

Creación de bot en Twitter

Eduardo Graván Serrano

Índice

Los puntos a tratar en esta presentación:

1. Introducción
2. Objetivos
3. Estado del arte
4. Desarrollo del sistema
5. Conclusiones
6. Trabajo futuro



Introducción

El uso de redes sociales y, en concreto, Twitter:

- Ha visto un incremento significativo en los últimos años. Twitter contaba con un total de 353 millones de usuarios activos mensualmente en 2020.
- Las redes sociales son medios de comunicación abiertos al público y utilizadas para compartir sus opiniones sobre diversos temas:
 - El análisis del uso de las redes sociales tiene un gran interés académico y social.
 - El análisis “manual” de la información publicada en redes sociales es impracticable debido al gran volumen de datos.
 - Necesidad de mecanismos de automatización en cuanto a la búsqueda y recuperación de información desde redes sociales.
- Posibilidad de manipular las corrientes de opinión de los usuarios a través de la publicación automatizada de mensajes.

Objetivos

En base a los puntos expuestos en la introducción, se tienen los siguientes objetivos:

- Creación de un sistema que permita a un usuario de la red social Twitter:
 - Realizar búsquedas en la plataforma a través de una interfaz gráfica de usuario.
 - Crear listas de seguimiento de usuarios que vea interesantes.
 - Posibilidad de exportar estos datos para su posterior análisis.
- Estudio del funcionamiento y operativa de las redes de bots en Twitter:
 - Búsqueda de información sobre las botnets en Twitter.
 - Creación de un sistema como prueba de concepto de una botnet funcional en la red social Twitter.

Estado del arte - Twitter

La red social Twitter:

- Permite la publicación de mensajes cortos (tweets) de hasta 280 caracteres.
 - Imágenes, vídeos, audios, etc.
- Otros usuarios de la plataforma pueden responder a estos tweets, creando hilos de mensajes.
- Los usuarios pueden reaccionar a un tweet de dos formas distintas:
 - Dando “me gusta” o “like”.
 - Dando “retweet”, lo que hace que el tweet se publique en el perfil de la persona que ha reaccionado.
 - Más adelante se añadió la opción de citar (retweet + nuevo tweet).

Estado del arte - Twitter (II)

Estos mensajes cortos o tweets:

- Pueden contener *hashtags*, formados por el carácter #. Los hashtags son utilizados para relacionar unos tweets con otros que hagan uso del mismo hashtag. (Ej. #ciberseguridad).
- Twitter recopila información sobre la afluencia de tweets que están haciendo uso de un hashtag en un determinado momento, creando lo que se conoce como tendencias (hashtags más populares).

Estado del arte - Métodos de interacción con Twitter

Existen dos métodos principales para la interacción con Twitter de forma automatizada:

- API oficial de desarrolladores.
- Métodos de scraping.

Estado del arte - API de Twitter

La API oficial de desarrolladores de Twitter:

- Requiere que un empleado de Twitter te acepte la petición para acceder a la API.
- Funciona como una API ReST por HTTP, permitiendo diferentes formas de autenticación.
- Al ser una forma de acceso autenticado, las acciones realizadas se harán mediante la cuenta de usuario asociada a la cuenta de desarrollador de Twitter.

Estado del arte - API de Twitter (II)

La API oficial de desarrolladores de Twitter:

- Esta API oficial tiene varias implementaciones en distintos lenguajes de programación (Ej: Tweepy para Python, Twitter4J para Java).
- Esta API permite realizar una gran cantidad de acciones como:
 - Realizar búsquedas en Twitter en base a parámetros definidos por el usuario (limitado a 7 días atrás desde la fecha en la que se realiza la búsqueda).
 - Recuperar información de cuentas de usuario.
 - Publicar tweets.
 - Gestión de “me gustas” y retweets.
 - Recuperar tweets publicados en tiempo real.

Estado del arte - API de Twitter (III)

Esta API cuenta con funcionalidades o endpoints **premium**:

- **Búsqueda 30-day**: permite realizar búsquedas de tweets con una antigüedad de hasta 30 días.
- **Búsqueda full-archive**: permite realizar búsquedas en el archivo de Twitter, esto es, sin limitación en cuanto a su fecha de publicación.
- **Account activity**: permite registrar una serie de cuentas para realizar un seguimiento de su actividad en la red social.

Esta API premium es de pago, aunque permite realizar ciertas consultas mensuales de forma gratuita.

Estado del arte - API de Twitter (IV)

En cuanto a limitaciones de la API:

- Al ser autenticada, todas las acciones que se hagan a través de ella se harán con la cuenta de Twitter asociada a la cuenta de desarrollador.
- Limitaciones de búsquedas normales a 7 días anteriores.
- Precios muy altos en las suscripciones de la API premium.

Estado del arte - Métodos de Scraping

El segundo método es a través de mecanismos de scraping web sobre la plataforma. Los principales métodos dentro del scraping son:

- Haciendo uso de la herramienta **Twint**.
- Utilizando soluciones de scraping web como puede ser el framework **Selenium**.

Estado del arte - Métodos de Scraping (II)

En cuanto a Twint:

- Tiene dos formas de operación:
 - Como librería de Python3.
 - Como herramienta independiente de línea de comandos.
- Permite realizar todo tipo de búsquedas parametrizadas por:
 - Nombre de usuario.
 - Palabras clave del tweet.
 - Fecha máxima y/o mínima.
 - Etcétera.
- Permite exportar y almacenar los resultados de diferentes formas como:
 - Base de datos SQLite.
 - Clúster de Elasticsearch.
 - Exportación a TXT, CSV y JSON.

Estado del arte - Métodos de Scraping (III)

En cuanto a puntos fuertes y limitaciones de Twint:

- No tiene límite de consultas de ningún tipo.
- Si una determinada búsqueda recupera más de 3200 resultados, solo puede devolver los primeros 3200.
- Al ser una herramienta que funciona de forma no autenticada en la plataforma, solo permite realizar búsquedas (no permite la publicación de tweets, dar “me gusta”, retweet, etc.).

Estado del arte - Métodos de Scraping (IV)

En cuanto a Selenium:

- Es un framework enfocado a la automatización de tests en aplicaciones web.
- Permite, mediante la ejecución de un webdriver, interactuar con páginas web mediante un navegador de una forma automatizada.
- Tiene librerías en distintos lenguajes de programación como Python o Java, permitiendo escribir código que interactúa con las páginas web a través del webdriver.

Estado del arte - Métodos de Scraping (V)

Debido a que mediante el uso de Selenium simplemente se está interactuando con la interfaz web de Twitter:

- Se pueden realizar búsquedas automatizadas sobre la plataforma con todas las opciones de parametrización que ofrece Twitter.
- Permite mandar tweets, gestionar “me gustas” y retweets, etc.
- En definitiva, permite realizar todas las acciones que podría realizar un usuario normal de la plataforma.

Estado del arte - Métodos de Scraping (VI)

Como contrapunto al uso de Selenium, Twitter actualizó sus términos de uso en 2017 prohibiendo ciertas acciones como:

- Autenticarse de forma automatizada en la plataforma mediante métodos que no sean la API oficial de desarrolladores.
- Publicar contenidos que caigan bajo la definición de spam.
- Dar “me gusta” de forma automatizada.
- Seguir o dejar de seguir cuentas en masa.

Si Twitter detecta este tipo de actividad en alguna cuenta de usuario, estas son baneadas permanentemente de la plataforma.

Estado del arte - Redes de bots en Twitter

En cuanto a los bots en la plataforma Twitter:

- Se estima que entre un 9% y un 15% de las cuentas totales de usuario son gestionadas totalmente por bots.
- Gran cantidad de código publicado que hace las veces de bot en Twitter, 20000 repositorios en GitHub.
- Pueden realizar gran cantidad de funciones, no necesariamente son maliciosos.

Estado del arte - Redes de bots en Twitter (II)

Las redes de bots en Twitter suelen tener los siguientes objetivos:

- Campañas de SPAM.
- Campañas de phishing.
- Publicitar información falsa, o inflar números en publicaciones políticas.

Su objetivo principal es que sus tweets lleguen a la mayor cantidad de gente posible. Para ello:

- Intentan conseguir que los usuarios sigan a las cuentas de bots.
- Envían mensajes directos a los usuarios.
- Principalmente, se adhieren a hashtags para publicitar sus tweets al máximo.

Estado del arte - Redes de bots en Twitter (III)

Durante el estudio de las redes de bots en Twitter:

- Se encontró mucha información sobre sus formas de actuación, mecanismos de detección de bots, etc.
- No se encontró prácticamente nada sobre la arquitectura real de estas redes y cómo funcionan internamente.
- El poco código de ejemplo encontrado no era aplicable a una situación real, ya que hacía uso de la API de desarrolladores de Twitter.

Desarrollo del sistema - Introducción

Como ya hemos anticipado previamente, el proyecto está separado en el desarrollo de dos sistemas independientes:

- Aplicación de escritorio capaz de realizar y guardar el resultado de búsquedas en la plataforma Twitter.
- Sistema capaz de hacer las veces de una red de bots en Twitter.

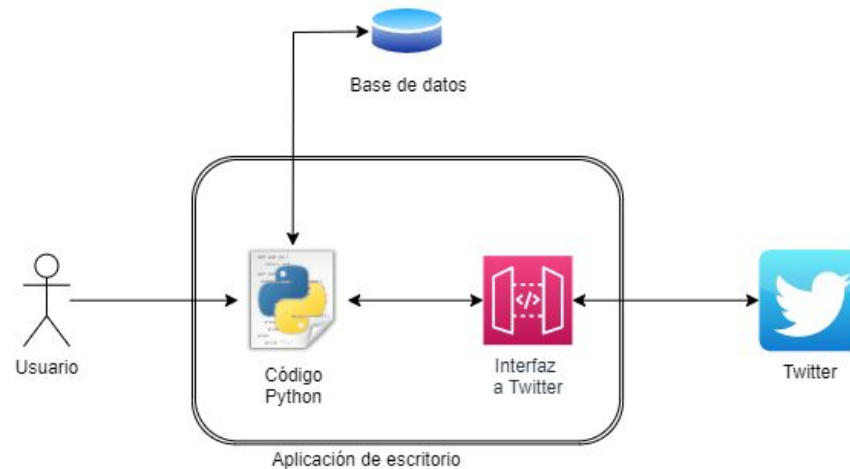
Desarrollo del sistema - Aplicación de escritorio

En cuanto a la aplicación de escritorio, está compuesta de las siguientes partes:

- Aplicación de escritorio.
- Interfaz de acceso a Twitter mediante la API.
- Base de datos.

Desarrollo del sistema - Aplicación de escritorio (II)

La arquitectura general del sistema es la siguiente:



Desarrollo del sistema - Aplicación de escritorio (III)

En cuanto a la base de datos:

- Esquema simple basado en una única tabla.
- Almacena información sobre las cuentas seguidas por el usuario desde la interfaz de la aplicación.
 - **twitter_handle**: tipo Text, PK.
 - **followed_timestamp**: tipo Time.
- Implementada en **SQLite**.

Desarrollo del sistema - Aplicación de escritorio (IV)

En cuanto a la aplicación de escritorio:

- Implementada en **Python3**.
- Uso del framework **Qt** para construir las interfaces gráficas de usuario.
- Implementación con la librería **Tweepy**, interactuando con la API oficial de desarrolladores de Twitter.

Desarrollo del sistema - Aplicación de escritorio (V)

La aplicación de escritorio permite realizar las siguientes acciones:

- **Búsquedas rápidas en Twitter:** recupera los 200 tweets más recientes que coincidan con los criterios de búsqueda con hasta 7 días de antigüedad. Sin límite de consultas en la API.
- **Búsquedas con la API premium:**
 - **API de archivo:** todos los tweets almacenados en la plataforma.
 - **API 30-day:** tweets publicados con hasta 30 días de antigüedad

Desarrollo del sistema - Aplicación de escritorio (VI)

Adicionalmente, la aplicación permite:

- Exportar los resultados de cualquiera de estos tres tipos de búsqueda en formato CSV para su posterior análisis en otras aplicaciones.
- Crear una lista de seguimiento de usuarios en base a los resultados de la consulta:
 - Con esta lista se pueden lanzar consultas rápidamente que recuperen tweets solo de un usuario en concreto, o de todos los usuarios de la lista.

Desarrollo del sistema - Aplicación de escritorio (VII)

Búsqueda con la API de archivo, limitando por fechas:

TFM

Búsqueda rápidaBúsqueda en profundidadCuentas seguidas

Búsqueda en profundidad en Twitter

Consulta:

Cuenta de usuario:

☒ Búsqueda en archivo









☐ Búsqueda 30 días

☒ Fecha - Desde:

☒ Fecha - Hasta:



Buscar

Imagen	Usuario	Seguimiento	Fecha	Tweet
	pablokdc		2008-06-20 23:55:15	Mañana a las 12 cojo un avion para Madrid y a 19:40 de Madrid salgo para Berlin hasta el 27 de junio que vuelvo para ver a Bob Dylan en Vigo
	periodistas		2008-06-20 23:42:31	«Quiero hacer Periodismo en la Carlos III, en Madrid» - Hoy Digital: «Quiero hacer Periodismo en la C.. http://tinyurl.com/6emo46
	324cat		2008-06-20 23:06:48	Les restes mortals dels dos militars espanyols morts a Bòsnia arriben a Madrid: Les despulles del tinent S.. http://tinyurl.com/6mg2ak
	mr_sparxx		2008-06-20 22:28:54	As you can imagine: no-party friday night, but cooperatively writing a script with my brothers in Madrid for a short-film contest.

Desarrollo del sistema - Aplicación de escritorio (VIII)






Panel de seguimiento de cuentas:

TFM

Búsqueda rápida Búsqueda en profundidad Cuentas seguidas

Panel de seguimiento de cuentas

Actualizar lista de seguimiento 🔍

Imagen	Usuario	Seguimiento	Creación	Buscar	Siguiendo desde
	ConectividadCO	✓	2010-04-05 14:41:39	🔍	2021-07-25 18:29:44
	ngelalleixa	✓	2012-10-25 09:29:40	🔍	2021-07-25 18:29:49
	EsGeeks	✓	2017-02-19 02:33:00	🔍	2021-07-25 18:29:53
	sanchezrum	✓	2011-07-02 15:32:53	🔍	2021-07-30 15:58:55
	chemaalonso	✓	2010-01-29 06:17:05	🔍	2021-08-02 14:00:03

Desarrollo del sistema - Botnet

Debido a la poca información encontrada sobre la arquitectura y funcionamiento de las redes de bots en Twitter reales, se decidió hacer una implementación de cero, siguiendo el esquema de una botnet de malware:

- Arquitectura Cliente - Servidor.
- Servidor de *Command & Control* (C2) que envía comandos.
- Bots independientes que ejecutan los comandos.

Desarrollo del sistema - Botnet (II)

Este sistema está implementado de la siguiente forma:

- Tanto el servidor como los clientes están escritos en **Python3**.
- El servidor hace uso de la **API de desarrolladores** de Twitter tanto para autenticarse en la plataforma como para recuperar tweets mediante la API de **stream**.
- Los bots se han implementado automatizando tareas sobre la interfaz web de Twitter mediante la librería de **Selenium** para Python.

Desarrollo del sistema - Botnet (III)

El flujo de ejecución del servidor C2 es el siguiente:

1. El servidor se autentica en Twitter a través de la API oficial de desarrolladores.
2. Se lanza un hilo encargado de escuchar conexiones UDP entrantes para registrarlos como bots en activo.
3. Se lanza un segundo hilo leyendo tweets de Twitter a través de la API de stream, en base a un parámetro configurable en el código.
4. Cuando un usuario publica un tweet que coincida con la búsqueda, se envía como comando a los bots que haya registrados a través de UDP.

Desarrollo del sistema - Botnet (IV)

El flujo de ejecución de los bots es el siguiente:

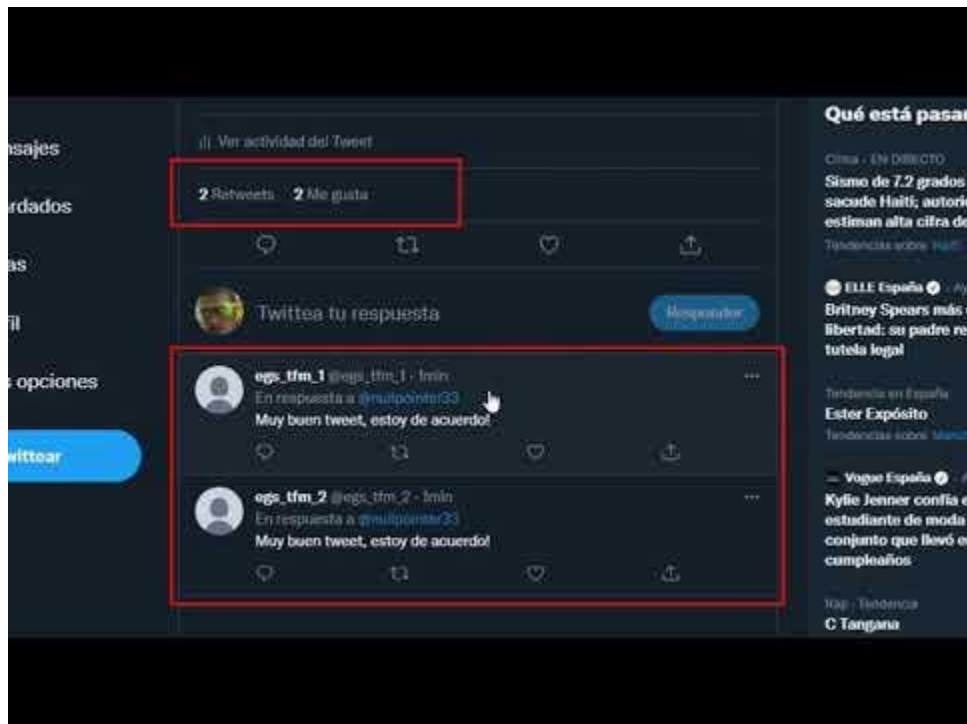
1. El bot recupera las credenciales de la cuenta de Twitter que utilizará desde el fichero de configuración que se le ha pasado por parámetro.
2. Inicializa el webdriver de Chrome para Selenium.
3. Se autentica en la red social con las credenciales de la cuenta mediante el formulario de login.

Desarrollo del sistema - Botnet (V)

El flujo de ejecución de los bots es el siguiente:

4. Una vez se ha autenticado, manda un mensaje UDP al servidor para registrarse en la lista de bots activos.
5. Al recibir confirmación del servidor, se queda a la espera de recibir comandos.
6. Cuando recibe un comando (tweet), accede a la URL a través del webdriver, le da “me gusta”, retweet y comenta un mensaje preestablecido en el código.

Desarrollo del sistema - Botnet (VI)



Conclusiones

Como conclusiones de la parte del proyecto relacionada con el desarrollo de la aplicación para búsquedas en Twitter:

- La API de desarrolladores de Twitter permite la automatización de tareas en la red social de una forma rápida y sencilla de implementar.
- Gran cantidad de implementaciones de esta API en diferentes lenguajes de programación.
- Grandes limitaciones en los rates de consultas, limitación a búsqueda en 7 días, API premium muy cara, etc.

Conclusiones (II)

En cuanto a la botnet en Twitter:

- Se debe remarcar lo sencillo que es desplegar una botnet simple pero eficaz y fácilmente escalable.
- La existencia de este tipo de herramientas de automatización atenta directamente contra la posibilidad que brinda las redes sociales de dar una plataforma a sus usuarios para compartir sus opiniones.
- Es especialmente peligroso teniendo en cuenta el nivel de penetración en la sociedad actual que tiene el uso de este tipo de redes sociales.

Trabajo futuro

En cuanto a la aplicación de escritorio:

- Incrementar la funcionalidad actual para realizar búsquedas, añadiendo mayor parametrización en función de geolocalización, etc.
- Suplir las deficiencias de la API, haciendo una implementación híbrida entre la API de desarrolladores y otras herramientas como Twint.
- Añadir funcionalidades en función de las necesidades del usuario (permitir la publicación de tweets, dar “me gusta” o retweet, etc.).

Trabajo futuro (II)

En cuanto a la botnet:

- Creación de un sistema para automatizar la creación de cuentas de Google, con el objetivo de crear cuentas de Twitter asociadas.
- Implementar mecanismos para detectar cuando un bot ha muerto para eliminarlo de la lista de bots activos desde el servidor C2.
- Estudiar los mecanismos de detección que tiene la plataforma para este tipo de comportamiento automatizado y adaptar el código de los bots en función de los resultados de este estudio.



¡Gracias!

¿Preguntas?