

Minería de Datos para Seguridad

Lectura 1: introducción a la Recopilación de Datos en Fuentes
Abiertas

Aldo Hernández, MEng

IPN SEPI-ESIME Culhuacan

Agosto, 2017

“Para encontrar señales en los datos, debemos aprender a reducir el ruido - no sólo aquel que reside en los mismos, sino el ruido que reside en nosotros. Es casi imposible para las mentes ruidosas percibir más que ruido en los datos”

— Stephen Few, Signal: Understanding What Matters in a World of Noise

Logística

- ▶ Sitio del curso (lecturas, tareas y ejercicios)
<https://github.com/aldhersu/dm2017>
- ▶ Ejercicios y tareas cuentan como participación
- ▶ Se evaluará con proyecto final y reporte en formato *IEEE Conference Template*
https://www.ieee.org/conferences_events/conferences/publishing/templates.html

Requisitos

- ▶ Python 3.5 ≥ <https://www.python.org/downloads/>
- ▶ Anaconda Framework
<https://www.continuum.io/downloads>
- ▶ VirtualBox <https://www.virtualbox.org/>
- ▶ Ubuntu, Debian o Mint Linux (para instalar en su VM)

¿Qué es un dato?

Una colección de **hechos** que pueden ser, cantidades, palabras, observaciones, incluso, pudiendo ser descripciones de eventos de la vida real

- ▶ **Datos cualitativos** es información descriptiva
- ▶ **Datos cuantitativos** es información numérica
 - ▶ **Datos discretos** pueden tomar solo ciertos valores
 - ▶ **Datos continuos** pueden tomar cualquier valor

Análisis Cualitativo y Cuantitativo

► Cualitativo

- ▶ Está siempre sonriente
- ▶ Tiene cabello castaño oscuro
- ▶ Es muy extrovertido

► Cuantitativo

- ▶ Mide 1.72 metros
- ▶ Tiene 51 años
- ▶ Gana 3,460,475 MXN al año



¿Qué es la Minería de Datos?

Es una **ciencia** enfocada al descubrimiento de estructuras para futuras predicciones en conjuntos grandes de datos

- ▶ **Aprendizaje no supervisado** Descubrir estructuras

Por ejemplo , dadas ciertas observaciones $\gamma_1, \dots, \gamma_n$, aprender estructuras de grupos subyacentes basadas en similitud

- ▶ **Aprendizaje supervisado** hacer predicciones

Por ejemplo, dadas ciertas observaciones $(\gamma_1, \varphi_1), \dots, (\gamma_n, \varphi_n)$, encontrar un modelo para predecir φ_i de γ_i

¿Qué es la Recolección de Datos en Fuentes Abiertas?

Denominada OSINT (Open Source Intelligence), es una **técnica** que permite recolectar información dispersa en diferentes fuentes abiertas

- ▶ Utiliza **cualquier canal abierto** para recuperar información
- ▶ **No necesita interacción directa** con el objetivo en cuestión
 - ▶ Reduce las posibilidades de ser descubierto

Ventajas de Recolección de Datos en Fuentes Abiertas

Es una **técnica** que permite recolectar información dispersa en diferentes fuentes abiertas

- ▶ Explorar redes y sitios sociales
- ▶ Explorar datos sensibles de grandes organizaciones (incluso si tienen algún mecanismo de defensa)
- ▶ Cada dato que tenemos relacionado a nuestro número móvil, correo electrónico, perfil en sitios y redes sociales **crea inteligencia**

Ciclo de la Inteligencia

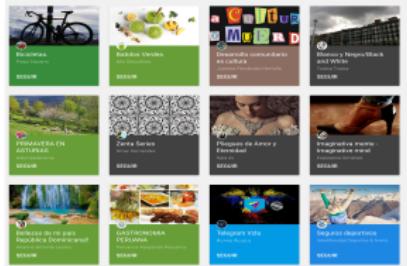
El ciclo de la inteligencia es aquella secuencia en la cual, se obtiene **información**, para ser transformada a **inteligencia** y después es presentada:

- ▶ Dirección
- ▶ Obtención
- ▶ Elaboración
- ▶ Difusión

¿Qué se necesita para realizar una búsqueda en Fuentes Abiertas?

- ▶ Navegador (de preferencia Opera , ya que tiene una VPN integrada)
- ▶ Protección mediante VPN o TOR
- ▶ Máquinas Virtuales (vBox de preferencia)
- ▶ Cifrar los datos a presentar (Veracrypt o Kleopatra)
- ▶ Un teléfono móvil (SIM Express)
- ▶ Crear perfiles anónimos en Redes Sociales

¿De dónde extraen y usan nuestros datos?



Google Plus

Find new customers now,
with Google AdWords



Google Analytics

Google maestría en seguridad

Todos Imágenes Mapas Vídeo Noticias Más Preferencias Herramientas

Cerca de 25,000 resultados (0.02 segundos)

Maestría en Ingeniería en Seguridad y Tecnologías de la Información
www.ipn.mx/sepe/estudios/maestrias/maestria-en-ingineria-en-seguridad-y-tecnologias-de-la-informacion.html • Traducir esta página
SEPESI ESI: Maestría en Ingeniería en Seguridad y Tecnologías de la Información en temas relacionados con la Seguridad y Transmisión de la Información... ATENCIÓN ALUMNOS MBTI:
Plan de estudio Requerimientos Lineas de investigación (LGAC)

SEPI ESIME Culiacán
www.ipn.mx/sepe/estudios/maestrias/maestria-en-seguridad-y-tecnologias-de-la-informacion.html • Traducir esta página
ESIME Culiacán - Edificio de galerías. Brevemente. La Sección de Estudios de Postgrado e Investigación (SEPI) de ESIME - DCE-MCIM-ICSE-MBTTI, seo.
Maestría en Ciencias de... Especialidad en Seguridad - Nuestra área

Convocatoria MBTI 2017-2019 - SEPI ESIME Culiacán - IPN
www.ipn.mx/sepe/estudios/maestrias/maestria-en-seguridad-y-tecnologias-de-la-informacion.html • Traducir esta página
El plan de estudios de la MBTI contempla un número de 54 créditos o más... se desarrollan dentro de las iniciativas de la SEPI de la ESIME Culiacán...

Google Ads

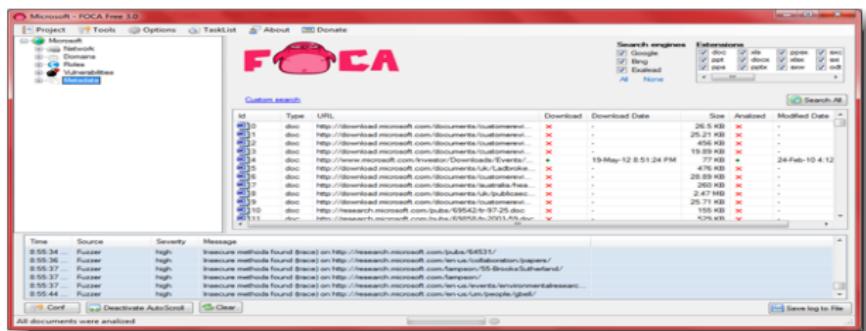
a hernandezs1325@alumno.ipn.mx

Google Search

¿De dónde se pueden extraer nuestros datos?

Se puede también extraer información a partir de los **metadatos**

- ▶ Cada metadato describe a otro dato
- ▶ Es contenido informativo de un objeto denominado recurso
- ▶ Se encuentran en archivos ofimáticos (.dox,.pdf,.xls, etc...) y también en imágenes



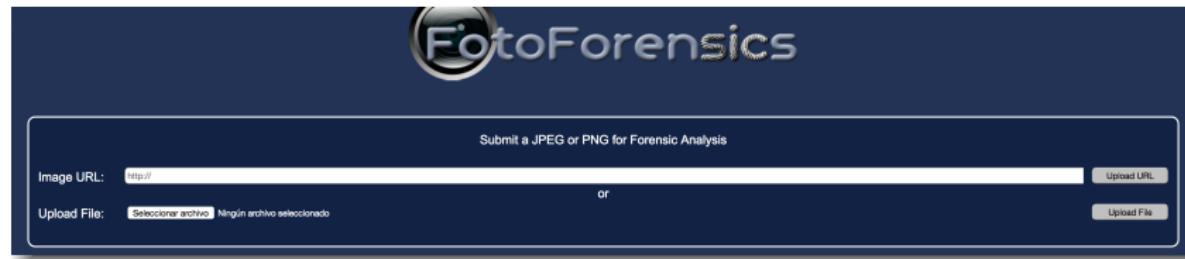
El software Foca permite extraer metadatos de diversos documentos, incluso de páginas web

Extracción de Metadatos



<http://www.exifviewer.org>

Extracción de Metadatos en Imágenes



<http://fotoforensics.com/>

¿Qué datos sociales podemos extraer?



Bill Gates
Co-chair, Bill & Melinda Gates Foundation
Seattle y alrededores, Estados Unidos | Filantropía

Actual Bill & Melinda Gates Foundation, Microsoft
Educación Harvard University
Sitios web Blog

influencer

Ve el perfil completo de Bill.
¡Es gratis!

Tus colegas, tus compañeros de clase y 500 millones más de profesionales están en LinkedIn.

[Ver el perfil completo de Bill](#)

LinkedIn comparte nuestros datos profesionales

¿Qué relaciones y afinidades sociales tenemos ?

Ricardo Flores Magón (Encuentro Reflexiones Anarquistas)

Add Friend Message More

Studied at ENAH - Escuela Nacional de Antropología e Historia
Lives in Mexico City, Mexico

About Photos Friends

Photos

CUENTRO REFLEXIONES ANARQUISTAS DEDICADO A RICARDO FLORES MAGÓN

See All Photos >

Facebook contiene nuestra información personal y vínculos con otros contactos

¿Qué opiniones personales tenemos ?

Tweets Sigiendo Seguidores
3.347 88 2,77 M

Tweets **Tweets y respuestas** **Multimedia**

Andrés Manuel @lopezobrador_ · 6 ago.
Reforma, apoyador del "nuevo frente" de la mafia en contra nuestra. Dice que voy a usar mi "dedito" en la elección de MORENA. No calumnien.

988 1,5K 2,1K

Andrés Manuel @lopezobrador_ · 6 ago.
Luego de la gira por Sudamérica, reiniciamos nuestros recorridos por el país.
Estuvimos en Pinos, Loreto y Ojocaliente, Zacatecas



334 879 1,6K

Twitter refleja información social acerca de contextos específicos

¿Qué relaciones y afinidades sociales tenemos ?

Ricardo Flores Magón (Encuentro Reflexiones Anarquistas)

Add Friend Message More

Studied at ENAH - Escuela Nacional de Antropología e Historia
Lives in Mexico City, Mexico

About Photos Friends

Photos

CUENTRO REFLEXIONES ANARQUISTAS DEDICADO A RICARDO FLORES MAGÓN

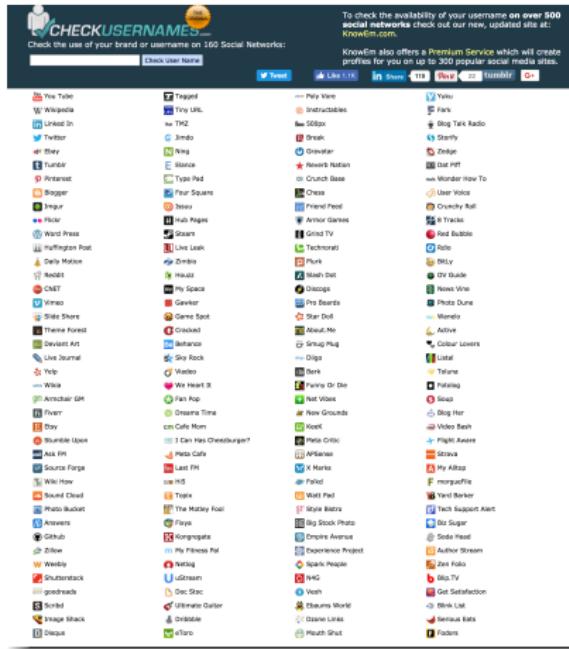
See All Photos >

Facebook contiene nuestra información personal y vínculos con otros contactos

¿Qué recursos se pueden explotar de las Fuentes Abiertas ?

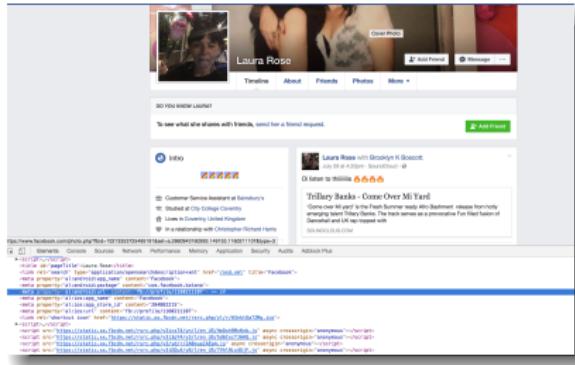
- ▶ APIs de Redes Sociales
- ▶ Web Scrapping
- ▶ Recon Avanzado
- ▶ ¿Dark/Deep Web?

Todos tenemos algo que nos identifica en la web



UsernamecCheck nos permite ver cuantos sitios sociales tienen nuestro username ocupado

Facebook y las búsquedas avanzadas (1/4)



Todos los perfiles de Facebook tienen un identificador único asociado
embebido como metadato

Facebook y las búsquedas avanzadas (2/4)

Podemos encontrar gran información relacionada a la actividad del usuario utilizando el endpoint */search/id – usuario/*

- ▶ publicaciones
- ▶ lugares visitados
- ▶ lugares gustados
- ▶ páginas gustadas
- ▶ fotos etiquetadas
- ▶ fotos comentadas
- ▶ fotos públicas
- ▶ aplicaciones utilizadas
- ▶ videos
- ▶ amigos

Facebook y las búsquedas avanzadas (3/4)

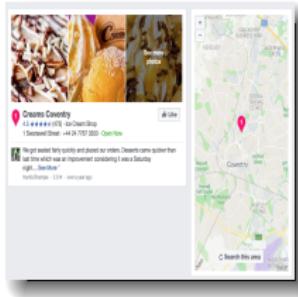
Podemos encontrar gran información relacionada a la actividad del usuario utilizando el endpoint `/search/id – usuario/`

- ▶ grupos
- ▶ colaboradores
- ▶ familiares
- ▶ grupos

Relaciones con otros usuarios como

- ▶ detalles en común
- ▶ amigos en común
- ▶ fotos en común
- ▶ lugares en común
- ▶ eventos en común
- ▶ grupos en común

Facebook y las búsquedas avanzadas (4/4)

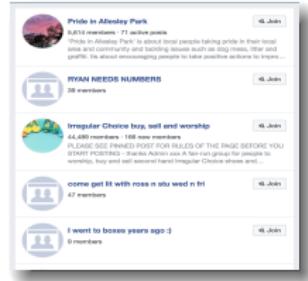


Lugares Visitados

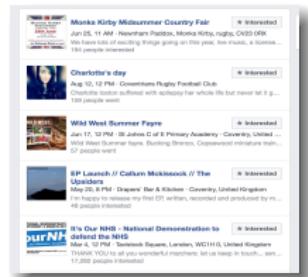


Fotos públicas

a hernandezs1325@alumno.ipn.mx



Grupos



Eventos

Twitter y los metadatos (1/3)

La información en Twitter puede incluir metadatos relacionados al perfil y a cada tweet

- ▶ media (imágenes y videos)
- ▶ enajenación
- ▶ ubicaciones geográficas (por referencia y GPS)
- ▶ orígenes de cabeceras de HTTP
- ▶ dominios relacionados
- ▶ bolsas de palabras
- ▶ orígenes geo-localizados de tendencias
- ▶ interacción con otros perfiles

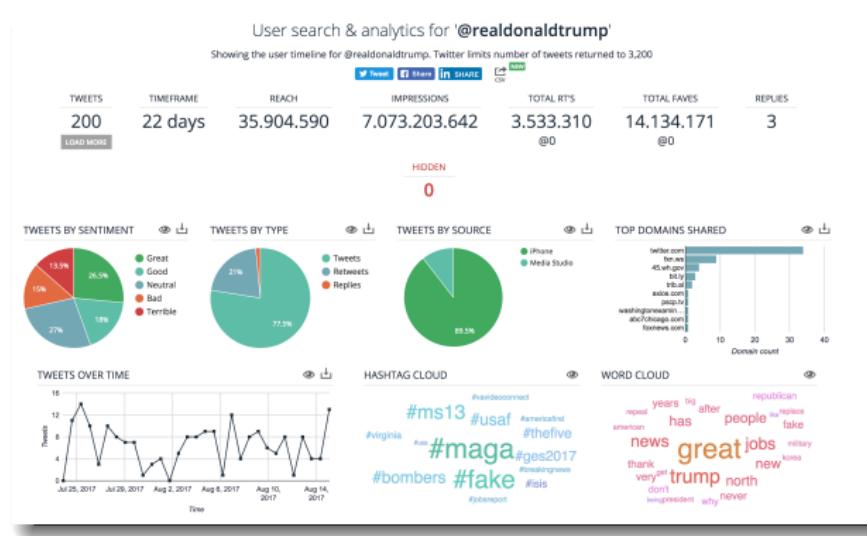
Twitter y los metadatos (2/3)



Se pueden monitorear los orígenes de hashtags en tiempo real

Twitter y los metadatos (3/3)

Entre la analíticas se puede encontrar diversos datos por cada metadato embebido



Ejemplo de metadatos en Twitter