

**91.25 Ciencia de Datos para la Toma de Decisiones**

**Trabajo de Aplicación**

**¨Modelo Predictivo para casos de Phishing¨**

Alumnos:

* Juan Cruz Camacho ()
* Tobías Pucci Romero ()
* Lucas Ken Kiriyama ()
* Brian Emanuel Slavkin (102820)

Cuatrimestre: 1º 2021

Docentes:

* Dr. Ing. Emilio Picasso
* Mag. Ing. Mariano Bonoli Escobar
* Mag. Ing. Xavier González

Resumen Ejecutivo

Indice

Introducción

Para comenzar la búsqueda de nuestro medido predictivo, debemos comenzar entendiendo que es el phishing y cual es la importancia de lograr evitar estas prácticas en el ámbito cibernético.

La palabra phishing quiere decir suplantación de identidad, se utiliza para denominar al conjunto de técnicas de ingeniería social que utilizan los ciberdelincuentes para obtener información confidencial o personal de otros usuarios de la red de manera fraudulenta y así tomar la identidad del estafado.

La practica mas frecuente de estas técnicas se realiza mediante correos electrónicos falsos los cuales aparentan no serlo, como podría ser un servicio de banco, subscripciones de paginas reconocidas por el individuo a estafar, redes sociales, etc. Estos emails derivan en enlaces o links los cuales también aparentan ser reales. Le exigen al usuario sus datos como pueden ser nombre de usuario, contraseñas o cuentas bancarias, mediante encuestas las cuales aparentan ser de los remitentes reales, en los cuales por lo general el estafado confía y así se concreta la obtención de datos del fraude. Otra técnica muy utilizada es mediante archivos adjuntos o botones los cuales generan una descarga la cual puede ser un virus o software del cual el remitente obtiene la información buscada.

Estos enlaces van a ser nuestro principal objeto de estudio, ya que no podemos evitar que el delincuente envíe los correos electrónicos ya que al identificarlos y bloquearlos siempre esta la posibilidad de crear uno nuevo.

Es importante también remarcar que los enlaces pueden llegar no solo por mail, sino que hoy en día estas técnicas mutaron otras vías, como mensajes de texto o redes sociales como WhatsApp, Facebook, Instagram, etc.

En el año 2011 en la Argentina se presento un proyecto de ley en el Senado de la nación en el cual se define el concepto:

*“La presente Ley tiene como objeto tipificar el delito de obtención ilegitima de datos confidenciales, conocido como “phishing” el cual se define como la capacidad de duplicar una página Web para hacer creer al visitante que se encuentra en el sitio Web original, en lugar del falso. Normalmente, se utiliza con fines delictivos enviando SPAM e invitando acceder a la página señuelo. El objetivo del engaño es adquirir información confidencial del usuario como contraseñas, tarjetas de crédito o datos financieros y bancarios.”[[1]](#footnote-1)*

<https://www.argentina.gob.ar/justicia/convosenlaweb/situaciones/phishing>

<https://phishtank.org/what_is_phishing.php>

<https://es.wikipedia.org/wiki/Phishing>

Los enlaces a los que nos estamos refiriendo son direcciones web, también denominada URL (Uniform Resource Locator). Estos son una serie de caracteres, número y símbolos que ayudan a localizar un recurso único en la Web, los cuales pueden ser paginas HTML, documentos CSS, imágenes, etc.

La composición de un URL es un concepto clave para el análisis de nuestro trabajo. Esto se puede observar en la siguiente imagen:



El protocolo (Scheme en inglés) es el primer componente que siempre se encuentra a la izquierda, la que indica como el navegador debe acceder al recurso al cual queremos llegar. Es una cuestión técnica. Los utilizados en las paginas web son http o https, y la diferencia entre estos es que el https impide que otros usuarios puedan interceptar la información confidencial que se transfiere entre el usuario y el servidor.

Luego viene el subdominio y el dominio que indican el servidor web que se solicita. El dominio es el nombre del sitio y el subdominio es una ramificación del dominio que sirve para crear distintas páginas. La extensión es la primera división que se realizó en internet, las cuales se comenzaron creando genéricas como *com*, *net*, *org* y una extensión para cada país del mundo. Como quedan pocos dominios libres en las extensiones fueron surgiendo con el tiempo nuevas.

Luego tenemos componentes optativos como lo pueden ser la carpeta y la subcarpeta también conocidos como el directorio el cual muestra al servidor el camino hacia donde esta el recurso. En caso de buscar dentro del dominio específicamente una página, imagen o video el URL incorpora el componente recurso.

Por último tenemos los parámetros, los cuales también son opcionales, y aparecen al final del URL separados mediante el ¨? ¨. Estos sirven a los programadores para incluir información adicional que utilizan las paginas web para enviar información a los servidores. Consiste, como se puede ver en el ejemplo de una clave y un valor (clave=valor).

<https://superadmin.es/blog/que-es/direccion-web-url/>

<https://developer.mozilla.org/en-US/docs/Learn/Common_questions/What_is_a_URL>

<https://es.godaddy.com/blog/diferencia-entre-http-y-https/>

Objetivo del análisis

Buscamos obtener un modelo que generalice bien antes nuevas observaciones, permitiendo detectar si una página tiene contenido legítimo o si este es un método de phishing web usando la menor información posible detectar.

Este servicio podría usarse como plugin (complemento) en buscadores web y en servicios de email o mensajería para prevenir a los usuarios el acceso a páginas potencialmente peligrosas. Una posibilidad es informar la probabilidad de que dicha página pueda tratarse de una actividad ilegítima a través de una predicción discreta.

Preparación del Data set

Para poder generar un modelo predictivo representativo, cuyas muestras representen tanto casos de phishing como paginas reales. Para esto, primero, primero recurrimos a la página <https://phishtank.org/> la cual se encarga de analizar URLs sospechosos proporcionados por los usuarios, en búsqueda de en caso de que sean fraudulentas sean inactivadas de la web. Esta pagina nos proporciona de manera gratuita una base de datos con 11.000 páginas webs inactivas que fueron verificadas como herramientas de phishing en formato CSV. De este archivo obtuvimos el URL y únicamente la variable Dummy de si es realmente phishing o no.

Para poder ampliar esta pequeña cantidad de muestras, utilizamos la herramienta de Web Scraping, en Python con la librería “Scrapy”, para ingresar a PhishTank y obtener muestras tanto “valid” (phishing) e “invalid” (no phishing). El método de scraping consistió en ingresar en la página <https://phishtank.org/phish_archive.php> la cual es un repositorio de URLs con la información que buscamos. Primero, se filtró la lista por valid y luego por invalid. Dependiendo del largo del URL tuvimos que crear un “re-scraping” en los casos en que no se podían visualizar de manera correcta en la tabla e ingresar en la página de la observación. Se obtuvieron alrededor de 60.000 observaciones “valid” y 76.000 “invalid”. Por lo que se generó un dataset de 147.000 muestras.

SE FILTRA O NO SE FILTRA

A partir de esto, se generaron las siguientes variables en función del URL total:

* URL completo (*url*)
* Phishing (variable respuesta, =0 si no es/=1 si es)
* Scheme (Protocolo del URL)
* Dominio completo (*domain\_complete*)
* Dominio del URL
* Subdominio del URL
* Sufijo del URL (*suffix*)
* Dominio y Subdominio (*domain\_subdomain*)
* Cantidad de puntos en el dominio (*dom\_n\_puntos*)
* Cantidad de guiones en el dominio (*dom\_n\_guion*)
* Cantidad de guiones bajos en el dominio (*dom\_n\_guionbajo*)
* Longitud del dominio completo (*dom\_len\_tot*)
* Longitud del dominio (*dom\_len*)
* Longitud del dominio (*dom\_len\_sub*)
* Cantidad de vocales en el dominio (*dom\_vocales*)
* Cantidad de consonantes en el dominio (*dom\_cons*)
* Cantidad de números en el dominio (*dom\_num*)
* Cantidad de caracteres diferentes en el URL (*dom\_car\_dif*)
* Dominio es un IP (*dom\_ip*): Variable Dummy si el dominio es un IP o no.
* Protocolo es http (*sch\_http*): Variable Dummy si el protocolo es http, en caso de no serlo es https.
* Longitud del sufijo (*suf\_len*)
* Sufijo .com (*suf\_com*): Variable Dummy si el sufijo es .com.
* Sufijo .net (*suf\_net*): Variable Dummy si el sufijo es .net.
* Sufijo .org (*suf\_org*): Variable Dummy si el sufijo es .org.
* Sufijo .other (*suf\_other*): Variable Dummy si el sufijo es .other.
* Sufijo .ru (*suf\_ru*): Variable Dummy si el sufijo es .ru.
* Sufijo .xyz (*suf\_xyz*): Variable Dummy si el sufijo es .xyz.

A partir de estas variables, observamos que muchas observaciones de phishing tenían dominios similares a paginas mundialmente reconocidas, y de las mas utilizadas. Esto era algo lógico, ya que al ser paginas fraudulentas se buscan asimilar con paginas conocidas, en búsqueda que el usuario no desconfíe y entregue información.

Es por esto, que recurrimos a un dataset de la página Kaggle[[2]](#footnote-2), basado en el repositorio de Alexa.com, una filial de Amazon.com que recoge información de usuarios de internet para generar estadísticas del número de visitas y enlaces a un sitio web. Con las estadísticas Alexa elabora un ranking de posicionamiento de la web. Este ranking muestra la paginas con mayores visitas a nivel mundial, las cuales consideramos confiables.

Para poder comparar nuestros URLs con los obtenidos de Alexa, utilizamos una medida de distancia entre strings: La distancia de Jaro-Winkler. Esta métrica nos explica que tan similares son los strings que se comparan. Esta es una variante de la distancia de Jaro original. Esta distancia se encuentra normalizada, por lo que todos sus valores se encuentran entre 0 y 1. Cuando las secuencias de caracteres son iguales, la distancia es igual a 1 en cambio si no son similares la distancia es igual a 0.[[3]](#footnote-3)

Con esta distancia generamos las variables de distancia con respecto a las primeras páginas del ranking y las que observamos como recurrentes en estafas:

* Distancia respecto de Amazon (*metric\_amazon*)
* Distancia respecto de Instagram (*metric\_instagram*)
* Distancia respecto de Google (*metric\_google*)
* Distancia respecto de WhatsApp (*metric\_whatsapp*)
* Distancia respecto de Twitter (*metric\_twitter*)
* Distancia respecto de Facebook (*metric\_facebook*)
* Distancia respecto de Yahoo (*metric\_yahoo*)
* Distancia respecto de Wikipedia (*metric\_wikipedia*)
* Distancia respecto de Baidu (*metric\_baidu*)
* Distancia respecto de PayPal (*metric\_paypal*)
* Distancia respecto de Mail (*metric\_mail*)
* Distancia respecto de SF Express (*metric\_sfexpress*)
* Distancia respecto de OneDrive (*metric\_onedrive*)
* Distancia respecto de Excel (*metric\_excel*)
* Distancia respecto de Square (*metric\_square*)
* Distancia respecto de Office 365 (*metric\_office365*)
* Distancia respecto de IRS de USA (*metric\_irs*)
* Distancia respecto de Tencent (*metric\_tencent*)
* Distancia respecto de Credit-Agricole (*metric\_* *creditagrecole\_s.a*)

1. Proyecto de ley incorporando el art. 157 ter al Código Penal, tipificando el delito de obtención ilegitima de datos confidenciales ("phishing"). N° de expediente: 2257/11 [↑](#footnote-ref-1)
2. <https://www.kaggle.com/cheedcheed/top1m> [↑](#footnote-ref-2)
3. <https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance> [↑](#footnote-ref-3)