WhatsApp Web: Public Group Message Scraper

Prepared by

Justin Strelka

&

Derek Holsapple

Advised by

Steven Beaty & Thyago Mota

January 25, 2021

Executive Summary

Governments across the globe are in a constant struggle with underground human trafficking organizations operating within their borders. Given this issue, Brighter Tomorrow Solutions, a non-profit organization offers its student resources to collaborate with both Dr. Steven Beaty and Dr. Thyago Mota. The desired outcome of efforts aims to address the possible connection between public domain WhatsApp groups used for soliciting victims and activity on the dark web. By doing so, we expect to find personally identifiable information that can be cross referenced with data collected from the anonymous dark web. Cross referenced findings could be of use to assist intelligence organizations across the globe in identifying individuals connected to underground human trafficking organizations.

Problem

The WhatsApp platform is a messaging service used by populations across the globe. It has been speculated that within the public group chat domains of the service underground human trafficking organizations conduct soliciting campaigns for new recruitments. Unwittingly, individuals are lured into promises of good working conditions, housing, and other life necessities that can be scarce amongst impoverished regions. Upon arrival the victims are forced to into slavery, living in inhumane conditions.

Solution

Through cross-referencing personally identifiable information extracted from public chatrooms on the WhatsApp Web application with traffic from the dark web, authorities may be able to allocate investigation resources in a more efficient manner. Currently, WhatsApp is owned by Facebook and the public does not have access to the native API. Brighter Tomorrow Solutions proposes using a browser automation software that can be used to programmatically extract data in large quantities from public chatrooms for cross-referencing.

Objectives

Brighter Tomorrow Solutions aims to achieve the following objectives for this project:

- Create an application using browser automation technology to programmatically extract message data from public chatrooms on WhatsApp Web.
- Design and create a GUI to drive the application and enhance UX.
- Implement a database infrastructure to allow for efficient querying and long-term data storage.

Team Members

- Justin Strelka
- Derek Holsapple

Software Engineering

- **Life cycle**
    - SCRUM (2-week sprints)

- **Project Management Tools**
    - GitHub Organization/Team
    - GitHub Project Kan-Ban board

- **Time Management Tools**
    - Toggl

- **TDD**
    - Use pytest library to write unit tests at the same time code is written.

Minimum Viable Product

- Application that can automatically extract message data from WhatsApp Web public groups.
- GUI to drive application and enhance UX.
- Organized database for long term data storage and efficient querying.
- Documentation in the form of a README.md file within the GitHub repo. Documentation will outline: software dependencies, installation, use of features.

Stretch Goals:
- Develop sample AI parser to cross-reference WhatsApp message data with data collected from the dark web.

Timeline

PHASE	DESCRIPTION	DEADLINE
Objective 0	Set-up disposable phone to register with WhatsApp Web account	Feb 12,2021
Objective 1	Create WhatsApp message extraction App	March 01,2021
Objective 2	Design GUI to enhance UX	April 01,2021
Objective 3	Implement database for efficient querying and long-term storage	May 01, 2021
Objective 4	Develop sample AI parser	TBD

Deployment Plan

OBJECTIVE 1
Create WhatsApp message scraping App.

ACTION	ASSIGNED TO	DEADLINE	PROGRESS
Create GitHub repo to store and host source code for delivery	Justin Strelka	1/23/21	Complete
Design application mapping template to display functionality	Justin Strelka/ Derek Holsapple	2/10/21	Upcoming
Write prototype code and unit tests to conduct baseline functionality	Justin Strelka/ Derek Holsapple	2/24/21	Upcoming
Refactor prototype code and unit testing suite for production release	??	2/28/21	Upcoming
Create documentation for deploying application and collecting necessary libraries	??	2/28/21	Upcoming

OBJECTIVE 2
Design GUI to enhance UX

ACTION	ASSIGNED TO	DEADLINE	PROGRESS
Create schematic for input options and how they relate	??	3/7/21	Upcoming
Design template for styling GUI	??	3/14/21	Upcoming
Develop code to create GUI along with unit tests	??	3/21/21	Upcoming
Test and refactor GUI where necessary	??	3/31/21	Upcoming

Implement database for efficient querying and long-term storage.

ACTION	ASSIGNED TO	DEADLINE	PROGRESS
Make decision for best use case technology.	??	4/2/21	Upcoming
Create database schema	??	4/10/21	Upcoming
Design interface to populate database with message data	??	4/30/21	Upcoming

Documentation

Within the project GitHub repo provide a README.md file that provides the user with instructions for using the application. The documentation will include a list of required software and libraries needed to run the application along with links for obtaining them. The documentation should also include the necessary steps for running the application on Windows, MAC, and Linux.

The documentation will provide detailed instruction on the behavior of all features built into the application and any default values when applicable.

Risks and Constraints

Risks:

- There is no guarantee that selenium will interact nicely with WhatsApp Web application.
- WhatsApp Web is a dynamic application and may be difficult to predict the html structure of the web page to precisely scrape message data.
- WhatsApp Web page my change its GUI without notice essentially causing the configuration of the WhatsApp message scraper to become outdated.
- It is possible we are noticed within a chatroom and targets may move to another platform requiring yet another message scraping tool to be configured.
- We have the potential to give up our own personal identifiable information while conducting research such as: IP address, Name, etc.

Constraints:

- Extraction application must interact with WhatsApp Web since the native API is not publicly available.
- Message history may or may not be visible using selenium. We may only be able to see future messages.