materia de Derechos de Autor	3; Libros, Streaming de Video, Radiodifusiones, Películas, Videojuegos
Falsedad	11; Falsificación de moneda, pasaportes, visas, obras de arte, documentos apócrifos
Revelación de secretos y acceso ilícito a sistemas y equipos de informática	96; DDOS, Cuentas Hackeadas, Servicios de Hackeo, Servicios de Spam, Malware, acceso a sistemas, vigilancia
Con conducta delictiva	514
Sin conducta delictiva	416
Total	930

Tabla 4.2 Número de redes onion por conducta delictiva estudiada.

La tabla 4.2, presenta el número de redes onion estudiadas de acuerdo con la conducta delictiva identificada anteriormente y encuadrada al Código Penal Federal de México, haciendo un total de 514 redes onion con posibles conductas delictivas y 416 redes onion sin conducta delictiva, dando un total de 930 redes.

En el anexo 3, se muestran las redes onion que se utilizan para formar categorías de palabras por delito, con el fin de identificar nuevas redes onion en base a estas.

4.5 Recuperar información contenida en las redes onion descubiertas, a través de la herramienta wget.

A partir de la muestra de 930 redes onion de un archivo HiddenServices.xlsx, se obtuvieron 514 redes con conductas delictivas y 416 sin conductas delictivas, lo que corresponde aproximadamente a un 55% de redes onion con conductas delictivas. Ahora corresponde realizar la recuperación de esta información.

El objetivo general del proceso de recuperación de datos es extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior, esto se puede ver como la minería de datos sin procesar toda la información de varios sitios web onion. Al copiar esta información de sitios web en redes onion, obtendremos una gran cantidad de tráfico en el servidor, lo que causará latencia en el servicio web proporcionado y algunos administradores de sistemas han configurado sus servidores para bloquear la IP que exija una gran carga en un determinado período de tiempo, lo cual es una problemática a ser resuelta. Para omitir esta protección, se puede usar una IP diferente para cada solicitud, a lo que TOR puede ayudar por la forma en que está programado.

La minería de datos con secuencias de comandos, la podemos realizar con *wget*, para extraer datos y canalizarlos a través de la red anónima de TOR y evitar el bloqueo de IP en la granja de servidores, Tor es un proxy SOCKS de tal forma que su información se envía a través de una red de forma anónima. El problema con TOR es que no ofrece un proxy http, que es lo que requiere wget. Así que para solucionar esto se instala el paquete de Privoxy, que le permitirá conectarse a TOR a través de un simple proxy HTTP.

Para que sea posible descargar información de las redes onion con la herramienta wget en linux, es necesario realizar un proceso de instalación

y configuración de programas como tor, privoxy, wgetc y selektor. En el Anexo 4 se muestra el procedimiento que se utilizó.

Posteriormente a esto se realiza la descarga de información de estas redes a través del comando wget en linux y crearemos una carpeta por cada Conducta Delictiva mostrada en la Tabla 4.1, Usaremos dos opciones para ello:

1. La primera que vamos a usar es con la opción de obtener las cabeceras de los archivos html y su correspondiente index.html a través de la siguiente instrucción:

wget -i DCLS.txt --save-headers

Donde DCLS.txt es un archivo de texto con la lista de redes onion de conductas delictivas Contra la Salud

2. Con la siguiente instrucción recuperamos hasta en cinco niveles la información de las redes onion indicadas en el archivo de texto "decdlpesp.txt", en este caso la información corresponde a Delitos en contra de las personas en su patrimonio, además de obtener las cabeceras de dichos archivos:

wget -r -i decdlpesp.txt --save-headers -o gnulog

De tal forma que terminamos por recuperar la información por conducta delictiva mostrada en la tabla 4.1 para las cabeceras de los "html" y para el total de archivos en cada red onion (Figura 4.4).

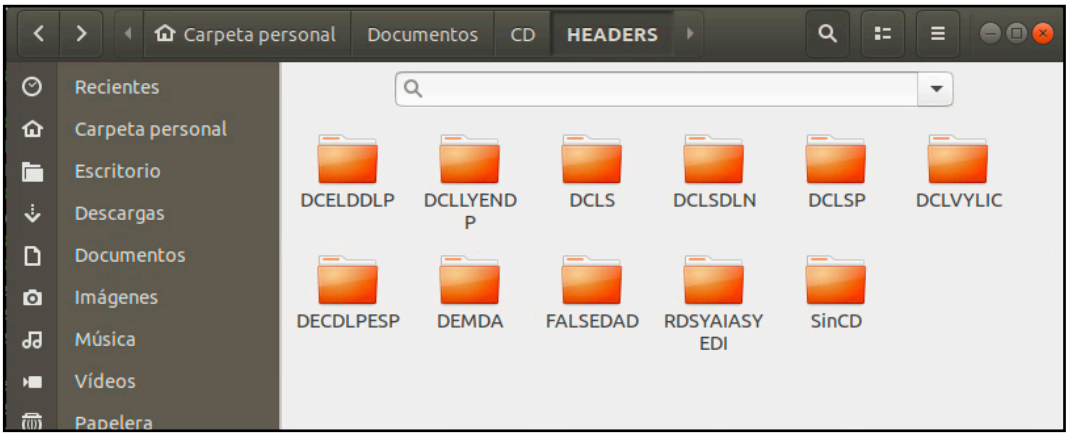


Figura 4.4. Redes onion recuperadas por conducta delictiva.

Cabe mencionar que la carpeta SinCD se refiere a Sin Conducta Delictiva, que son las redes onion que no contienen estos posibles delitos más sin embargo existen dentro de nuestra muestra de datos y será importante su información para su uso posterior.

4.6 Modelado de los datos obtenidos

4.6.1. Formato para archivos html recuperados

Se obtuvieron una serie de archivos index.html conteniendo su cabecera y su estructura html, se necesita convertirlos a formato de texto, quitarles sus etiquetas html y quitarle sus caracteres de control convirtiéndolos a formato “UNIX” para poder tratarlos posteriormente, para ello se realizó lo siguiente:

1. Instalar el paquete html2text que convierte archivos html a texto y les quita sus etiquetas html, ya instalado convertimos los archivos html recuperados en texto, para ello ejecutamos lo siguiente en la línea de comandos:

```
apt install html2text
html2text index.html.* > index.txt.*
```

2. Como existen caracteres de control que cada sistema operativo maneja con sus ficheros, necesitamos convertir los archivos recuperados a un formato de Unix, para esto utilizaremos el comando dos2unix como se muestra a continuación:

dos2unix index.html.*

4.6.2. Creación de CSV con información de cabeceras por cada red onion

Para poder realizar una categorización de las semejanzas y diferencias que existen entre los diferentes tipos de conductas delictivas se pretende crear archivos de texto separados por comas que incluyan las características de sus cabeceras, así como su valor de similitud Jaccard, valor de similitud coseno, media, mínima, máxima, mediana y varianza. Para ello se procede con el siguiente procedimiento:

1. Para recuperar las características de las cabeceras de los archivos index.html por delito con las que ya se cuenta cómo se puede ver en la figura 4.5, se puede usar la herramienta awk y generar un archivo de texto separado por punto y coma con estas características.

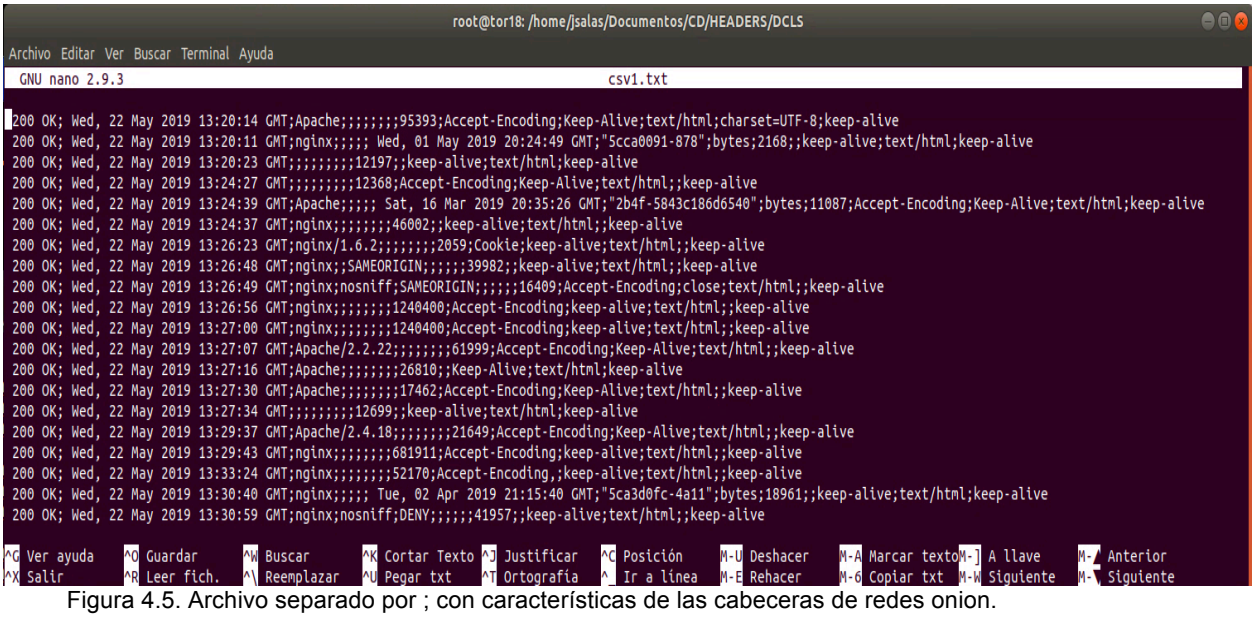


Figura 4.5. Archivo separado por ; con características de las cabeceras de redes onion.

Para lograr esto realizar lo siguiente:

```
# awk -f cmd.txt index.html.1
```

Donde cmd.txt es un archivo de texto que contiene el resultado de la extracción de características del encabezado del archivo html dado. Podemos observar el código de este archivo en el Anexo 5.

Se generó un script que automáticamente convierte todos los archivos index a formato “UNIX” y posteriormente extrae las cabeceras y genera un archivo separado por puntos y comas que recopila las características de las cabeceras de todos los archivos de entrada “index”.

Posteriormente se procedió a correr el script para todas las carpetas clasificadas por delito para obtener las cabeceras con las características en archivos separados con comas para su análisis:

```
# ./proc1.sh
```

Resultando así el archivo csv1.txt para cada carpeta mostrado en la figura 4.5.

4.7 Algoritmo de Clasificación

4.7.1. Cálculo de similitud de Jaccard y Coseno

Una vez teniendo los archivos de texto separados por coma con las características de sus cabeceras de las redes onion con conductas delictivas, se requiere adjuntar a cada registro el cálculo del índice de Jaccard y la similitud Coseno, esto con el fin de saber que tanto se parecen dos archivos, uno con conductas delictivas y otro sin conductas delictivas, e incluso entre diferentes conductas delictivas y entre ellos mismos.

Para llevar a cabo esto se ejecuta un programa en RStudio, así que se necesita instalar junto con sus librerías adicionales:

sudo apt-get install RStudio

Dentro de RStudio instalamos las librerías text2vec, stringr y data.table:

```
> install.packages(‘text2vec')
> install.packages('stringr')
> install.packages('data.table')
```

En la figura 4.6 se muestra el programa CalculateSimilarity.R, las figuras 4.7 y 4.8 presentan dos carpetas bad y good, en la carpeta bad tenemos 3 archivos html y su correspondiente txt (sin etiquetas html) con conductas delictivas; y en la carpeta good tenemos un dominio sin conducta delictiva html y txt.

El programa comienza almacenando en good el contenido del archivo de texto de la carpeta good, y en bad el contenido de los tres archivos contenidos en la carpeta bad, posteriormente crea dos vocabularios, uno bueno y uno malo conteniendo las palabras de los mismos, y finalmente calcula el índice de Jaccard y la Similaridad de Coseno. En el anexo 6 podemos ver el código de este programa.

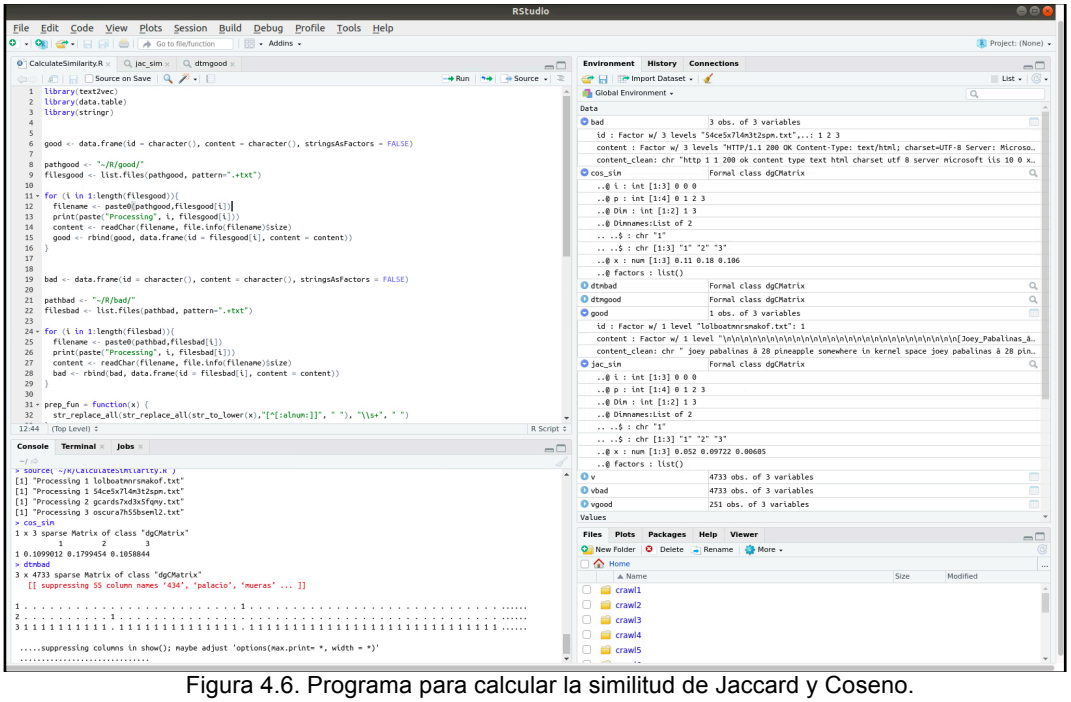


Figura 4.6. Programa para calcular la similitud de Jaccard y Coseno.

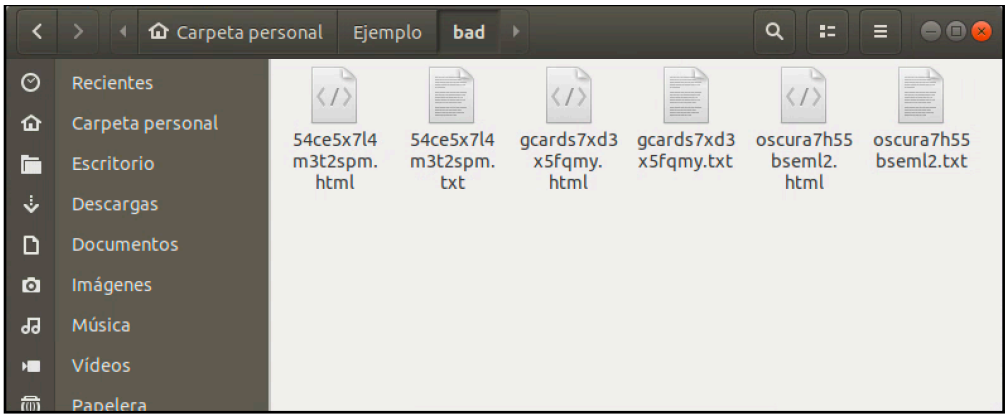


Figura 4.7. Dominios con conductas delictivas.

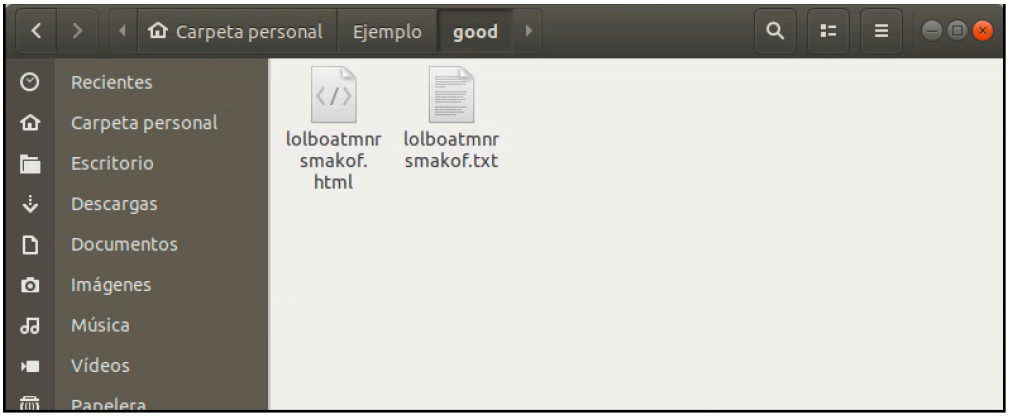


Figura 4.8. Dominio sin conducta delictiva.

Podemos ver en la Consola el resultado de comparar los archivos de texto de las carpetas good y bad. Al ejecutar la función sim2, se compara

el archivo de texto lolboatmnrsmakof.txt que no contiene conducta delictiva contra los archivos 54ce5x7l4m3t2spm.txt, gcards7xd3x5fqmy.txt y oscura7h55bseml2.txt que si contienen conductas delictivas.

```
jac_sim = sim2(dtmgood, dtmbad, method = "jaccard", norm = "none")
```

El resultado es:

```
> jac_sim
1 x 3 sparse Matrix of class "dgCMatrix"
     1           2           3
1 0.052 0.09722222 0.006048387
```

lo cual indica que no existe mucha similitud entre el archivo sin conducta delictiva contra los archivos con conducta delictiva, esto conforme al cálculo de similitud de Jaccard. La fila 1 es el archivo sin conducta delictiva de la carpeta good que se compara con los 3 archivos de la carpeta bad, aquí el que tiene mayor acercamiento es el segundo con 0.097 y el que menos se parece es el primero con 0.052.

En el caso del cálculo de similitud con Coseno, el resultado es el siguiente:

```
> cos_sim
1 x 3 sparse Matrix of class "dgCMatrix"
         1           2           3
1 0.1099012 0.1799454 0.1058844
```

Podemos notar que de igual forma no son parecidos, sin embargo, los valores son más altos que en el caso de Jaccard.

En la figura 4.9 podemos notar que en el archivo sin conductas delictivas existen 252 palabras y cuáles son las que más se repiten.

En la figura 4.10 vemos la frecuencia de palabras de los 3 archivos con conductas delictivas.

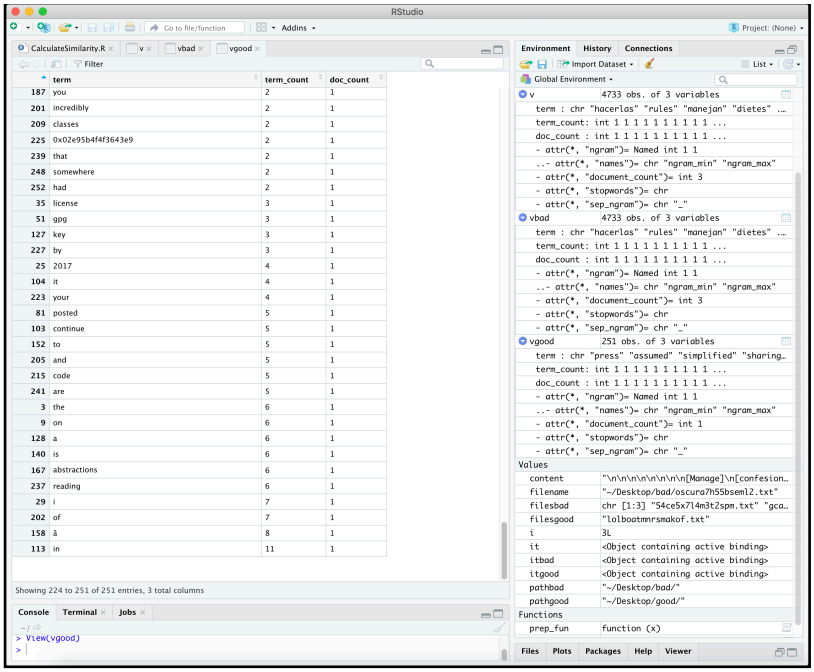


Figura 4.9. Frecuencia de palabras para archivo sin conductas delictivas.

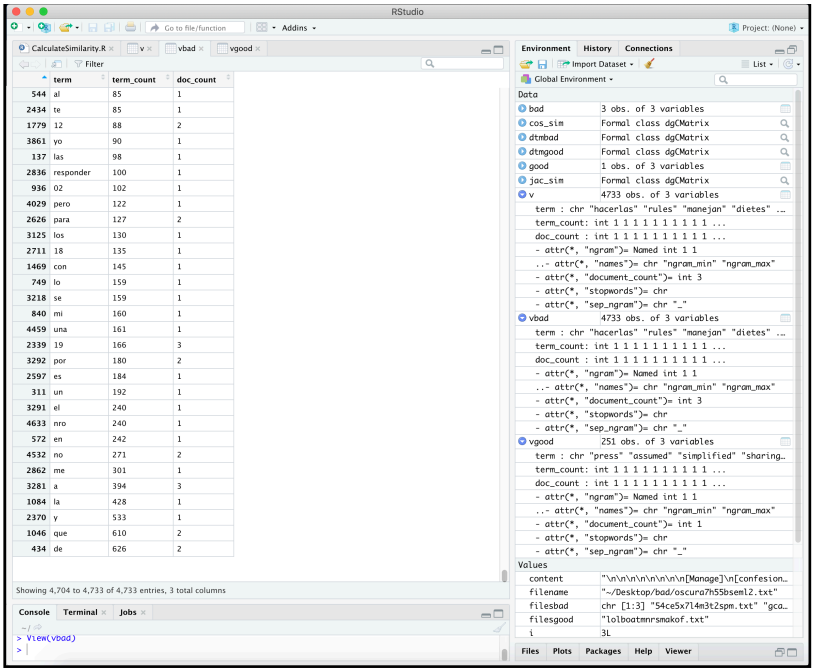


Figura 4.10. Frecuencia de palabras para archivos con conductas delictivas.

Por otro lado, si comparamos los 3 archivos con conductas delictivas, obtenemos lo siguiente:

```
> jac_sim
3 x 3 sparse Matrix of class "dgCMatrix"
           1          2          3
1 1.00000000 0.10933333 0.01367484
2 0.10933333 1.00000000 0.01000435
3 0.01367484 0.01000435 1.00000000
```

Aquí se puede interpretar que al comparar el primer archivo con el mismo su resultado es de 1, es decir idéntico, cuando se compara con el segundo tiene una similitud de 0.109 lo cual nos dice que no es muy similar, y el tercero es de 0.013 lo cual indica que es menos similar.

El resultado con Coseno es:

```
> cos_sim
3 x 3 sparse Matrix of class "dgCMatrix"
          1           2           3
1 1.00000000 0.09113093 0.23174007
2 0.09113093 1.00000000 0.04181909
3 0.23174007 0.04181909 1.00000000
```

Los resultados muestran que la similitud del primero contra el segundo archivo es de 0.09 siendo esta muy baja, y el primero contra el tercero es de 0.23 también, es decir hay poca similitud, pero más alto que el anterior lo cual contrasta con el resultado de Jaccard que fue a la inversa.

Una vez que se comprenden los resultados de estos valores de similitud, es interesante comparar archivos con una conducta delictiva similar para ver cómo se comporta Jaccard y Coseno. En este caso se utilizará un conjunto de archivos de texto con la conducta delictiva de armas contra la conducta delictiva de Carding, los resultados se mostrarán en el siguiente capítulo.

5.Resultados

5.1 Resultados presentados en 2019

En esta sección se presentan los resultados obtenidos con respecto a la metodología vista en el capítulo anterior, se cuenta con 14 redes onion de Armas, 15 redes onion de Carding, 42 redes onion de pornografía infantil (Tablas 5.1.1, 5.1.2 y 5.1.3), de los cuales se presentan sus totales en imágenes y en texto, tomando en cuenta que la información obtenida es

pública dentro de estas redes onion y no se pagó por suscripciones o registros para acceder al contenido:

El primer resultado obtenido es una página web con acceso restringido, la cual contiene la clasificación de conductas delictivas y no delictivas identificadas en la red privada de TOR, esto a través del estudio de Ontología de Delitos de Delincuencia Organizada en México, al que se supervisó y corrigió la información ingresada, (Figura 5.1.1), esta página se está usando por investigadores del área de Prevención de Delitos Electrónicos de la Dirección General Científica de la Guardia Nacional para consulta y gestión en investigaciones y ciberpatrullaje.

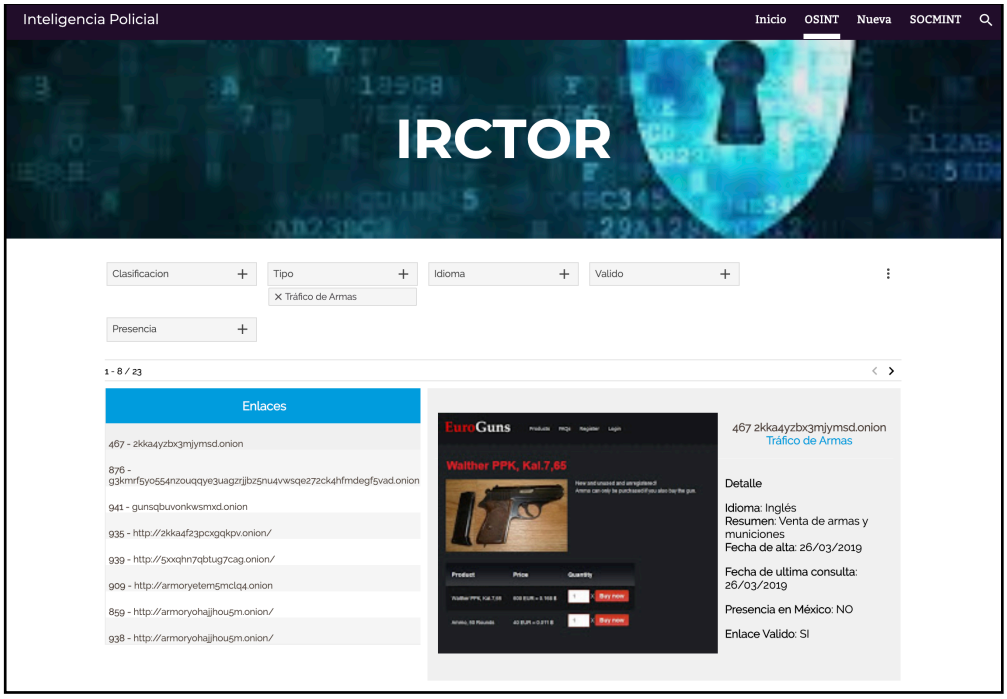


Figura 5.1.1. Página Web con conductas delictivas y no delictivas.

Red onion	Texto KB
2kka4f23pcxgqkpv.onion	8
5xxqhn7qbtug7cag.onion	25
darkgunweeep2epr.onion	19

Red onion	Texto KB
drkseidwayn6uc5x.onion	6
eugunnravvopif7i.onion	7
f6pxr3iqw7iziuc2.onion	272
gunsay5oixhyvj64.onion	7
gunsdtk47tolcrre.onion	9
gunsganjkiexjkew.onion	1
gunshopzpqbe4kgl.onion	92
gunstry2lzpvf47i.onion	7
luckp47s6xhz26rn.onion	7
pistolcqex2ecr5r.onion	10
weaponstrxqniqrt.onion	23
Total 14	493

Tabla 5.1.1. Redes onion de Armas.

Red onion	Texto KB
2222222faw2zy5t7.onion	45
2222222jukyyqtf6.onion	45
bucepafkui6lyblt.onion	2
cardsunwqrzhg5cw.onion	6
ccgalaxyoehif6qj.onion	36
creditclap4h3w6b.onion	12
ddrcb4qzjlv37e63.onion	2
fridumprbu2u4iys.onion	3
g5b5erkjomqen6nm.onion	29

Red onion	Texto KB
golden7djzq32zh4.onion	408
marketcvwplqswqq.onion	26
plasticmavm3fw7q.onion	26
vkjgulnzzgh5gnlf.onion	116
vkkzd55b7bidntmk.onion	38
xsqp76ka66qgue2s.onion	125
Total	919

Tabla 5.1.2. Redes onion de Carding.

Red onion	Imágenes	Texto KB
http://222222c6hrrxbydu.onion	50	1,291
http://3dboysn3o5d7dk3i.onion/	1	1
http://666666677563g5vi.onion	1	1
http://7hk64iz2vn2ewi7h.onion	8	507
http://abodxeycuklpva2v.onion/	1	2
http://akaoxb4ek4hm7vau.onion	29,192	1,260
http://bobt6q5lcdw6r466.onion/	55	3
http://childhsifechod5u.onion/	7	2
http://dadfuckiiqttgnjk.onion/	3	4
http://darksdsp6iexyidx.onion	20	1
http://dwpornbupmqnw4wv.onion	13	14
http://erolandgpitb7vjn.onion/	2,064	141
http://exclxmadj7tdy54l.onion/	27	11

Red onion	Imágenes	Texto KB
http://familybw6azkhjsc.onion	3	2
http://h3clhio3nera3sxw.onion/	164	187
http://imh3odalu2eanx5v.onion/	7	5
http://itu5h4f7shmamz2x.onion	1	6
http://jabber.jungswtfwgjwile2.onion/	1	1
http://jjclubumn7vkhyuw.onion/	6	117
http://lcpcs2q6y7umo426.onion/	10	15
http://litlg3cs3frsmh2j.onion/	50	101
http://lolitmhfkpif7sky.onion/	48	25
http://m4hzynbjgypfdqnq.onion/	5,009	142
http://magic5zudmw7witu.onion/	3	7
http://magiccp6ifgzlafa.onion	2	12
http://mda3nxsigriahnxq.onion/	6,989	552
http://newiomdqqdgtblp7.onion/	48	2
http://nh7zph33i3hkuurc.onion/	48	8
http://onicenyhxmhpaoyg.onion	3	36
http://pb327s7brxdgfqqk.onion/	4	107
http://pedohub2wdnufl2f.onion/	5	2
http://pedohubqgav4fubr.onion/	5	4
http://qxq4i6bu3ylmtueu.onion/	166	1,518
http://rapedbigpr4j7mhl.onion/	3	3
http://rfwtogljhrrzxyrl.onion/	1,734	197
http://sedsjdatpyp6swwq.onion/	20	68

Red onion	Imágenes	Texto KB
http://smallmf3tnhp5whg.onion	6,708	1
http://teen7tfmwtpsbfk3.onion/	4	2
http://teenxxxbtl7wsllp.onion	0	21
http://wrcp7wfcoakkr22d.onion/	11	4
http://xonion7ul44qs2aj.onion/	53	2,457
http://yiopwkqqh3vr65cv.onion/	4	2
http://yiswhxmewwothck6.onion	488	130
http://younga7oiicwmcex.onion	6	8
http://youngchgifiry3y2.onion	12	6
http://younggijdyyhuura.onion	0	3
http://younggIqkq62ovno.onion	0	1
http://zooscaqbvd5wfjul.onion/	11	5
Total	**53,068**	**8,995**

Tabla 5.1.3. Redes onion de pornografía infantil.

Para estudiar el comportamiento de estas conductas delictivas, se ocuparon 3 grupos de datos de conductas delictivas: armas y carding y pornografía infantil, se dividió en varios pasos el análisis de estos resultados, los cuales se listan a continuación:

1. En la figura 5.1.2, se muestran los archivos de cada conducta delictiva:

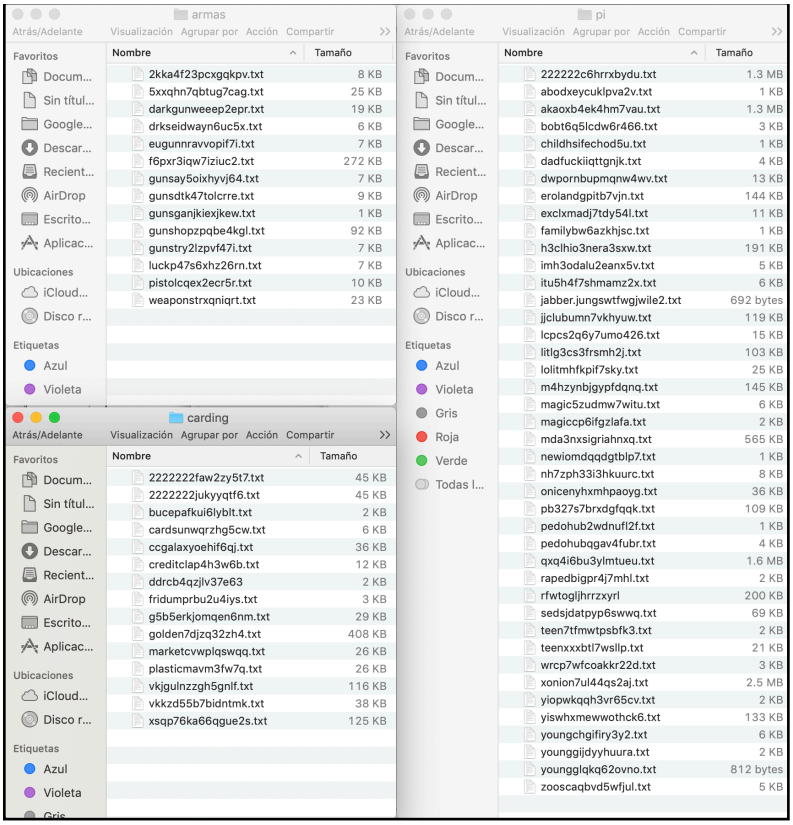
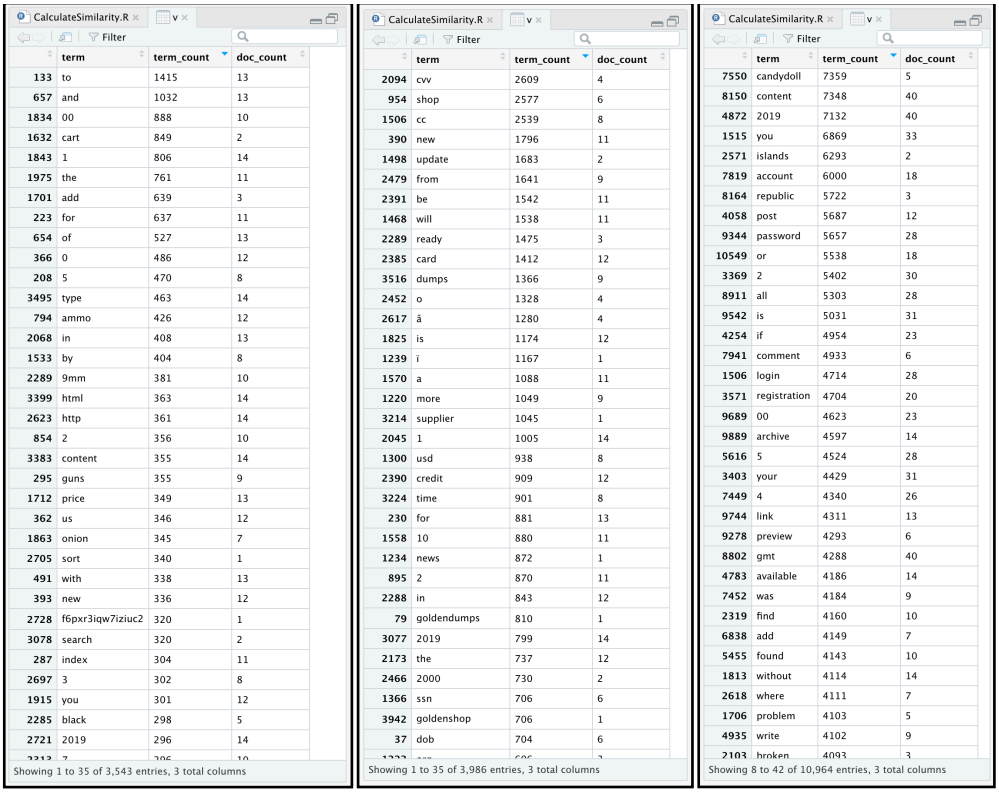


Figura 5.1.2. Archivos por conducta delictiva.

2. En la figura 5.1.3, se muestran los atributos a estudiar estadísticamente para el aprendizaje automático, que nos permitirá clasificar una nueva red onion en alguna de estas tres categorías o en ninguna: el primer campo es el consecutivo de los términos encontrados, donde un término se refiere a una palabra, número o símbolo especial; el segundo campo corresponde al término, el tercero a la frecuencia del término de cada conducta delictiva ; y el cuarto campo es el número de archivos en el que se encuentra el término, esto fue obtenido con el programa CalculateSimilarity.R en la herramienta R-Studio como se vio en el capítulo anterior.



CalculateSimilarity.R | v

Filter

	term	term_count	doc_count
133	to	1415	13
657	and	1032	13
1834	00	888	10
1632	cart	849	2
1843	1	806	14
1975	the	761	11
1701	add	639	3
223	for	637	11
654	of	527	13
366	0	486	12
208	5	470	8
3495	type	463	14
794	ammo	426	12
2068	in	408	13
1533	by	404	8
2289	9mm	381	10
3399	html	363	14
2623	http	361	14
854	2	356	10
3383	content	355	14
295	guns	355	9
1712	price	349	13
362	us	346	12
1863	onion	345	7
2705	sort	340	1
491	with	338	13
393	new	336	12
2728	f6pxr3iqw7iziuc2	320	1
3078	search	320	2
287	index	304	11
2697	3	302	8
1915	you	301	12
2285	black	298	5
2721	2019	296	14

Showing 1 to 35 of 3,543 entries, 3 total columns

CalculateSimilarity.R | v

Filter

	term	term_count	doc_count
2094	cvv	2609	4
954	shop	2577	6
1506	cc	2539	8
390	new	1796	11
1498	update	1683	2
2479	from	1641	9
2391	be	1542	11
1468	will	1538	11
2289	ready	1475	3
2385	card	1412	12
3516	dumps	1366	9
2452	o	1328	4
2617	â	1280	4
1825	is	1174	12
1239	ï	1167	1
1570	a	1088	11
1220	more	1049	9
3214	supplier	1045	1
2045	1	1005	14
1300	usd	938	8
2390	credit	909	12
3224	time	901	8
230	for	881	13
1558	10	880	11
1234	news	872	1
895	2	870	11
2288	in	843	12
79	goldendumps	810	1
3077	2019	799	14
2173	the	737	12
2466	2000	730	2
1366	ssn	706	6
3942	goldenshop	706	1
37	dob	704	6

Showing 1 to 35 of 3,986 entries, 3 total columns

CalculateSimilarity.R | v

Filter

	term	term_count	doc_count
7550	candydoll	7359	5
8150	content	7348	40
4872	2019	7132	40
1515	you	6869	33
2571	islands	6293	2
7819	account	6000	18
8164	republic	5722	3
4058	post	5687	12
9344	password	5657	28
10549	or	5538	18
3369	2	5402	30
8911	all	5303	28
9542	is	5031	31
4254	if	4954	23
7941	comment	4933	6
1506	login	4714	28
3571	registration	4704	20
9689	00	4623	23
9889	archive	4597	14
5616	5	4524	28
3403	your	4429	31
7449	4	4340	26
9744	link	4311	13
9278	preview	4293	6
8802	gmt	4288	40
4783	available	4186	14
7452	was	4184	9
2319	find	4160	10
6838	add	4149	7
5455	found	4143	10
1813	without	4114	14
2618	where	4111	7
1706	problem	4103	5
4935	write	4102	9
2103	broken	4093	3

Showing 8 to 42 of 10,964 entries, 3 total columns

Figura 5.3. Frecuencias por palabras de conductas delictivas de a) armas; b) Carding; c) Pornografía infantil.

3. Exportamos desde R.Studio los datos de cada conducta delictiva en archivos de Excel con el siguiente procedimiento:

> source('~/Desktop/R/CalculateSimilarityArmas.R')
[1] "Processing 1 2kka4f23pcxgqkpv.txt"
[1] "Processing 2 5xxqhn7qbtug7cag.txt"
[1] "Processing 3 darkgunweeep2epr.txt"
[1] "Processing 4 drkseidwayn6uc5x.txt"
[1] "Processing 5 eugunnravvopif7i.txt"
[1] "Processing 6 f6pxr3iqw7iziuc2.txt"
[1] "Processing 7 gunsay5oixhyvj64.txt"
[1] "Processing 8 gunsdtk47tolcrre.txt"
[1] "Processing 9 gunsganjkiexjkew.txt"
[1] "Processing 10 gunshopzpqbe4kgl.txt"
[1] "Processing 11 gunstry2lzpvf47i.txt"
[1] "Processing 12 luckp47s6xhz26rn.txt"
[1] "Processing 13 makegunq4r36mego.txt"
[1] "Processing 14 pistolcqex2ecr5r.txt"
[1] "Processing 15 weaponstrxqniqrt.txt"
> View(v)
> armas=v
> write.csv(armas,file="FrecArmas.csv")

> source('~/Desktop/R/CalculateSimilarityCarding.R')
[1] "Processing 1 2222222faw2zy5t7.txt"
[1] "Processing 2 2222222jukyyqtf6.txt"

[1] "Processing 3 bucepafkui6lyblt.txt"
[1] "Processing 4 cardsunwqrzhg5cw.txt"
[1] "Processing 5 ccgalaxyoehif6qj.txt"
[1] "Processing 6 creditclap4h3w6b.txt"
[1] "Processing 7 ddrcb4qzjlv37e63.txt"
[1] "Processing 8 fridumprbu2u4iys.txt"
[1] "Processing 9 g5b5erkjomqen6nm.txt"
[1] "Processing 10 golden7djzq32zh4.txt"
[1] "Processing 11 marketcvwplqswqq.txt"
[1] "Processing 12 plasticmavm3fw7q.txt"
[1] "Processing 13 vkjgulnzzgh5gnlf.txt"
[1] "Processing 14 vkkzd55b7bidntmk.txt"
[1] "Processing 15 xsqp76ka66qgue2s.txt"
> View(v)
> carding=v
> write.csv(carding,file="FrecCarding.csv")

> source('~/Desktop/R/CalculateSimilarityCardingpi.R')
> source('~/Desktop/R/CalculateSimilarityPI.R')
[1] "Processing 1 222222c6hrrxbydu.txt"
[1] "Processing 2 abodxeycuklpva2v.txt"
[1] "Processing 3 akaoxb4ek4hm7vau.txt"
[1] "Processing 4 bobt6q5lcdw6r466.txt"
[1] "Processing 5 childhsifechod5u.txt"
[1] "Processing 6 dadfuckiiqttgnjk.txt"
[1] "Processing 7 dwpornbupmqnw4wv.txt"
[1] "Processing 8 erolandgpitb7vjn.txt"
[1] "Processing 9 exclxmadj7tdy54l.txt"
[1] "Processing 10 familybw6azkhjsc.txt"
[1] "Processing 11 h3clhio3nera3sxw.txt"
[1] "Processing 12 imh3odalu2eanx5v.txt"
[1] "Processing 13 itu5h4f7shmamz2x.txt"
[1] "Processing 14 jabber.jungswtfwgjwile2.txt"
[1] "Processing 15 jjclubumn7vkhyuw.txt"
[1] "Processing 16 lcpcs2q6y7umo426.txt"
[1] "Processing 17 litlg3cs3frsmh2j.txt"
[1] "Processing 18 lolitmhfkpif7sky.txt"
[1] "Processing 19 m4hzynbjgypfdqnq.txt"
[1] "Processing 20 magic5zudmw7witu.txt"
[1] "Processing 21 magiccp6ifgzlafa.txt"
[1] "Processing 22 mda3nxsigriahnxq.txt"
[1] "Processing 23 newiomdqqdgtblp7.txt"
[1] "Processing 24 nh7zph33i3hkuurc.txt"
[1] "Processing 25 onicenyhxmhpaoyg.txt"
[1] "Processing 26 pb327s7brxdgfqqk.txt"
[1] "Processing 27 pedohub2wdnufl2f.txt"
[1] "Processing 28 pedohubqgav4fubr.txt"
[1] "Processing 29 qxq4i6bu3ylmtueu.txt"
[1] "Processing 30 rapedbigpr4j7mhl.txt"
[1] "Processing 31 rfwtogljhrrzxyrl.txt"
[1] "Processing 32 sedsjdatpyp6swwq.txt"
[1] "Processing 33 teen7tfmwtpsbfk3.txt"
[1] "Processing 34 teenxxxbtl7wsllp.txt"
[1] "Processing 35 wrcp7wfcoakkr22d.txt"
[1] "Processing 36 xonion7ul44qs2aj.txt"
[1] "Processing 37 yiopwkqqh3vr65cv.txt"
[1] "Processing 38 yiswhxmewwothck6.txt"
[1] "Processing 39 youngchgifiry3y2.txt"
[1] "Processing 40 younggijdyyhuura.txt"
[1] "Processing 41 youngglqkq62ovno.txt"
[1] "Processing 42 zooscaqbvd5wfjul.txt"
> View(v)
> pi=v

> write.csv(pi,file="FrecPI.csv")

---

4. El número de términos de armas es de 4,451, el número de términos de carding es de 3,987, y el número de términos de pornografía infantil es de 10,979.

En la figura 5.1.4, se muestra el programa de distribución realizado el programa Orange3.

En la figura 5.1.5, se observa la distribución de las palabras de armas.

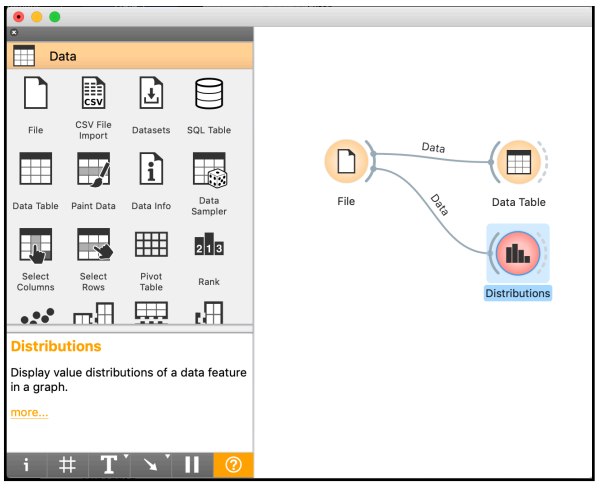


Figura 5.1.4. Programa en Orange3, con gráfica de Distribución.

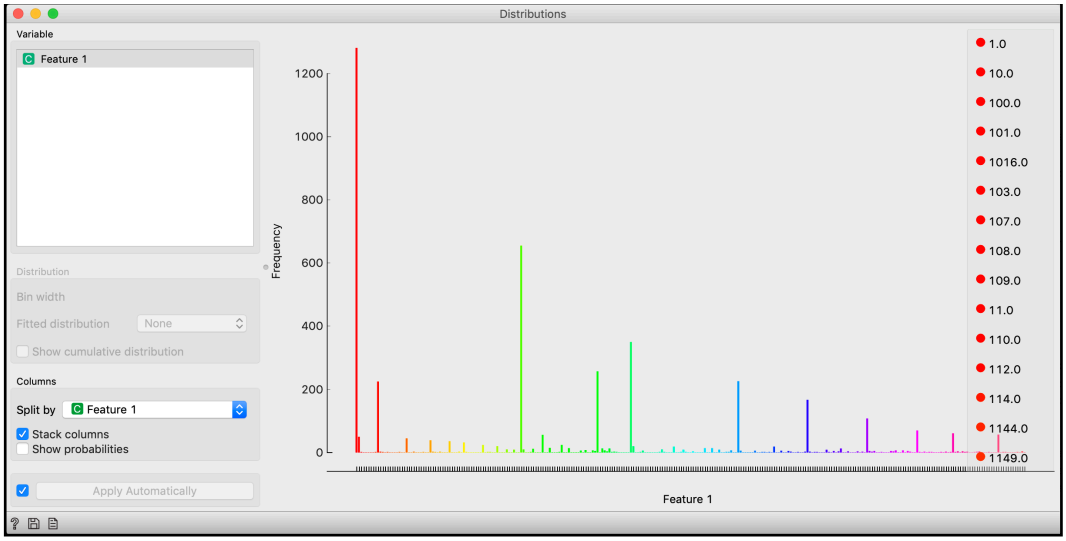


Figura 5.1.5. Distribución de la frecuencia de palabras de armas

En la figura 5.1.6, se observa la distribución de las palabras de carding.

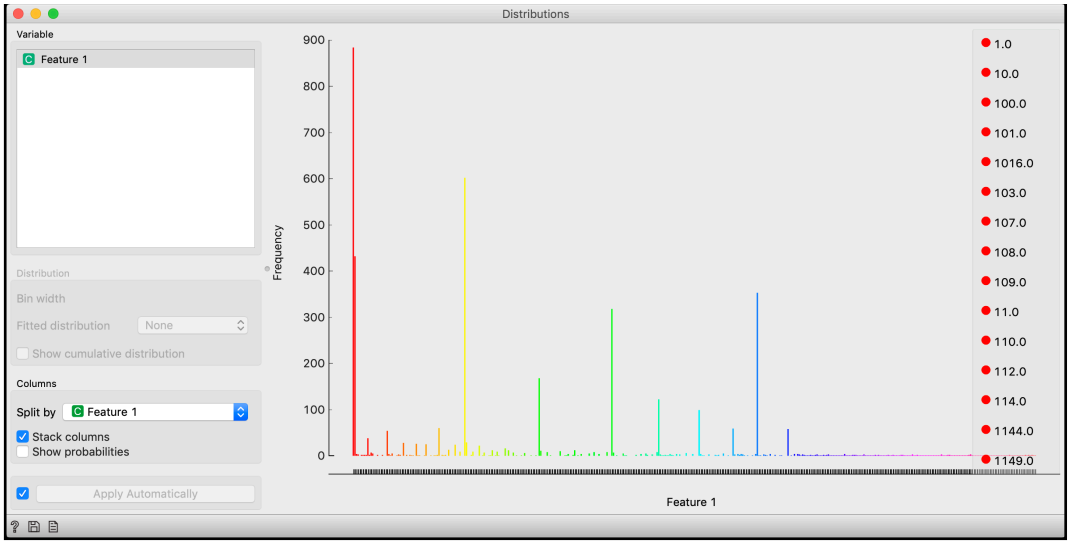


Figura 5.1.6. Distribución de la frecuencia de palabras de carding

En la figura 5.1.7, se observa la distribución de las palabras de armas.

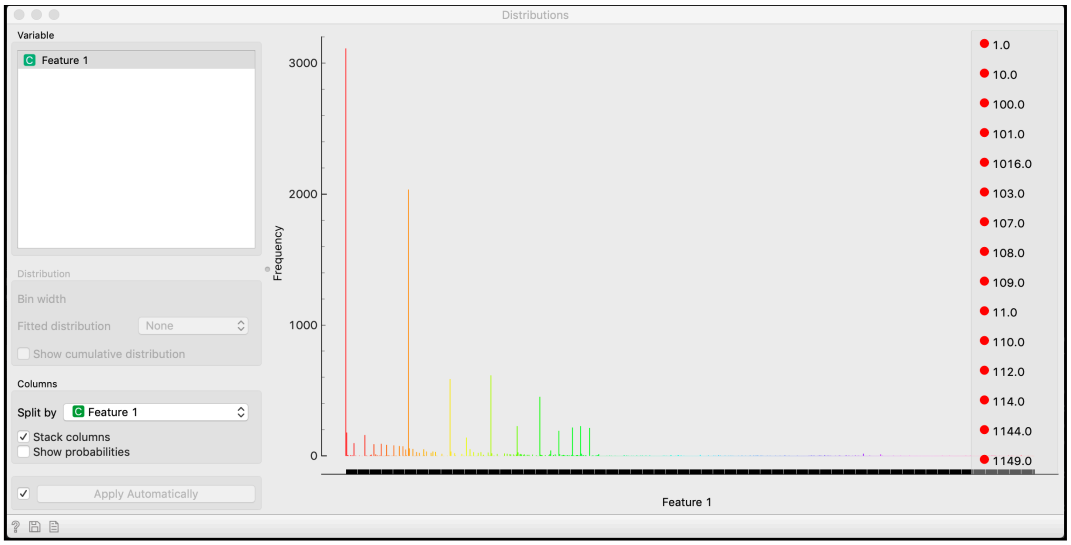


Figura 5.1.7. Distribución de la frecuencia de palabras de pornografía infantil.

5. La comparación entre la conducta delictiva de armas y carding tuvo el siguiente resultado, se ejecutó el programa CalculateSimilarity.R apuntando a las carpetas de armas y Carding, el primer cálculo se realizará con la Similitud de Jaccard, como se muestra a continuación:

```
> source('~/Desktop/R/CalculateSimilarity.R')
[1] "Processing 1 2kka4f23pcxgqkpv.txt"
[1] "Processing 2 5xxqhn7qbtug7cag.txt"
[1] "Processing 3 darkgunweeep2epr.txt"
[1] "Processing 4 drkseidwayn6uc5x.txt"
[1] "Processing 5 eugunnravvopif7i.txt"
[1] "Processing 6 f6pxr3iqw7iziuc2.txt"
[1] "Processing 7 gunsay5oixhyvj64.txt"
[1] "Processing 8 gunsdtk47tolcrre.txt"
[1] "Processing 9 gunsganjkiexjkew.txt"
[1] "Processing 10 gunshopzpqbe4kgl.txt"
[1] "Processing 11 gunstry2lzpvf47i.txt"
[1] "Processing 12 luckp47s6xhz26rn.txt"
[1] "Processing 13 pistolcqex2ecr5r.txt"
[1] "Processing 14 weaponstrxqniqrt.txt"
[1] "Processing 1 2222222faw2zy5t7.txt"
[1] "Processing 2 2222222jukyyqtf6.txt"
[1] "Processing 3 bucepafkui6lyblt.txt"
[1] "Processing 4 cardsunwqrzhg5cw.txt"
[1] "Processing 5 ccgalaxyoehif6qj.txt"
[1] "Processing 6 creditclap4h3w6b.txt"
[1] "Processing 7 ddrcb4qzjlv37e63.txt"
[1] "Processing 8 fridumprbu2u4iys.txt"
[1] "Processing 9 g5b5erkjomqen6nm.txt"
[1] "Processing 10 golden7djzq32zh4.txt"
[1] "Processing 11 marketcvwplqswqq.txt"
[1] "Processing 12 plasticmavm3fw7q.txt"
[1] "Processing 13 vkjgulnzzgh5gnlf.txt"
[1] "Processing 14 vkkzd55b7bidntmk.txt"
[1] "Processing 15 xsqp76ka66qgue2s.txt"
> jac_sim
15 x 15 sparse Matrix of class "dgCMatrix"
  [[ suppressing 15 column names ‘1’, ‘2’, ‘3’ ...]]

1  1.00000000 0.98613518 0.07096774 0.11661808 0.10786699 0.13738441 0.06636501
2  0.98613518 1.00000000 0.07073955 0.12116788 0.10948905 0.13852243 0.06616541
3  0.07096774 0.07073955 1.00000000 0.22746781 0.06235012 0.12058824 0.18229167
4  0.11661808 0.12116788 0.22746781 1.00000000 0.12272727 0.24484536 0.14236111
5  0.10786699 0.10948905 0.06235012 0.12272727 1.00000000 0.20066519 0.06536697
6  0.13738441 0.13852243 0.12058824 0.24484536 0.20066519 1.00000000 0.11578947
7  0.06636501 0.06616541 0.18229167 0.14236111 0.06536697 0.11578947 1.00000000
8  0.11695138 0.11811024 0.10778443 0.21025641 0.19111111 0.20638298 0.10427807
```

9 0.11219081 0.11297440 0.05270270 0.11802030 0.82614057 0.22613065 0.06476684
10 0.09825103 0.09927984 0.03243918 0.07134146 0.15994094 0.08749266 0.04235727
11 0.10792079 0.10990099 0.07575758 0.15062112 0.12973884 0.17415730 0.10000000
12 0.10934394 0.11133201 0.07432432 0.14953271 0.13006757 0.17489422 0.10048622
13 0.10985117 0.11126860 0.05130687 0.10924370 0.16406250 0.14655943 0.05909944
14 0.13649564 0.13513514 0.07264297 0.15697674 0.23978686 0.22222222 0.08223201
15 0.09629630 0.09743590 0.04034582 0.08643710 0.16125541 0.11632653 0.04424157

1 0.11695138 0.11219081 0.09825103 0.10792079 0.10934394 0.10985117 0.13649564
2 0.11811024 0.11297440 0.09927984 0.10990099 0.11133201 0.11126860 0.13513514
3 0.10778443 0.05270270 0.03243918 0.07575758 0.07432432 0.05130687 0.07264297
4 0.21025641 0.11802030 0.07134146 0.15062112 0.14953271 0.10924370 0.15697674
5 0.19111111 0.82614057 0.15994094 0.12973884 0.13006757 0.16406250 0.23978686
6 0.20638298 0.22613065 0.08749266 0.17415730 0.17489422 0.14655943 0.22222222
7 0.10427807 0.06476684 0.04235727 0.10000000 0.10048622 0.05909944 0.08223201
8 1.00000000 0.18550369 0.12737293 0.14424411 0.14166667 0.11383538 0.17804552
9 0.18550369 1.00000000 0.14562118 0.12694064 0.12728938 0.16736111 0.24782188
10 0.12737293 0.14562118 1.00000000 0.11052632 0.10778128 0.13241807 0.14368727
11 0.14424411 0.12694064 0.11052632 1.00000000 **0.78852459** 0.13142438 0.17484663
12 0.14166667 0.12728938 0.10778128 **0.78852459** 1.00000000 0.13088235 0.17297851
13 0.11383538 0.16736111 0.13241807 0.13142438 0.13088235 1.00000000 0.23625097
14 0.17804552 0.24782188 0.14368727 0.17484663 0.17297851 0.23625097 1.00000000
15 0.09616678 0.15852273 0.13822725 0.11182670 0.11137163 0.69386282 0.21140732

1 0.09629630
2 0.09743590
3 0.04034582
4 0.08643710
5 0.16125541
6 0.11632653
7 0.04424157
8 0.09616678
9 0.15852273
10 0.13822725
11 0.11182670
12 0.11137163
13 0.69386282
14 0.21140732
15 1.00000000
>

Como podemos observar vamos a comparar 14 archivos de armas contra 15 archivos de Carding, al comparar los valores entre los archivos vemos que la similitud entre estas dos conductas delictivas es muy baja, rondando entre 0.0403 y 0.211 la mayoría de los valores. ahora procedemos a aplicar la similitud de coseno obteniendo lo siguiente:

```
> cos_sim
15 x 15 sparse Matrix of class "dgCMatrix"
  [[ suppressing 15 column names ‘1’, ‘2’, ‘3’ ...]]

1  1.00000000 0.99993341 0.08239628 0.1457955 0.1257251 0.2185829 0.04495534
2  0.99993341 1.00000000 0.08172240 0.1461511 0.1257111 0.2185076 0.04508745
3  0.08239628 0.08172240 1.00000000 0.5451372 0.1163284 0.2767280 0.25299725
4  0.14579548 0.14615110 0.54513722 1.0000000 0.2932853 0.5405926 0.20793898
5  0.12572514 0.12571109 0.11632840 0.2932853 1.0000000 0.4574202 0.11898841
6  0.21858290 0.21850761 0.27672800 0.5405926 0.4574202 1.0000000 0.14536248
7  0.04495534 0.04508745 0.25299725 0.2079390 0.1189884 0.1453625 1.00000000
8  0.13650876 0.13655420 0.21176529 0.4069401 0.5631549 0.5107290 0.10773258
9  0.12863744 0.12863932 0.08618648 0.2964465 0.9701087 0.4974246 0.08870421
10 0.09499757 0.09511643 0.05847403 0.1050908 0.2106260 0.1572368 0.08850978
11 0.19232369 0.19246837 0.34817848 0.4180680 0.2302590 0.4056542 0.61085291
12 0.19237558 0.19251657 0.35034164 0.4174923 0.2320882 0.4061302 0.61208890
13 0.08430652 0.08417447 0.07358292 0.1181976 0.1467222 0.1475845 0.11100338
14 0.20929752 0.20930311 0.31348331 0.5938564 0.5423944 0.6771489 0.13888863
15 0.09028335 0.09016292 0.08397378 0.1389065 0.1693596 0.1745357 0.11152533

1  0.1365088 0.12863744 0.09499757 0.1923237 0.1923756 0.08430652 0.2092975
2  0.1365542 0.12863932 0.09511643 0.1924684 0.1925166 0.08417447 0.2093031
3  0.2117653 0.08618648 0.05847403 0.3481785 0.3503416 0.07358292 0.3134833
4  0.4069401 0.29644649 0.10509076 0.4180680 0.4174923 0.11819762 0.5938564
5  0.5631549 0.97010872 0.21062602 0.2302590 0.2320882 0.14672223 0.5423944
6  0.5107290 0.49742463 0.15723676 0.4056542 0.4061302 0.14758454 0.6771489
7  0.1077326 0.08870421 0.08850978 0.6108529 0.6120889 0.11100338 0.1388886
8  1.0000000 0.59772076 0.31668219 0.2880035 0.2883640 0.13339390 0.6164361
9  0.5977208 1.00000000 0.21027051 0.2103520 0.2111034 0.15652575 0.5811662
10 0.3166822 0.21027051 1.00000000 0.1148085 0.1195149 0.07506019 0.2292809
11 0.2880035 0.21035201 0.11480848 1.0000000 0.9965831 0.13601940 0.4175297
12 0.2883640 0.21110340 0.11951490 0.9965831 1.0000000 0.13798020 0.4164329
13 0.1333939 0.15652575 0.07506019 0.1360194 0.1379802 1.00000000 0.1757675
        14 0.6164361 0.58116623 0.22928088 0.4175297 0.4164329 0.17576747 1.0000000
15 0.1588681 0.18083422 0.08577464 0.1450484 0.1469961 0.99718245 0.2089370
```

```
1  0.09028335
2  0.09016292
3  0.08397378
4  0.13890652
5  0.16935964
6  0.17453568
7  0.11152533
8  0.15886814
9  0.18083422
10 0.08577464
11 0.14504835
12 0.14699608
13 0.99718245
14 0.20893704
15 1.00000000
>
```

Al aplicar la similitud de coseno se obtiene una similitud entre 0.0449 y 0.597, en este caso se observa que los valores son más altos con respecto a la similitud de Jaccard.

6. En las figuras 5.1.8, 5.1.9 y 5.1.10 se observan el gráfico de dispersión de armas, de carding y de Pornografía Infantil:

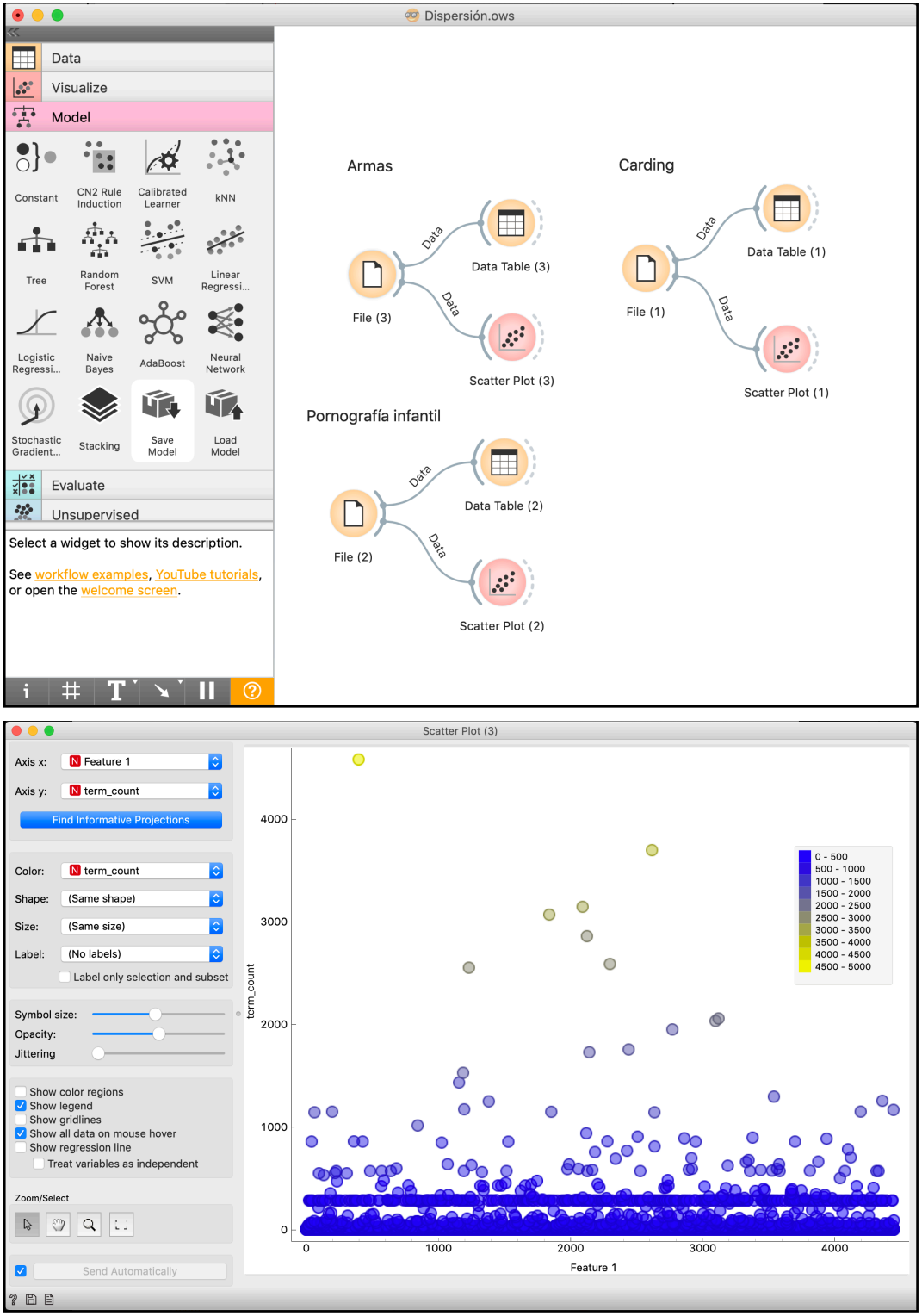


Figura 5.1.8. a) Programa de dispersión en R-Studio, b) Gráfica de la dispersión en la conducta delictiva de Armas.

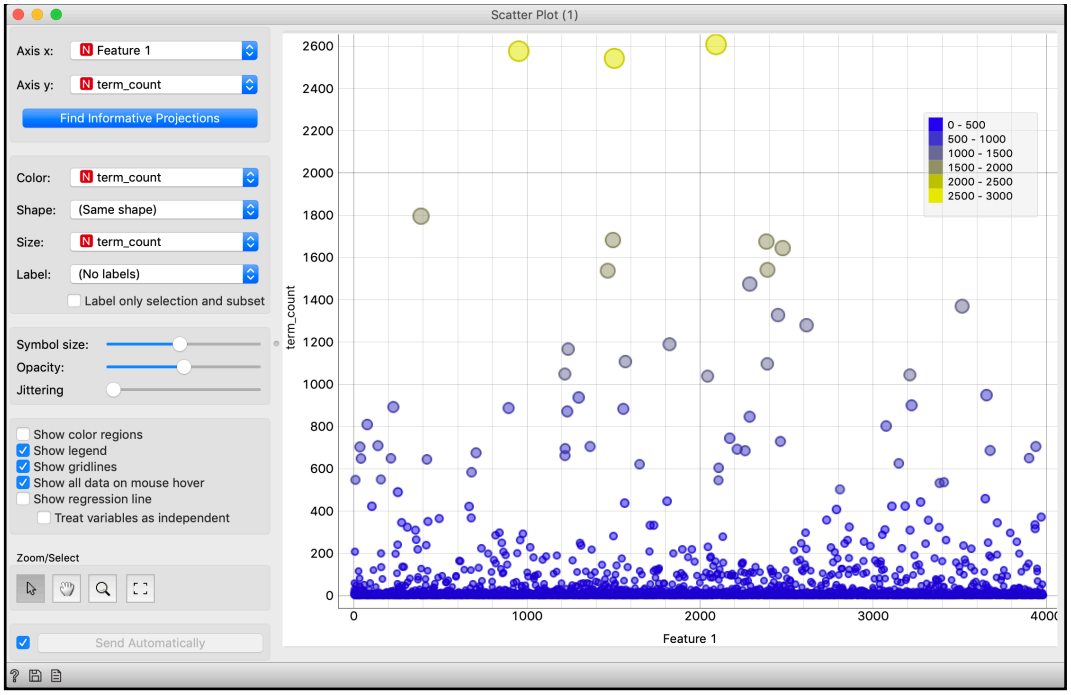


Figura 5.1.9. Gráfica de la dispersión en la conducta delictiva de Carding.

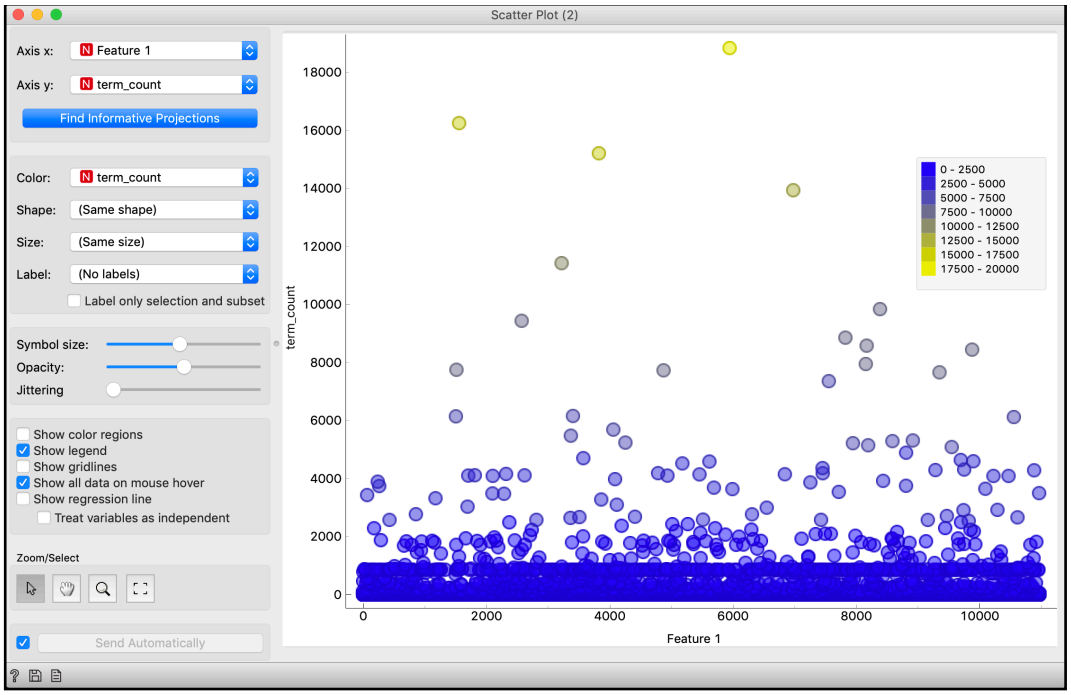


Figura 5.1.10. Gráfica de la dispersión en la conducta delictiva de Pornografía Infantil.

Observamos que el patrón es parecido aun que los rangos son distintos y la dispersión es distinta de los términos de mayor frecuencia en contrario de los que menos frecuencia tienen.

7. En la figura 5.1.11, 5.1.12 y 5.1.13, se muestran los términos ya pre procesados de armas, Carding y pornografía infantil.

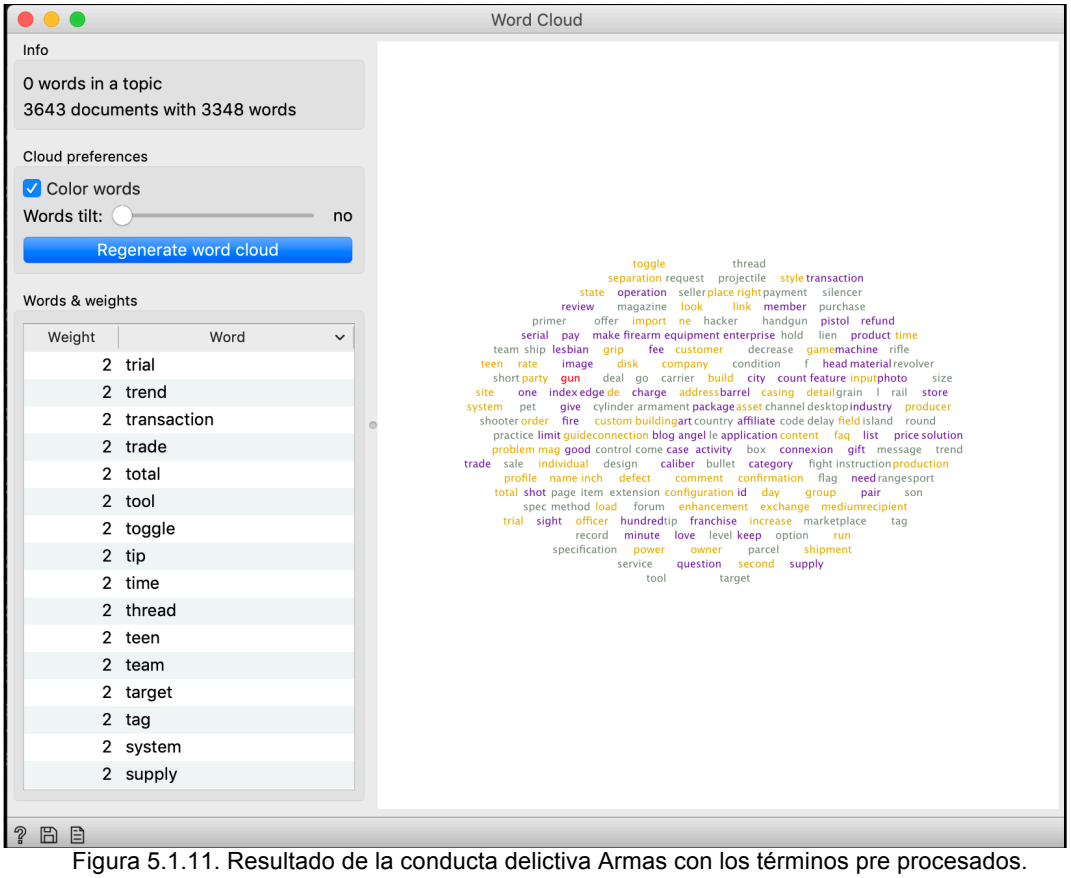


Figura 5.1.11. Resultado de la conducta delictiva Armas con los términos pre procesados.

Figura 5.1.12. Resultado de la conducta delictiva de Carding con los términos pre procesados.

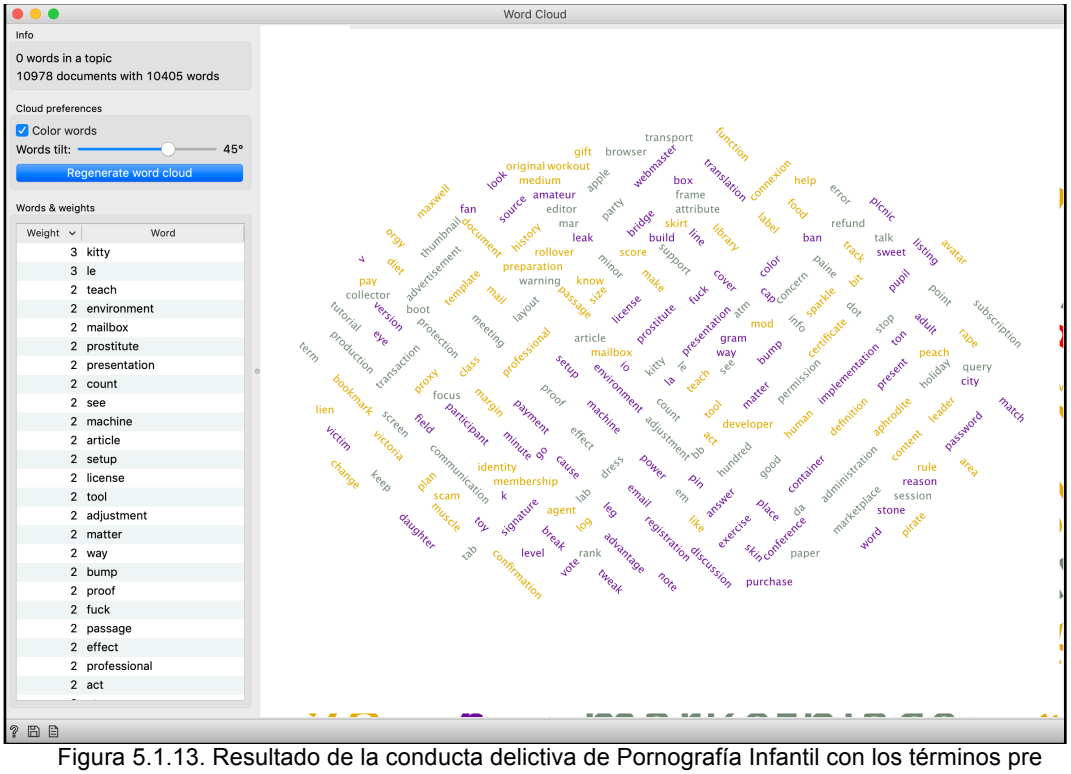


Figura 5.1.13. Resultado de la conducta delictiva de Pornografía Infantil con los términos pre procesados.

5.2 Resultados presentados en 2022

En esta sección se muestran los resultados de la identificación y clasificación de las conductas delictivas en TOR, dirigido para obtener información que sirva en las investigaciones que se llevan a cabo en el área de Ciberdelitos de la Dirección General Científica de la Guardia Nacional. En la figura 5.2.1. Podemos observar una página web la cual contiene el resultado de lo recopilado de páginas en TOR con conductas delictivas y no delictivas siguiendo la metodología del capítulo 4, desde enero y febrero de 2022 a la fecha. El investigador puede realizar filtros por Tipo y Subtipo de Conducta Delictiva como se puede apreciar en las figuras 5.2.2 y 5.2.3, en la figura 5.2.4 se muestra el filtro por Idioma y por fechas de alta de la página en el sistema y última consulta de la misma, así como los datos procesados de la página en cuestión.

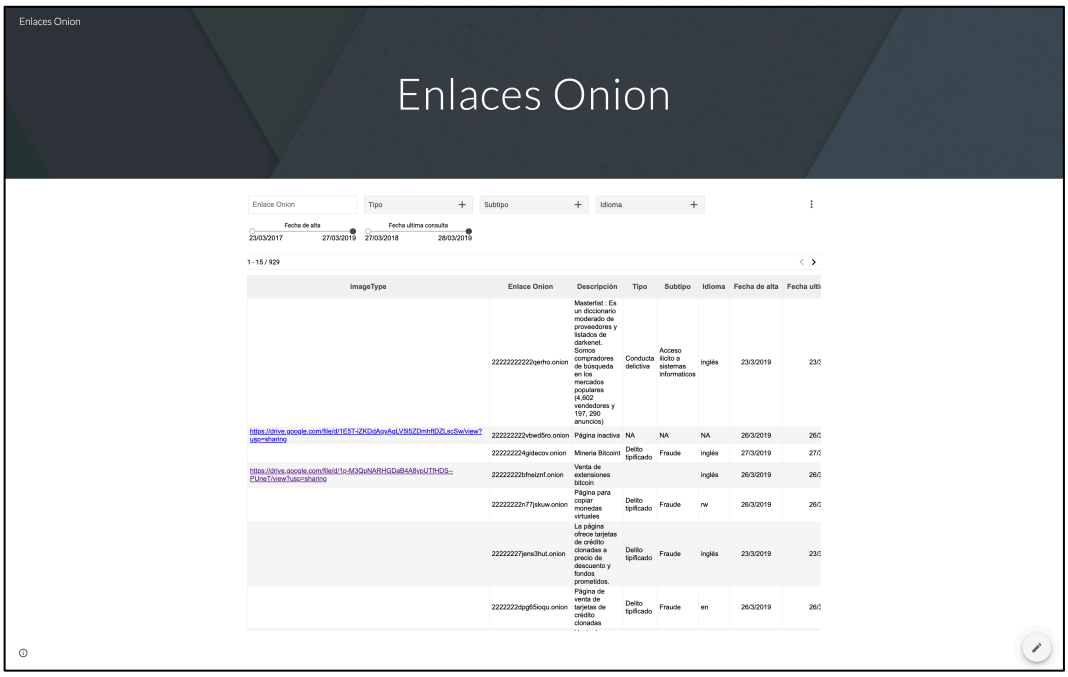


Figura 5.2.1. Resultados de páginas web en la red Tor con conductas delictivas y no delictivas.

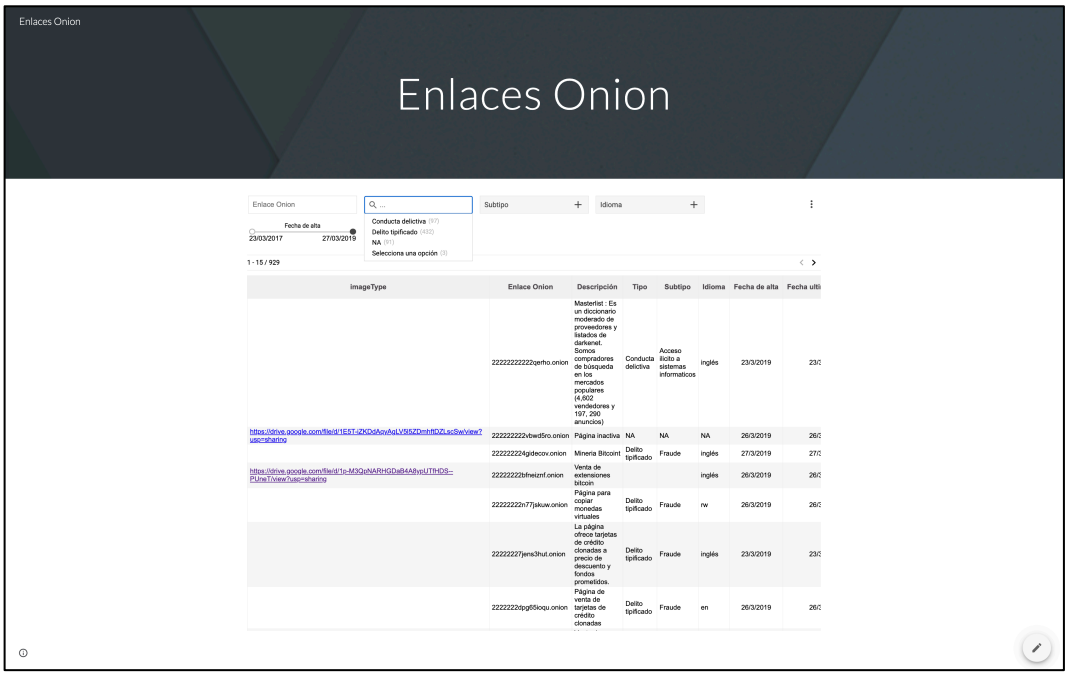


Figura 5.2.2. Filtro por Tipo de Conducta Delictiva.

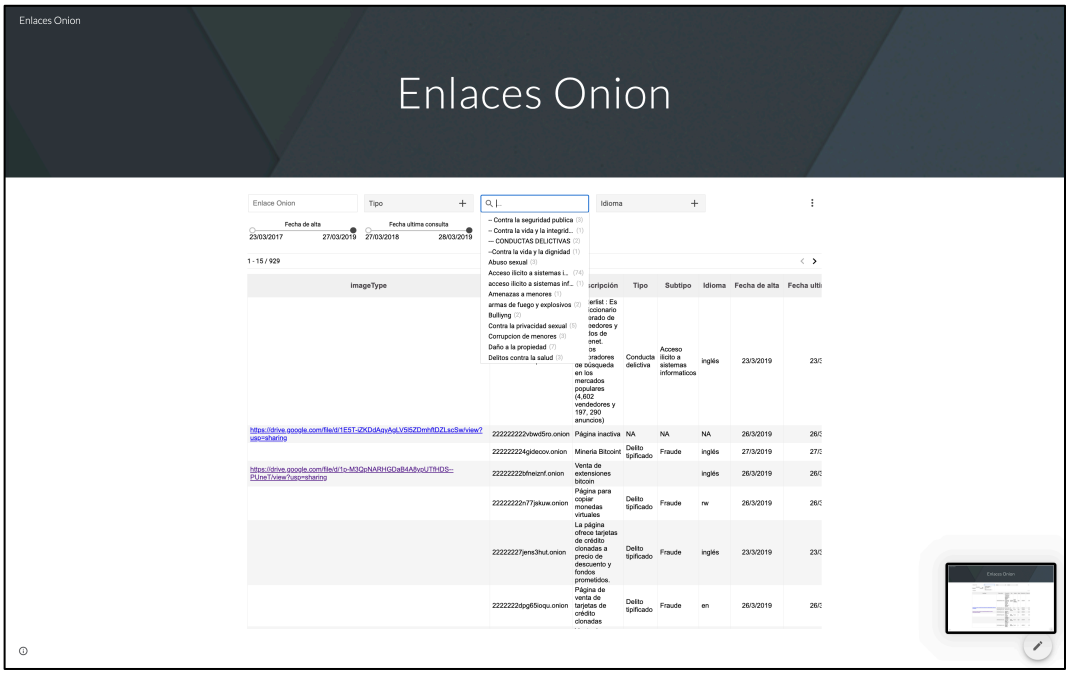


Figura 5.2.3. Filtro por Subtipo de Conducta Delictiva.

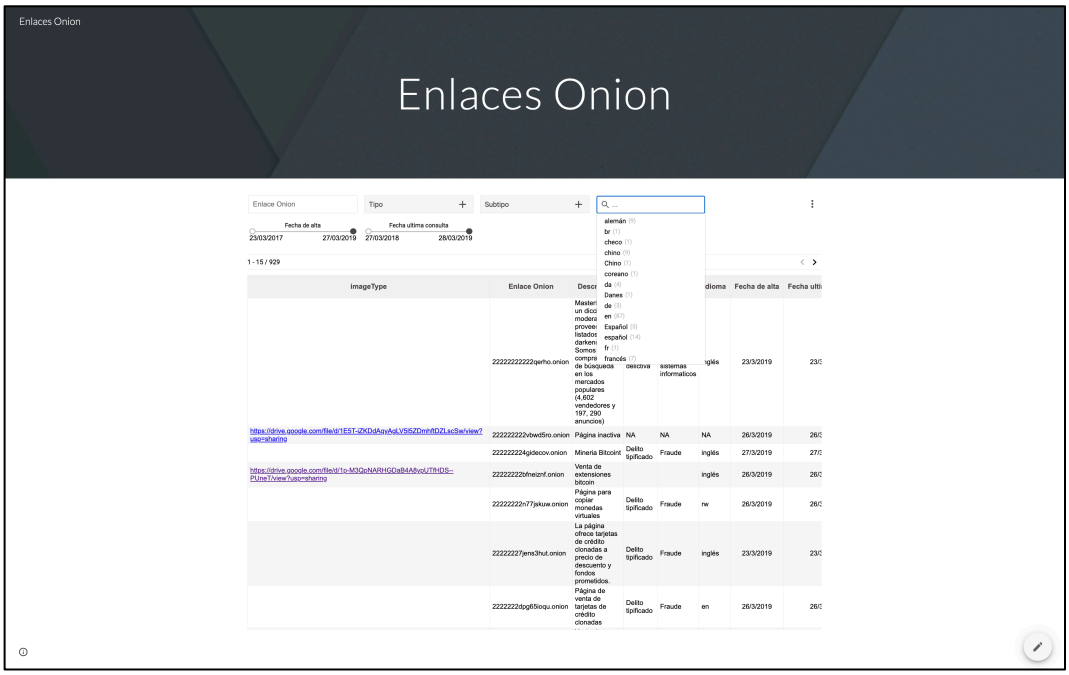


Figura 5.2.4. Filtro por Idioma en la página Web.

De la misma forma, se creó una página web con las estadísticas resultantes en la investigación en el periodo del 1º de enero al 22 de febrero del 2022, las cuales podemos apreciar en la Figura 5.2.5.

1. En la primera gráfica se muestran el número total de redes onion por día en un periodo del 1º de enero al 22 de febrero del 2022, se puede apreciar que la primera semana de enero tiene un promedio de 2,386 redes, sin embargo, a partir de la segunda semana y has-

ta el 15 de febrero baja la cantidad de páginas a 2,065, para volver a subir hasta 2,422 redes.

2. En la segunda gráfica se muestra la frecuencia con que se repiten las redes onion dentro de la tabla de datos, siendo 53 la más alta y 6 la más baja.
3. La tercera gráfica presenta el número de redes por idioma, resultando en un 84.2% de redes en inglés con 95,925, cabe resaltar que en el idioma de español se cuenta con 465 páginas, siendo esto el 0.4% del total.
4. La cuarta gráfica muestra el estatus de código de estado de las páginas estudiadas en este periodo, resultando un total de 109,306 redes con un estatus de éxito al ser consultadas, un total de 4,671 páginas con algún tipo de error en la solicitud como solicitud de contraseñas, permisos, no existencia de la página, entre otros y 260 páginas con errores del servidor que aloja las páginas.
5. Por último, en la quinta gráfica se puede ver la frecuencia de los títulos de las redes onion, siendo la de “Onion Dir – Adult” la que más se repite con 45,212 veces, seguida por las redes de Facebook por idioma con 1,866 veces, cabe destacar que el título de “LoliPorn” que contiene el delito de pornografía infantil se repite 510 veces.

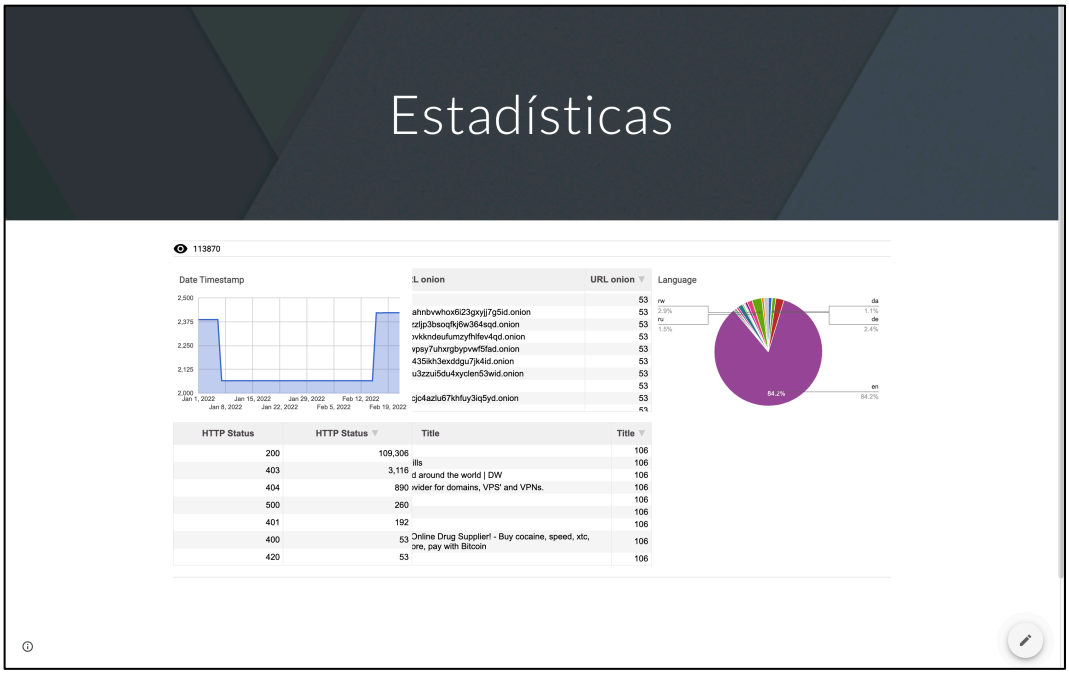


Figura 5.2.5. Gráficas de resultados.

En base a este análisis se encontraron las siguientes conductas delictivas:

URL		Conducta Delictiva
http://onionsnunpziixmcunux46t3t7rjcvt5pjy6a52kaoaidcwimqarjid.onion	Onion Dir - Adult	Contra la Libertad Sexual

fastdogsealtbyvd.onion	Store	Contra la salud
http://hackedabm6ejvmh7xiiqdq3ptflorhdvjurhheqbovlu64xi3zbrlpad.onion/	Hacked databases store	Filtración de datos
silkroadkaxmspva.onion	Silk Road 4	Delitos contra las Tecnologías de la Información
jn6weomv6klvnwdwcgu55miabpwklsmmyaf5qrkt4miif4shrqmvdhqd.onion	Rent-A-Hacker - Hire a hacker for every job you can imagine, from DDOS to completely ruining people or destroy reputation of a company or individual	Delitos contra las Tecnologías de la Información
marketcvwplqswqq.onion	Plastic Marketplace	Contra la Seguridad Financiera
plasticuou2uakdu.onion	Plastic Money | Your easy way to get money	Contra la Seguridad Financiera
6bdg5y5ab55p6jdojtl2llv4rm5gvykbz5isdlwkcq7kxy6l22jb2aad.onion	CREDITCARD CENTER	Contra la Seguridad Financiera
http://countrfazzak77uykjhrje5bj6pmzrdnacziws5ncpsszfdk3przgfyd.onion/	Euro & USD Counterfeits	Contra la Seguridad Financiera
http://k6m3fagp4w4wspmdt23fldnwrmknse74gmxosswvaxf3ciasficpenad.onion/	UK Guns and Ammo Store - Buy guns and ammo in the UK for Bitcoin.	Contra laSalud
ccpalsto5gglun22.onion	ccPal - CCs, CVV2s, PayPals, Ebay accounts and more - buy stolen creditcards with bitcoin	Contra la Seguridad Financiera
crackersccqxsmdb.onion	PirateCRACKERS | Home	Delitos contra las Tecnologías de la Información
http://cfactorbod7y6vmeexqldozgtlzxq4fjbpe4pridcdn3barehdfrz5ad.onion	Counterfeit Factory	Contra la Seguridad Financiera
dumpsmarslfo4h5g.onion	Dumps Market	Contra la Seguridad Financiera
accountlwghunkeg.onion	PAYPAL	Contra la Seguridad Financiera
weapon5cd6o72mny.onion	Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since 2001	Contra laSalud
streetddxedw5thp.onion	21 Dump Street	Contra la Seguridad Financiera
platypus77f3ujfw.onion	KryptoPlatypus • PayPal Cashout Service	Contra la Seguridad Financiera
stationd-mosxiwo63u4ymtgso6ak2vuuuydp7crssnwhklkns75obaqd.onion	Child Porno Station | Curated PTHC and child porno collections | Jailbait photos and videos | Lolitas | Lolilust | pedo porno | little daughters | join now	Contra la Libertad Sexual
san-tat7kpllt6iyvqbr7q4amdv6dzrh6paatvyrzl7ry3zm72zigf4ad.onion	HOME | CL0P^_- LEAKS	Delitos contra las Tecnologías de la Información
replicaf6cjadwxs.onion	HQER - High Quality Euro Counterfeits - best counterfeit bank notes in europe	Contra la Seguridad Financiera
hackergruemqvew6.onion	Hacking	Delitos contra las Tecnologías de la Información
mollyworup44gri7.onion	Mollyworld No.1 Provider of Crystals and Pills	Contra la Salud
pzaboystoravp2rz.onion	PZA Boy Stories	Contra la Libertad Sexual
5figq755l7c55eopjphypkpfj5b4ap5nm6rvie2tygcwtafbjhv3p3id.onion	PedoBum | Upload photos	Contra la Libertad Sexual
drugszun7tvsgsaa.onion	Peoples Drug Store - The Darkweb's Best Online Drug Supplier! - Buy cocaine, speed, xtc, mdma, heroin and more at peoples drug store, pay with Bitcoin	Contra la Salud

6.Conclusiones

Se desarrolló una metodología de identificación, recuperación, clasificación y presentación de los datos que ofrece la Dark Web a partir de la red TOR para descubrir datos relevantes de delitos cibernéticos que ayuden en el proceso de investigación de autoridades competentes.

Esta metodología puede aplicarse a diversos escenarios con otras redes de la Dark Web, como por ejemplo i2p, Freenet, Ares, Emule, entre otras, e incluso en páginas de la red indexada como RaidForums que es un sitio de filtraciones de información sustraída a empresas y gobiernos, de tal forma que:

1. Se estudien e implementen mecanismos para identificar páginas o redes objetivo.
2. Investigar, diseñar e implementar modelos para extraer la información contenida en estas redes objetivo.
3. Se utilice la Ontología de términos diseñada en esta tesis o se elabore una nueva a partir de un documento como una Ley local o Federal, que ayude a comprender las conductas delictivas que se quieran clasificar posteriormente.
4. Modelar los datos obtenidos para su posterior clasificación.
5. Utilizar el algoritmo de clasificación de Jaccard y Coseno utilizado en esta tesis para comparar la similitud de los datos obtenidos y modelados u otro algoritmo que permita llevar a cabo su clasificación.
6. Presentación de resultados, utilizando herramientas como R-Studio en la cual se obtiene el programa CalculateSimilarity.R para saber la frecuencia de los términos en las diferentes redes onion y se realizó la comparación entre la conducta delictiva de armas y carding utilizando la Similitud de Jaccard y Coseno con el programa CalculateSimilarity.R apuntando a las carpetas de armas y Carding, se obtiene que la similitud entre estas dos conductas delictivas es

muy baja, también se utilizaron herramientas como Orange3 para presentar la distribución de la frecuencia de palabras (en este caso las palabras fueron armas, carding, pornografía infantil), además gráficos de dispersión de armas, de carding y de Pornografía Infantil. Por último se crean dos páginas web, la primera contiene el resultado de lo recopilado de páginas en TOR con conductas delictivas y no delictivas siguiendo la metodología del capítulo 4, desde enero y febrero de 2022 a la fecha obteniendo filtros por Tipo y Subtipo de Conducta Delictiva, por Idioma y por fechas de alta de la página en el sistema y última consulta de la misma, así como los datos procesados de la página en cuestión, la segunda página contiene las estadísticas resultantes en la investigación, en la primera gráfica se muestran el número total de redes onion, en la segunda gráfica se muestra la frecuencia con que se repiten las redes onion dentro de la tabla de datos, la tercera gráfica presenta el número de redes por idioma, la cuarta gráfica muestra el estatus de código de estado de las páginas estudiadas en este periodo, Por último, en la quinta gráfica se puede ver la frecuencia de los títulos de las redes onion, y se muestra una tabla con las redes con conductas delictivas que mas se repiten.

Lo anterior sirve para que investigadores de ciberdelitos puedan obtener información de conductas delictivas en la Red de TOR, que sirvan como líneas de investigación, ciberinteligencia e incluso junto con agencias internacionales persecutoras de delitos cibernéticos, incautar estas páginas de la Deep Web.

Anexo 1

Para implementar este proyecto podemos seguir el siguiente procedimiento, recordando que corre sobre el Sistema Operativo de Linux:

I. Descargar e Instalar el proyecto de OnionScan de la página: https://github.com/s-rah/onionscan
II. Ejecutar el programa de SelekTOR, que brinda una interface gráfica para gestionar y controlar los proxys a los que tenemos que conectarnos para poder navegar a través de Tor. (Figura A1.1).
III. Ejecutar el programa de Python (mostrado al final del Anexo 1) para escanear redes onion y obtenerlas en un archivo de texto *"onion_master_list.txt"*, como se puede ver en la figura A1.2, a mayo de 2018 el número de redes onion resultó de 8,386.
IV. Muchas de estas redes onion ya no se encuentran activas y otras si lo están, por lo que se procedió a elaborar un programa en Python con el nombre *"VerifyOnionSite.py"* el cuál probará cuales de estas redes funcionan y cuales no y las activas las escribe en el archivo *"onionVerifiedList.txt"*, a este programa lo llamará el Script "*verifyListTor.sh*" (Figura A1.3), el cual controla el inicio y el fin de la lista de redes onion del archivo de texto. El resultado a mayo de 2018 de la verificación de redes onion activas es de 1,143.

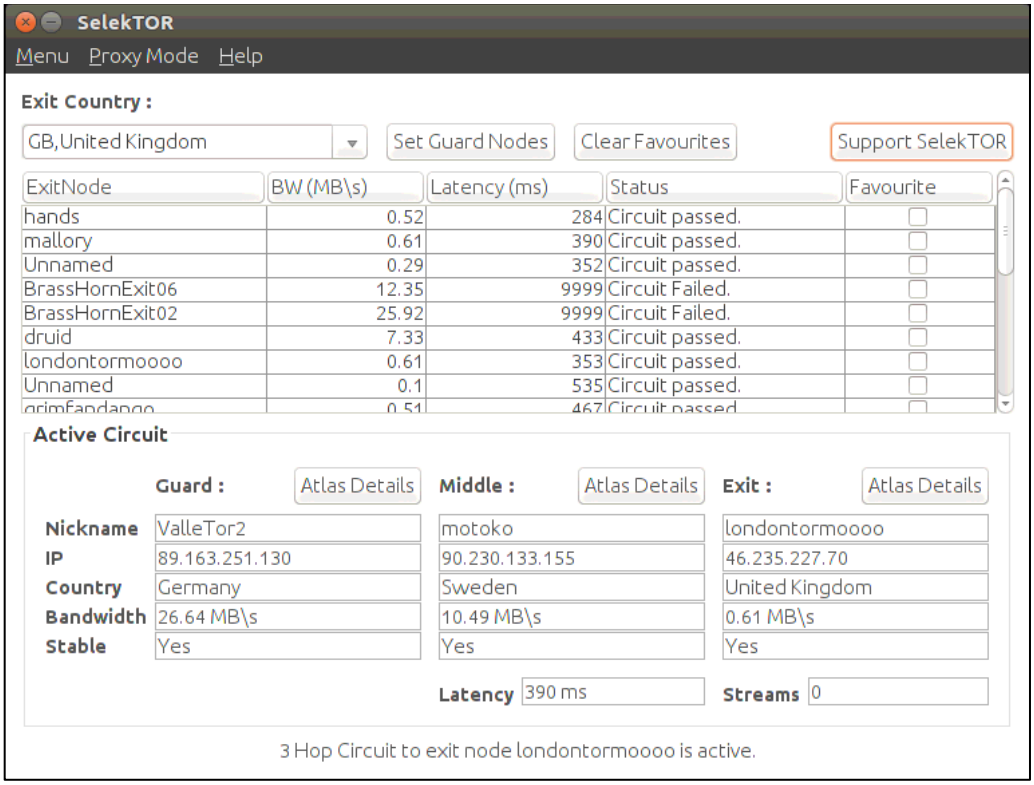


Figura A1.1. Programa SelekTOR

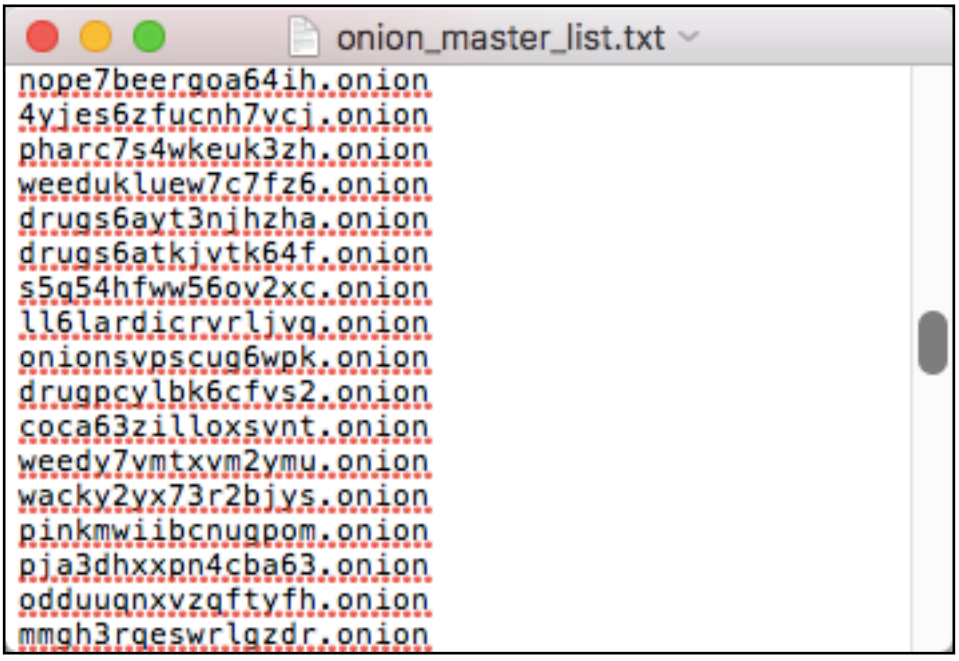


Figura A1.2. Archivo de texto con redes onion resultantes del proyecto onion.

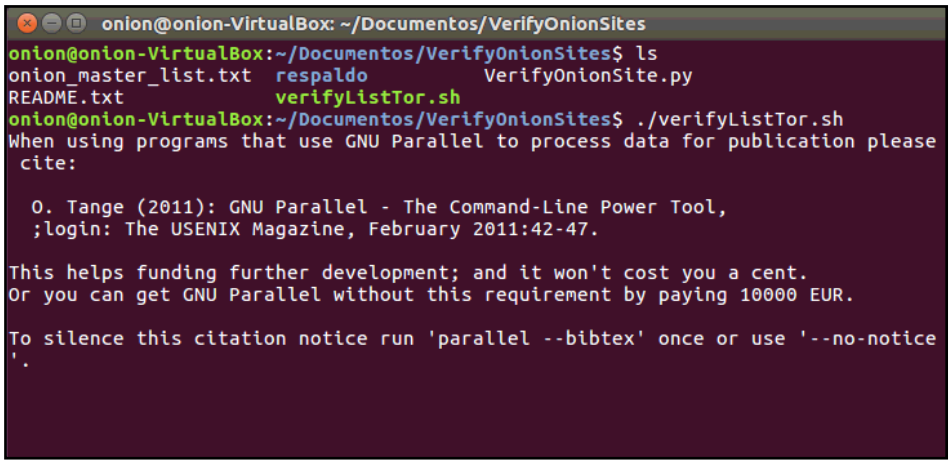


Figura A1.3. Ejecución de Script "*verifyListTor.sh*"

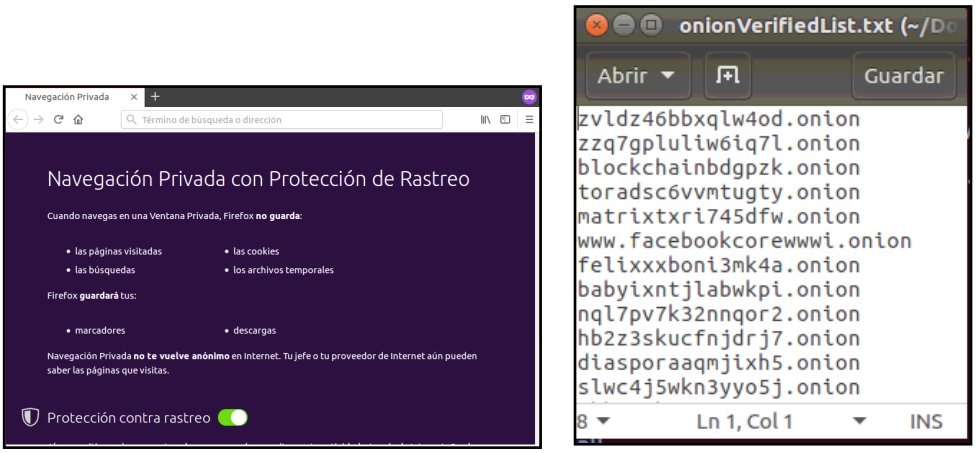


Figura A1.4. Programa *VerifyOnionSite.py* verificando red onion y en caso de estar activa la escribe en "*onionVerifiedList.txt*"

Programa en Python para obtener redes onion en un archivo de texto

```
from stem.control import Controller
from stem import Signal
from threading import Timer
from threading import Event

import codecs
import json
import os
import random
import subprocess
import sys
import time

onions         = []
session_onions = []

identity_lock  = Event()
identity_lock.set()
```

```
    #
    # Grab the list of onions from our master list file.
    #
    def get_onion_list():

            # open the master list
            if os.path.exists("onion_master_list.txt"):
                    with open("onion_master_list.txt","rb") as fd:
                            stored_onions = fd.read().splitlines()
            else:
                    print "[!] No onion master list. Download it!"
                    sys.exit(0)

            print "[*] Total onions for scanning: %d" % len(stored_onions)
            return stored_onions
    #
    # Stores an onion in the master list of onions.
    #
    def store_onion(onion):
            print "[++] Storing %s in master list." % onion
            with codecs.open("onion_master_list.txt","ab",encoding="utf8") as fd:
                    fd.write("%s\n" % onion)
            return
    #
    # Runs onion scan as a child process.
    #
    def run_onionscan(onion):
            print "[*] Onionscanning %s" % onion
            # fire up onionscan
            process                 =                 subprocess.Popen(["onionscan","--webport=0","--jsonReport","--
simpleReport=false",onion],stdout=subprocess.PIPE,stderr=subprocess.PIPE)
            # start the timer and let it run 5 minutes
            process_timer = Timer(300,handle_timeout,args=[process,onion])
            process_timer.start()

            # wait for the onion scan results
            stdout = process.communicate()[0]

            # we have received valid results so we can kill the timer
            if process_timer.is_alive():
                    process_timer.cancel()
                    return stdout

            print "[!!!] Process timed out!"

            return None

    #
    # Handle a timeout from the onionscan process.
    #
    def handle_timeout(process,onion):

            global session_onions
            global identity_lock

            # halt the main thread while we grab a new identity
            identity_lock.clear()

            # kill the onionscan process
            try:
                    process.kill()
                    print "[!!!] Killed the onionscan process."
            except:
                    pass

            # Now we switch TOR identities to make sure we have a good connection
            with Controller.from_port(port=9051) as torcontrol:

                    # authenticate to our local TOR controller
                    torcontrol.authenticate("PythonRocks")

                    # send the signal for a new identity
                    torcontrol.signal(Signal.NEWNYM)
```

```
                # wait for the new identity to be initialized
                time.sleep(torcontrol.get_newnym_wait())

                print "[!!!] Switched TOR identities."

        # push the onion back on to the list
        session_onions.append(onion)
        random.shuffle(session_onions)

        # allow the main thread to resume executing
        identity_lock.set()

        return


#
# Processes the JSON result from onionscan.
#
def process_results(onion,json_response):
        global onions
        global session_onions

        # create our output folder if necessary
        if not os.path.exists("onionscan_results"):
                os.mkdir("onionscan_results")

        # write out the JSON results of the scan
        with open("%s/%s.json" % ("onionscan_results",onion), "wb") as fd:
                fd.write(json_response)

        # look for additional .onion domains to add to our scan list
        scan_result = ur"%s" % json_response.decode("utf8")
        scan_result = json.loads(scan_result)

        if scan_result['identifierReport']['linkedOnions'] is not None:
                add_new_onions(scan_result['identifierReport']['linkedOnions'])

        if scan_result['identifierReport']['relatedOnionDomains'] is not None:
                add_new_onions(scan_result['identifierReport']['relatedOnionDomains'])

        if scan_result['identifierReport']['relatedOnionServices'] is not None:
                add_new_onions(scan_result['identifierReport']['relatedOnionServices'])


        return

#
# Handle new onions.
#
def add_new_onions(new_onion_list):

        global onions
        global session_onions

        for linked_onion in new_onion_list:

                if linked_onion not in onions and linked_onion.endswith(".onion"):

                        print "[++] Discovered new .onion => %s" % linked_onion

                        onions.append(linked_onion)
                        session_onions.append(linked_onion)
                        random.shuffle(session_onions)
                        store_onion(linked_onion)

        return

# get a list of onions to process
onions = get_onion_list()

# randomize the list a bit
random.shuffle(onions)
```

```
session_onions = list(onions)

count = 0

while count < len(onions):

        # if the event is cleared we will halt here
        # otherwise we continue executing
        identity_lock.wait()

        # grab a new onion to scan
        print "[*] Running %d of %d." % (count,len(onions))
        onion  = session_onions.pop()

        # test to see if we have already retrieved results for this onion
        if os.path.exists("onionscan_results/%s.json" % onion):

                print "[!] Already retrieved %s. Skipping." % onion
                count += 1

                continue

        # run the onion scan
        result = run_onionscan(onion)

        # process the results
        if result is not None:

                if len(result):
                        process_results(onion,result)

                        count += 1
```

Programa en Python "*VerifyOnionSite.py"* para verificar la disponibilidad de las redes onion obtenidas en el archivo

I. PRE-REQUISITOS

- Instalar la herramienta Parallel con la siguiente instrucción:
  sudo apt-get install parallel

II. PROGRAMA DE PROCESAMIENTO POR LOTES sh

- Ejecutar el script verifyListTor.sh:
  ./verifyListTor.sh

  Nota: Revisar que tenga permisos de ejecución, en caso de no tenerlo ejecutar la siguiente instrucción:
  chmod +x verifyListTor.sh

  Nota 2: El proceso hace uso de Firefox, por lo que se debe tener instalado el Geckodriver para su funcionamiento.

III. ARCHIVO verifyListTor.sh

```
#!/bin/bash

# Variable con la cual se verifica que se ha llegado al final de la lista de urls .onion
isFileEmpty=1
# Limites inferior y superior para obtener un segmento de la lista de urls. onion
limInf=1
limSup=10

# Variable con la cual se recorren los límites para obtener el siguiente segmento de la lista
step=$limSup

while [ $isFileEmpty -ne 0 ]; do
        # Se obtiene el segmento de acuerdo a los límites y se envían a un archivo en temporal.
   sed -n ' '$limInf','$limSup'p ' onion_master_list.txt > /tmp/asd.txt
        # Se verifica que el archivo temporal no este vacío.
   # En caso de estar vacío terminal el ciclo.
```

```
    isFileEmpty=$(wc -l < /tmp/asd.txt)
    if [ $isFileEmpty -eq 0 ]; then
        break
    fi
        # Se incrementan los límites de acuerdo a la variable step
        # para obtener el siguiente segmento de la lista.
    limInf=$[$limInf+$step]
    limSup=$[$limSup+$step]


        # Con el comando parallel se ejecuta el script VerifyOnionSite.py
        # tantas veces como urls existan en el archivo temporal de forma paralela.
    parallel -a /tmp/asd.txt python3 VerifyOnionSite.py
done
```

IV. Archivo VerifyOnionSite.py

```
import os
import sys
import time
import json
import random
from datetime import datetime

from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support.ui import Select
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.action_chains import ActionChains

from selenium.common.exceptions import TimeoutException

sOnionUrl = sys.argv[ 1 ]

options = webdriver.FirefoxOptions()
options.add_argument( '--private-window' )

profile = webdriver.FirefoxProfile()
profile.set_preference( 'network.proxy.type', 5 )
profile.set_preference( 'network.proxy.socks_remote_dns', True )

try:
    browser = webdriver.Firefox( log_path="/dev/null", firefox_options=options, firefox_profile=profile,  )
except:
    sys.exit( 1 )

browser.minimize_window()
browser.set_page_load_timeout( 30 )
try:
    browser.get( 'http://' + sOnionUrl )
    with open( 'onionVerifiedList.txt', "a" ) as onionVerifiedList:
        onionVerifiedList.write( sOnionUrl + '\n' )
except:
    pass
```

Anexo 2

Se puede crear una configuración avanzada para realizar investigación de forma segura, de tal forma que se guarden las páginas onion seleccionadas, así como sus fotografías e información adicional. Para ello se utilizarán dos máquinas virtuales Linux para proporcionar una forma de acceder a la red Tor con Hunchly. Tendrá la máquina virtual Whonix Gateway que pasará TODO el tráfico a la red Tor y una máquina de investigación donde tendrá instalado Chrome y donde realizará su trabajo de investigación, esto proporcionará un nivel mucho más alto de seguridad y anonimato.

Para llevar a cabo este procedimiento se debe seguir el siguiente procedimiento:

I. Descargar e instalar Virtual Box.
II. Descargar Mint Linux ISO y ejecutarlo como máquina virtual en Virtual Box (Figura A2.1).
III. Dentro del sistema operativo de Linux Mind 18 “Sarah”, instalar el Google Chrome (Figura A2.2).
IV. Descargar e instalar Hunchly (Figura A2.3).
V. Descargar y ejecutar Whonix Gateway Virtualbox Image en Virtual Box (Figura A2.4).
VI. Minimizar Whonix Gateway Virtualbox y abrir la máquina virtual de Investigator Hunchly y modificar en Configuración, Red (IP Address: 10.152.152.11; Netmask: 255.255.192.0; Gateway:10.152.152.10; DNS Server: 10.152.152.10) (Figura A2.5).
VII. Conectarse a la Red interna con el nombre de Whonix, posteriormente abrir el Google Chrome con la URL de https://check.torproject.org (Figura A2.6).
VIII. En la parte superior derecha del navegador tenemos el icono de Hunchly, al cual al seleccionarlo podemos empezar a capturar

la(s) páginas que deseemos, las cuales se guardaran y administrarán en el Dashboard, también podemos guardar selectores, notas y etiquetas, al presionar el botón de Dashboard vemos un administrador de investigaciones (Figura A2.7), en el cual podemos capturar las páginas onion vistas, así como fotos, archivos adjuntos, e información de la página (Figura A2.8).

IX. Por último, podemos generar reportes y exportar la información capturada (Figura A2.9).

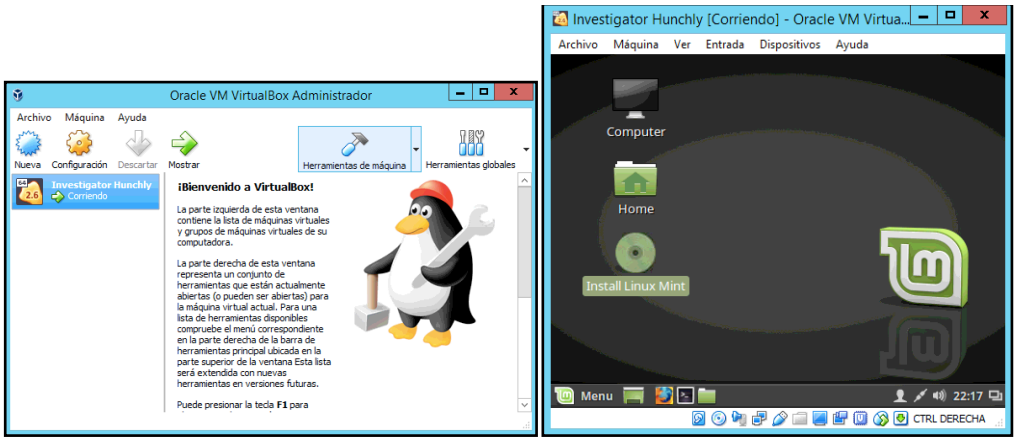


Figura A2.1. Ejecución de Mint Linux en VirtualBox.

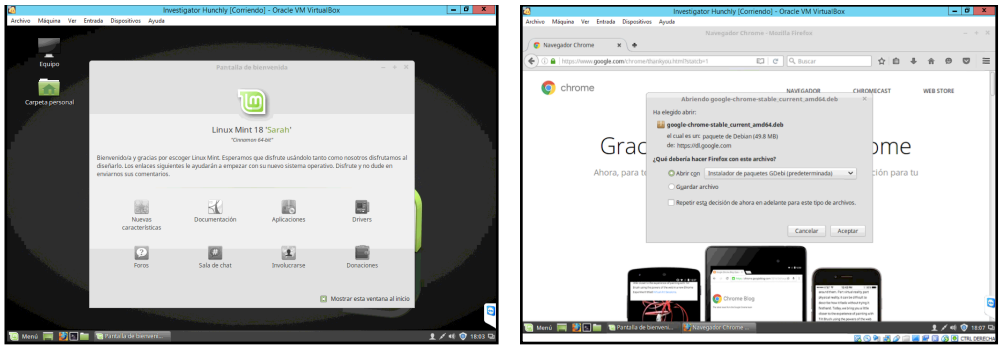


Figura A2.2..Sistema Operativo Linux Mint 18; Instalación de Google Chrome.

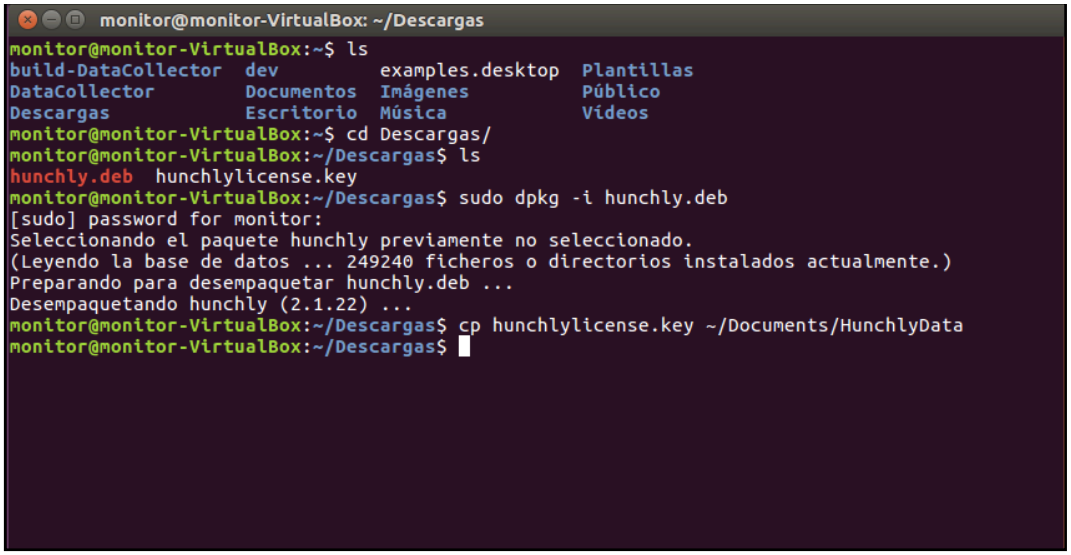


Figura A2.3. Instalación de Hunchly

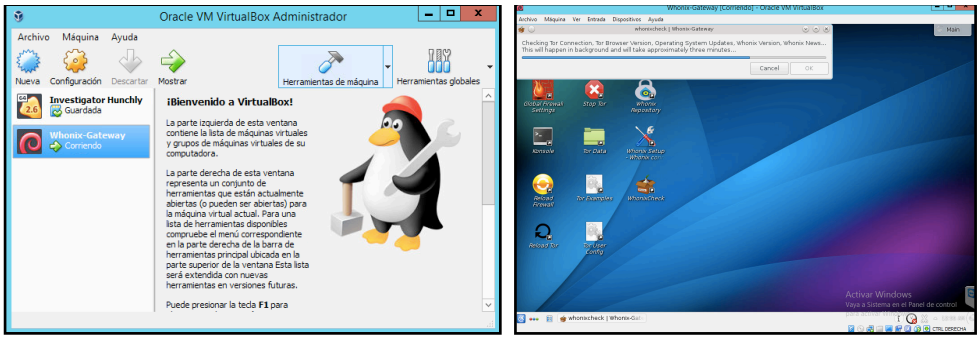


Figura A2.4. Ejecución de Whonix Gateway Virtualbox Image en Virtual Box.

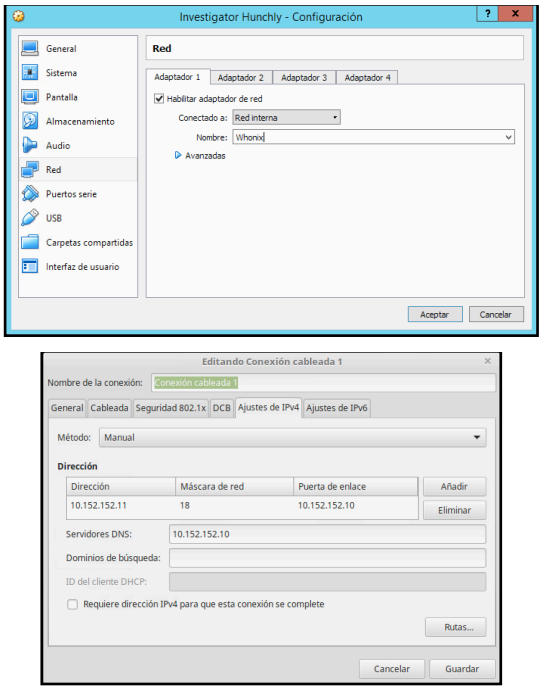


Figura A2..5. Configuración de red en Investigator Hunchly en Virtual Box.

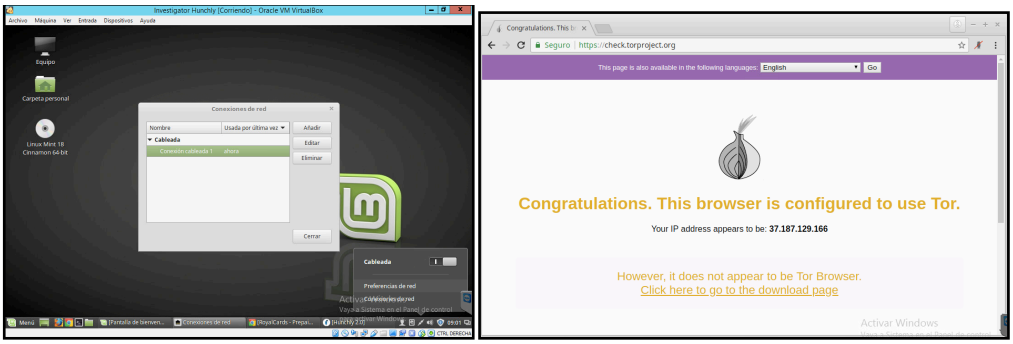


Figura A2.6. Conexión a red interna desde Investigation Hunchly y abrir https://check.torproject.org en Google Chrome.

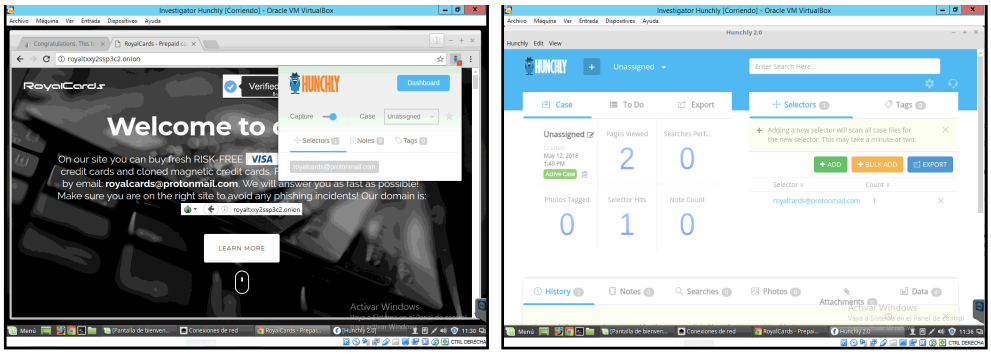


Figura A2.7.Dashboard de Hunchly.

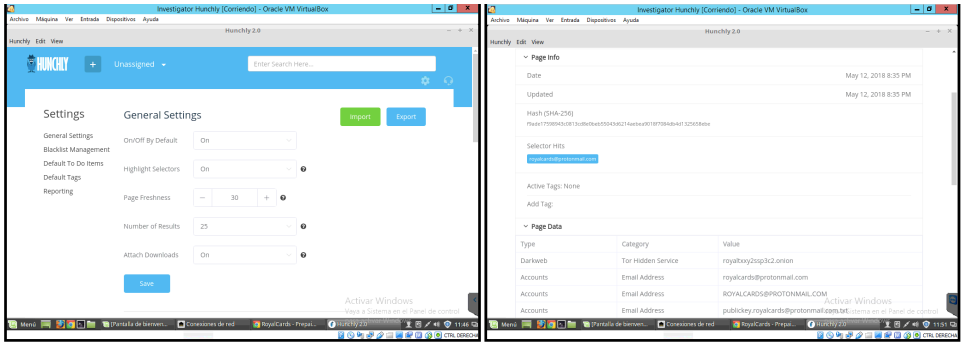


Figura A2.8. Activación de características e Información de la página onion.

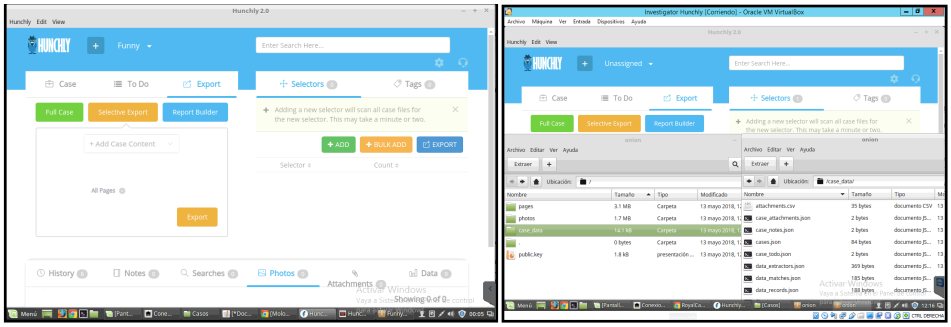


Figura A2.9. Reportes y Datos exportados.

Anexo 3

Este anexo tiene la intención de listar muestras de las redes onion por conductas delictivas que se tomaron en cuenta para esta investigación.

Contra la Salud

Red onion	Descripción
armoryohajjhou5m.onion	Armory - Buy weapons for bitcoin
2kka4f23pcxgqkpv.onion	Euro Guns - Number one guns dealer in onionland - Buy guns and ammo for Bitcoin
yourdyh7n6z46uog.onion	Your Drug
elherfvmliyagg7b.onion	ElHerbolario
weaponstrxqniqrt.onion	Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since 2001
pharmacpr5lpfin5.onion	BitPharma - biggest european .onion drug store - Cocaine for Bitcoins, Psychedelics for Bitcoins, Prescriptions for Bitcoins, Viagra
drugszun7tvsgsaa.onion	Peoples Drug Store - The Darkweb's Best Online Drug Supplier! - Buy cocaine, speed, xtc, mdma, heroin and more at peoples drug store, pay with Bitcoin
mollyworup44gri7.onion	Mollyworld No.1 Provider of Crystals and Pills
cjakglmv3vidqwgt.onion	US Pharmacy
elevateeo6usyjlq.onion	Elevate

Contra la Libertad de Tránsito:

Red onion	Descripción
redrooocf3it3g3d.onion	Red Room
rape2izpkoc6lfld.onion	Real Rape
gewaltics7teim6i.onion	Gefilmte Vergewaltigung - Echte Rape Sex Porno Videos.
timf7jxjoflkybdd.onion	Bondage Porn Sites Rip's
pu23yyt5u5p3xd23.onion	Pure Porn - The Best Hacked Porn Accounts
amputefruj4rzgz5.onion	Amputees Porn
gjgbztmown7d6usl.onion	Rape and murder! Shock photo and video! Killing people.
vaultjdmqoxebbav.onion	Vault of Sex and Dead

Contra la Libertad Sexual:

Red onion	Descripción
ditxyqfbmhilaloi.onion	Photos LS Models pedos Child porn
wannarcgxjn5hqc4.onion	Wanna Fuck me - Teen Girls Teen Sex Teen Rape
cpstoreizehdhenw.onion	CP Store. Buy underage sex child porn kids preteen jailbait pthc ptsc content for bitcoins
itu5h4f7shmamz2x.onion	CP videos and photos Child porn Children pedo kids
animalirgsuecrvn.onion	Animal Porno FREE FREE
qa4t6wjhl4gzl5in.onion	Teen Deepthroat | Amateur porn private collection of Young Girls!
youngd4apm5ua5r2.onion	The Youngest Girls - Check out tons of Porn Videos with Super Young Girls
222222avkcjpcbwi.onion	18 X Girls - 18 Year Old Girls - Adult Videos and XXX Porn Pictures - Teen Sex ::
destroplh4zwowdk.onion	Destroyed Daughters - Private Extreme Teen Porn Video Community
alicedbdh5xixwai.onion	AliceDB - Child models photo collection
jd6jgx744yyhfwki.onion	Dear
Lolimknaduomuzdr.onion	Loli Lust

Contra la Seguridad Nacional:

Red onion	Descripción
6dvj6v5imhny3anf.onion	CyberGuerrilla leAkboX
lu4qfnnkbnduxurt.onion	CyberGuerrilla Autonomous Nexus
264nglqbtqlabsxl.onion	CyberGuerrilla soApboX
qf7bzc2hcmmooqnb.onion	Namaste - CyberGuerrilla Nexus - BroadCast Station
5slxqzbtjz5uu4pt.onion	Internetional CyberGuerrillA Column

Contra la Seguridad Financiera:

Red onion	Descripción
creditlwkdnptwla.onion	Venta de tarjetas de crédito y de transferencias de dinero
prvtzone7mq377pw.onion	Foro relacionados con tarjetas de crédito
r4u6jtmqzuedlgle.onion	Clonación de tarjetas de crédito
buyccoq36hlj6etg.onion	Venta de tarjetas de crédito y cuentas bancarias
ccxdnuwmk25iqtas.onion	Clonación de tarjetas de crédito

utvjqkyc4ejhzkwu.onion	Falsificación de cuentas de banco en Europa
777o6suetmexlesv.onion	Compra de tarjetas de crédito
moneyrnr22vgcil6.onion	Trabajan con tarjetas de crédito pre pagadas
bnwcards4xuwihpj.onion	Tarjetas clonadas en Euros y Dólares
u4oh5loaqrl3gotd.onion	Clonación de cuentas de PayPal
l6quosmt2ffwphvf.onion	Venta de datos de tarjetas de crédito y debido

Contra las Tecnologías de la Información:

Red onion	Descripción
www.2ogmrlfzdthnwkez.onion	Ataque de denegación de servicio (DDOS)
hackeroql4l2mejs.onion	Rent-A-Hacker - Hire a hacker for every job you can imagine, from DDOS to completely ruining people or destroy reputation of a company or individual
paxhumana5oopssw.onion	Pax Humana
ctworld7doy2422v.onion	CTW Forums – Powered by XMB
rhe4faeuhjs4ldc5.onion	ANONYMOUS'z FORUM
lu4qfnnkbnduxurt.onion	CyberGuerrilla Autonomous Nexus

En la siguiente tabla se lista el comportamiento por fecha de páginas nuevas, desactivadas y activas de la red Tor dentro de la muestra de Hunchly

Fecha	Nuevas	Inactivas	Activas
02/06/19	3	1006	4513
01/06/19	1	1181	4584
31/05/19	0	1043	4569
30/05/19	1	981	4636
29/05/19	0	968	4611
28/05/19	1	953	4668
27/05/19	1	961	4566
26/05/19	0	958	4577

Fecha	Nuevas	Inactivas	Activas
25/05/19	0	1023	4528
24/05/19	2	975	4576
23/05/19	2	996	4537
22/05/19	3	1092	4582

Anexo 4

Este anexo tiene la intención de mostrar el proceso de instalación y configuración para usar la herramienta wget, con el objetivo de recuperar sitios onion en linux.

Paso 1: Instalar privoxy

root@tor18:/home/jsalas# **sudo apt-get install -y tor tor-geoipdb privoxy**

Paso 2: Detener privoxy y editar el archivo de configuración de privoxy

root@tor18:/home/jsalas# **sudo systemctl stop privacy**
root@tor18:/etc# **cd /etc/privoxy/**
root@tor18:/etc/privoxy# **nano config**
agregar:
listen-address localhost:8118
forward-socks5 / 127.0.0.1:9054 .

Paso 3: Editar el archivo de configuración de wgetrc y agregar http://localhost:8118 en http_proxy=

root@tor18:/etc# **cd /etc**
root@tor18:/etc# **nano wgetrc**

Tune HTTPS security (auto, SSLv2, SSLv3, TLSv1, PFS)
#secureprotocol = auto
http_proxy = http://localhost:8118

Paso 4: Instalar Tor y detenerlo

root@tor18:/etc# sudo apt-get install tor

root@tor18:/etc# sudo nano /etc/default/tor
root@tor18:/etc# sudo service tor stop

Paso 5: Descargar e instalar Selektor

root@tor18:/home/jsalas/Descargas# cd /home/jsalas/Descargas
root@tor18:/home/jsalas/Descargas# tar -zxvf selektor-3.13.73_all.tar.gz
root@tor18:/home/jsalas/Descargas# cd selektor-3.13.73_all
root@tor18:/home/jsalas/Descargas/selektor-3.13.73_all# sudo ./install.sh

Paso 6: Iniciar selector desde la interface gráfica

Paso 7: Iniciar privoxy

root@tor18:/home/jsalas/crawl1# **sudo systemctl start privoxy**

¡Listo!

Una de las fuentes usadas fue:
http://www.dejonck.be/2013/05/data-mining-using-wget-with-tor-for.html?m=1

Anexo 5

Archivo cmd1.txt que contiene las características de las cabeceras de los archivos html recuperados y es llamado por awk.

```
BEGIN {
OFS = ";"
}
{
if ( $1 ==  "HTTP/1.1" )
{
for (i = 2; i <= 3; i++)
http = http " " $i;
}
if ( $1 == "Date:" )
{
for (i = 2; i <= 7; i++)
fecha = fecha " " $i;
}
if ( $1 == "Server:" )
{
server = $2;
}
if ( $1 == "X-Content-Type-Options:" )
{
content = $2;
}
if ( $1 == "X-Frame-Options:" )
{
frame = $2;
}
if ( $1 == "X-Xss-Protection:" )
{
xss = $2;
}
if ( $1 == "Referrer-Policy:" )
{
referrer = $2;
}
if ( $1 == "Last-Modified:" )
{
for (i = 2; i <= 7; i++)
last = last  " " $i;
}
if ( $1 == "ETag:" )
{
```

```
etag = $2;
}
if ( $1 == "Accept-Ranges:" )
{
ranges = $2;
}
if ( $1 == "Content-Length:" )
{
l = $2;
}
if ( $1 == "Vary:" )
{
vary = $2;
}
if ( $1 == "Connection:" )
{
conn = $2;
}
if ( $1 == "Content-Type:" )
{
ConType = $2;
}
if ( $1 == "Proxy-Connection:" )
{
Proxy = $2;
}
}
END {
print http, fecha, server, content, frame, xss, referrer, last, etag, ranges, l, vary, conn, ConType, Proxy
}
```

Anexo 6

El programa CalculateSimilarity.R calcula la similitud de Jackard y Coseno, como podemos ver en la figura 4.8. Posteriormente ejecutaremos el programa desde RStudio:

> source(“~/R/CalculateSimilarity.R")

```
library(text2vec)
library(data.table)
library(stringr)
good <- data.frame(id = character(), content = character(), stringsAsFactors = FALSE)

pathgood <- "~/R/good/"
filesgood <- list.files(pathgood, pattern=".+txt")

for (i in 1:length(filesgood)){
  filename <- paste0(pathgood,filesgood[i])
  print(paste("Processing", i, filesgood[i]))
  content <- readChar(filename, file.info(filename)$size)
  good <- rbind(good, data.frame(id = filesgood[i], content = content))
}

bad <- data.frame(id = character(), content = character(), stringsAsFactors = FALSE)

pathbad <- "~/R/bad/"
filesbad <- list.files(pathbad, pattern=".+txt")

for (i in 1:length(filesbad)){
  filename <- paste0(pathbad,filesbad[i])
  print(paste("Processing", i, filesbad[i]))
  content <- readChar(filename, file.info(filename)$size)
  bad <- rbind(bad, data.frame(id = filesbad[i], content = content))
}

prep_fun = function(x) {
  str_replace_all(str_replace_all(str_to_lower(x),"[^[:alnum:]]", " "), "\\s+", " ")
}

good$content_clean = prep_fun(good$content)
bad$content_clean = prep_fun(bad$content)

itgood = itoken(good$content_clean, progressbar = FALSE)
vgood = create_vocabulary(itgood)
itbad = itoken(bad$content_clean, progressbar = FALSE)
vbad = create_vocabulary(itbad)

it = itoken(bad$content_clean, progressbar = FALSE)
v = create_vocabulary(it) #%>% prune_vocabulary(doc_proportion_max = 0.1, term_count_min = 5)
vectorizer = vocab_vectorizer(v)


dtmgood = create_dtm(itgood, vectorizer)
dim(dtmgood)

dtmbad = create_dtm(itbad, vectorizer)
dim(dtmbad)
```

```
# Jaccard similarity
jac_sim = sim2(dtmgood, dtmbad, method = "jaccard", norm = "none")
dim(jac_sim)
jac_sim[]

# Cosine similarity
cos_sim = sim2(dtmgood, dtmbad, method = "cosine", norm = "l2")
dim(cos_sim)
cos_sim[]
```

Figura 4.8. . Programa CalculateSimilarity.R.

Bibliografía

1. Aditya K Sood, Sherali Zeadally, and Rohit Bansal. (14 de July de 2017). Cybercrime at a Scale: A Practical Study of Deployments of HTTP-Based Botnet Command and Control Panels. IEEE.
2. Ahmed T. Zulkarnine, Richard Frank, Bryan Monk, Julianna Mitchell, Garth Davies (17 de November de 2016). Surfacing collaborated networks in dark web to find illicit and criminal content. IEEE.
3. Altayar, M. S. (24 de April de 2017). A comparative study of anti-cybercrime laws in the Gulf Cooperation Council countries. IEEE.
4. Andres Baravalle, M. S. (15 de December de 2016). Mining the Dark Web: Drugs and Fake Ids. IEEE.
5. Andrew J. Park, Brian Beck, Darrick Fletche, Patrick Lam, and Herbert H. Tsang. (24 de December de 2016). Temporal analysis of radical dark web forum users. IEEE.
6. Arredondo, N. P. A. (2009). *Método semisupervisado para la clasificación automática de textos de opinión* (Doctoral dissertation, Instituto Nacional de Astrofísica, Óptica y Electrónica).
7. Balduzzi M., C. V. (s.f.). Cybercrime in the Deep Web. Black Hat EU, Amsterdam 2015. Trend Micro.
8. Berghel, H. (2017). Which Is More Dangerous the Dark Web or the Deep State? IEEE.
9. BrightPlanet. (March de 2013). UNDERSTANDING THE DEEP WEB IN 10 MINUTES. Obtenido de http://cdn2.hubspot.net/hub/179268/file-377288418-pdf/docs/understandingthedeepweb_20130311.pdf?t=1493920037822
10. Cámara de Diputados, C. (2012). Constitución Política de los Estados Unidos Mexicanos. . México.
11. Cassou, R. J. (2009). Delitos informáticos en México. Revista Número 28 del Instituto de la Judicatura Federal. http://www.ijf.cjf.gob.mx/publicaciones/revista/28/Delitos_inform%C3%A1ticos.pdf.
12. Chakrabarti, S. V. (1999). Focused Crawling: A New Approach to Topic-Specifi c Resource Discovery. In Proceedings of the Eighth World Wide Web Conference. Toronto, Canada.
13. Chen, H. (2011). Dark web: Exploring and data mining the dark side of the web (Vol. 30). Springer Science & Business Media.
14. Chen, H. a. (2003). Comparison of Three Vertical Search Spiders. ((. 2.–3. Annual Review of Information Science and Technology, Ed.)
15. Cheong, F. C. (1996). Internet Agents: Spiders, Wanderers, Brokers, and Bots. Indianapolis, IN: New Riders Publishing.
16. Christos Iliou, George Kalpakis, Theodora Tsikrika, Stefanos Vrochidis and Ioannis Kompatsiaris. (4 July de 2017). Hybrid focused crawling on the Surface and the Dark Web. EURASIP Journal on Information Security.
17. Comisión Nacional de Seguridad. Atribuciones de la División Científica. (2017). Obtenido de http://cns.gob.mx/portalWebApp/wlp.c?__c=fdd

18. Conti, M., Crane, S., Frassetto, T., Homescu, A., Koppen, G., Larsen, P., ... & Sadeghi, A. R. (2016). Selfrando: Securing the tor browser against de-anonymization exploits. Proceedings on Privacy Enhancing Technologies 2016(4), 454-469.
19. de Organizaciones, L. G. (s.f.). Ley Federal de Instituciones de Fianzas. Ley para Regular las Agrupaciones Financieras.
20. Denic, M. N. (2017). GOVERNMENT ACTIVITIES TO DETECT, DETER AND DISRUPT THREATS ENUMERATING FROM THE DARK WEB. Fort Leavenworth, Kansas.
21. Federal, D. (2001). Código Penal Federal. (E. L. Lozano., Ed.)
22. Florescu, D. L. (1998). Database Techniques for the World-Wide Web: A Survey. (2. 5.–7. SIGMOD Record, Ed.)
23. Gao, X., & Wu, S. (2018, August). Hierarchical Clustering Algorithm for Binary Data Based on Cosine Similarity. In 2018 8th International Conference on Logistics, Informatics and Service Sciences (LISS) (pp. 1-6). IEEE.
24. Ghappour, A. (April de 2017). Searching Places Unknown: Law Enforcement Jurisdiction on the Dark Web. SSRN.
25. Graeber, C. (October de 2016). The Man Who Lit The Dark Web. Popular Science.
26. Helena Piccinini, Marco A. Casanova, Luiz André P. P. Leme, Antonio L. Furtado . (2014). Publishing deep web geographic data. Springer.
27. Hunchly. (2018). Hunchly. Obtenido de https://www.hunch.ly/
28. Hurlburt, G. (26 de April de 2017). Shining Light on the Dark Web. IEEE, 50.
29. Investigation, F. B. (s.f.). Internet Crime Complaint Center. Obtenido de https://www.ic3.gov/preventiontips.aspx
30. KhalidAl-RowailyaMuhammadAbulaishbNurAl-Hasan HaldarcMajedAl-Rubaian . (14 de September de 2015). BiSAL – A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security. ELSEIVER.
31. Kirkpatrick, K. (March de 2017). Financing the Dark Web. ACM, 60.
32. Krawezik, G. P., Kogge, P. M., Dysart, T. J., Kuntz, S. K., & McMahon, J. O. (2018, September). Implementing the Jaccard Index on the Migratory Memory-Side Processing Emu Architecture. In 2018 IEEE High Performance extreme Computing Conference (HPEC) (pp. 1-6). IEEE.
33. Lewis, S. J. (s.f.). Discovering the Dark Web. Obtenido de OnionScan: https://onionscan.org/
34. Martijn Spitters, Femke Klaver, Gijs Koot, Mark van Staalduinen. (9 de September de 2015). Authorship Analysis on Dark Marketplace Forums. IEEE.
35. Matti Nasi, Atte Oksanenb, Teo Keipia and Pekka Rasanen . (6 de July de 2015). Cybercrime victimization among young people: a multi-nation study. Journal of Scandinavian Studies in Criminology and Crime Prevention .
36. Mittal, S. (30 de August de 2017). Enough Law of Horses and Elephants Debated...Let's Discuss the Cyber Law Seriously. SSRN.
37. Pant, G. S. (2002). Exploration versus Exploitation in Topic Driven Crawlers. In Proceedings of the WWW Workshop on Web Dynamics.
38. Rafiuddin, M. F. B., Minhas, H., & Dhubb, P. S. (2017, September). A dark web story in-depth research and study conducted on the dark web based on forensic com-

puting and security in Malaysia. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* (pp. 3049-3055). IEEE.

39. Raghavan, S. a.-M. (2001). Crawling the Hidden Web. (I. P. Databases., Ed.)
40. Rechtman, Y. (June de 2017). Shifting the Risk of Cybercrime. The CPA Journal.
41. Regner Sabillon, Jeimy Cano, Victor Cavaller, Jordi Serra-Ruiz. (17 de November de 2016). Cybercriminals, cyberattacks and cybercrime. IEEE.
42. Report on Cyber Crime Investigation. (2011). Obtenido de http://www.htcia.org/wp-content/uploads/2011survey_report.pdf
43. Senado de la República. (s.f.). Obtenido de http://www.senado.gob.mx/sgsp/gaceta/63/1/2015-10-27-1/assets/documentos/Inic_PRI_Ley_Delitos_Informaticos.pdf
44. Somayyeh Aghababaei, Masoud Makrehchi . (16 de October de 2016). Mining Social Media Content for Crime Prediction. IEEE.
45. Sonali Gupta, Komal Kumar Bhatia. (June de 2014). A Comparative Study of Hidden Web Crawlers. International Journal of Computer Trends and Technology (IJCTT).
46. Wadhwa, A., & Arora, N. (June de 2017). A Review on Cyber Crime: Major Threats and Solutions. International Journal of Advanced Research in Computer Science.
47. Weimann, G. (2016). Going Dark: Terrorism on the Dark Web. Routledge.
48. WIECZNER, J. (22 de August de 2017). THE 21ST-CENTURY BANK ROBBERY. FORTUNE.
49. Xuefeng Xian, Pengpeng Zhao, Victor S. Sheng, Ligang Fang, Caidong Gu, Yuanfeng Yang, and Zhiming Cui . (6 de April de 2016). Stratification-Based Outlier Detection over the Deep Web. Computational Intelligence and Neuroscience.
50. Yan Wang, J. L. (2016). Crawling ranked deep Web data sources. Springer Science+ Business Media New York.
51. Yan Wang, J. L. (3 de September de 2016). Crawling ranked deep Web data sources. SPRINGER.
52. Zhou Li, Sumayah Alrwais Yinglian Xie, Fang Yu, XiaoFeng Wang. (25 de June de 2013). Finding the Linchpins of the Dark Web: a Study on Topologically Dedicated Hosts on Malicious Web Infrastructures. IEEE.