

# Undercover Recovery



Logan Chamberlain, Tyler Lukacz  
CSCE A470 Capstone  
Client: FBI Anchorage Alaska VC1



# Operational Need

- Client - Anchorage FBI VC1 - currently manually reviews social media sites looking for suspicious content.
  - Time Consuming.
  - Ineffective.
  - Infeasible.
- Client requires an automated tool.



# Solution Approach

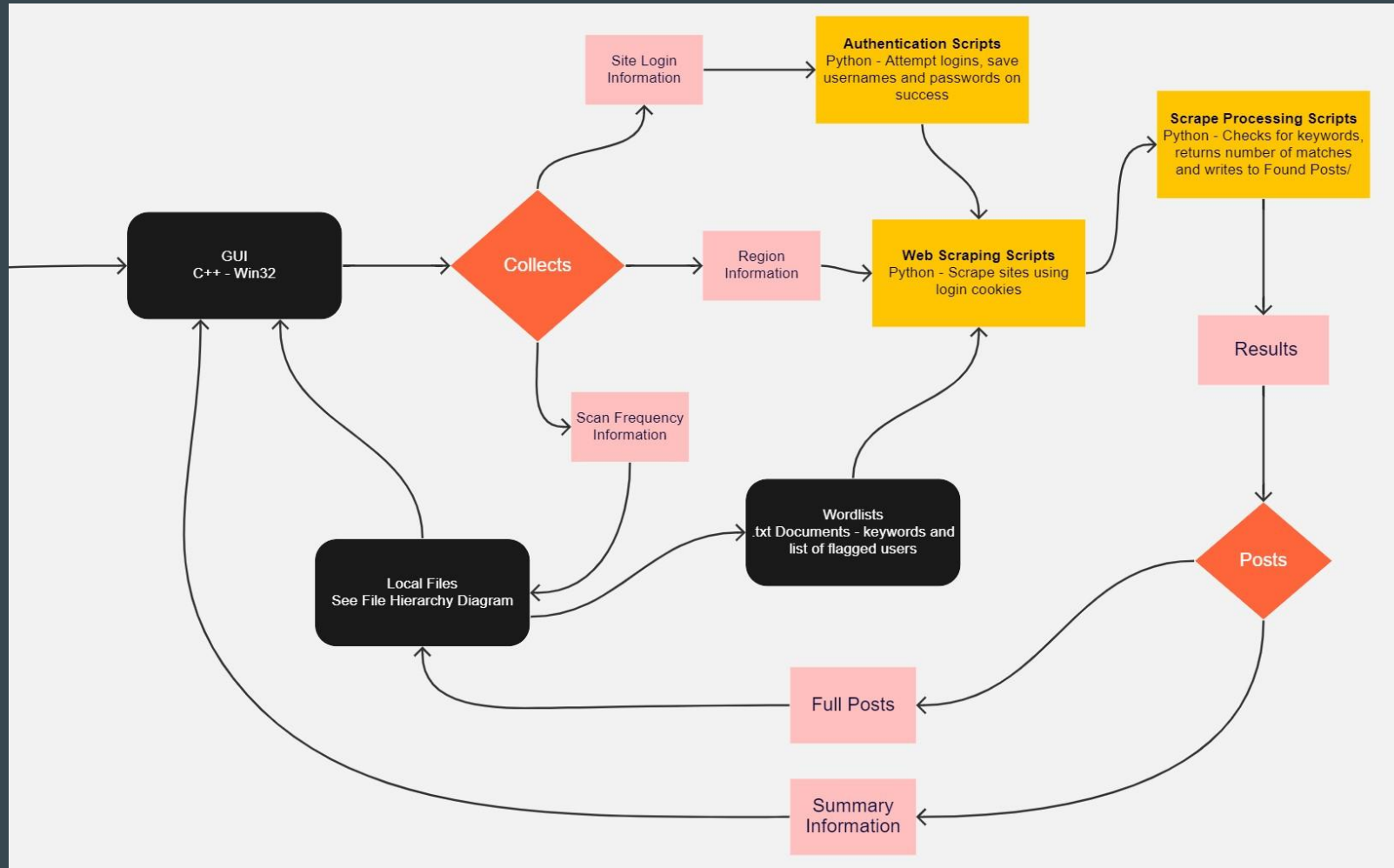
A desktop application capable of scraping social media sites looking for keywords and posts made by flagged authors, limited to Alaska.



- Application runs a script to authenticate logins to social media sites.
- Application runs scripts to pull web content for each target site and saves flagged content.

# Project Components

- The Product is constructed of 3 parts:
  - A graphical interface that collects required information.
  - The python scripts for authentication and scraping of web content.
  - A file structure for saving the suspicious posts meeting region and keyword criteria.



# GUI Specifications

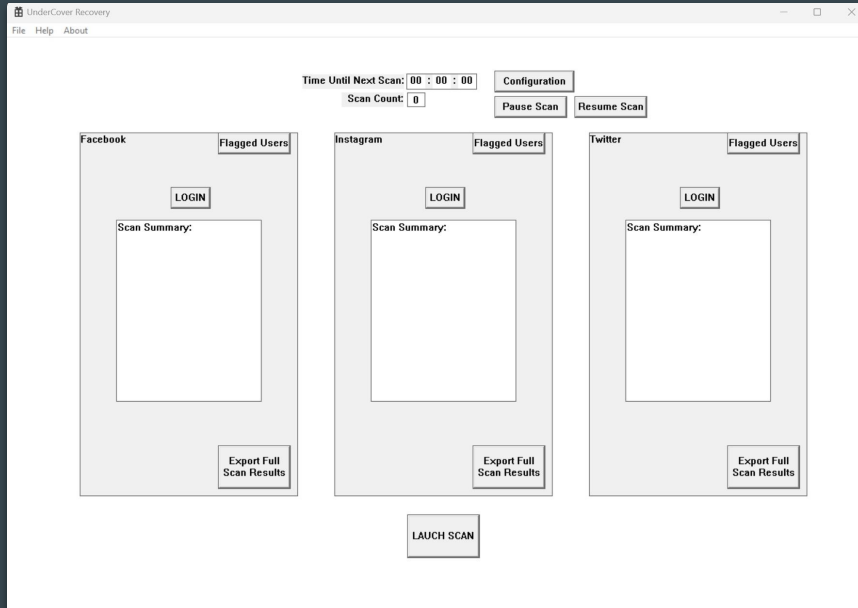
- The GUI is written in C++, using the Win32 Windows API, runs python scripts via timer and frequency configurations.
- The GUI saves login information and allows choice of keywords and locations to scrape.
- Graphic Design was built using the 24-Bit Bitmap image format supported by the Win32 API.

# GUI Specifications

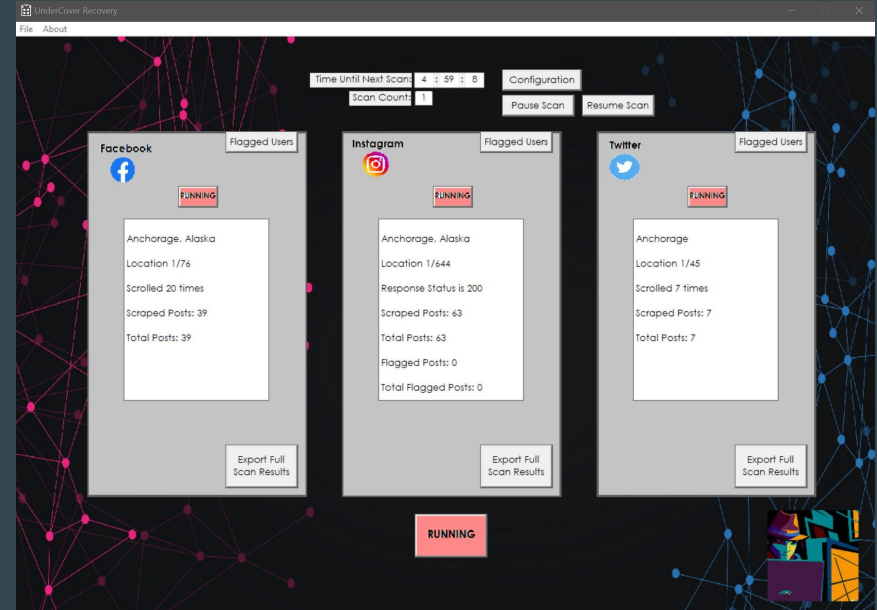
- The GUI displays a countdown between scans.
- The GUI relays the progress of the website scraping in real time.
- The results of the scan are saved and accessible as .html documents.

# GUI Design

Preliminary GUI



Finalized Product





# GUI - Configuration Window

Configuration

Scan Frequency: [05] h [00] m [00] s

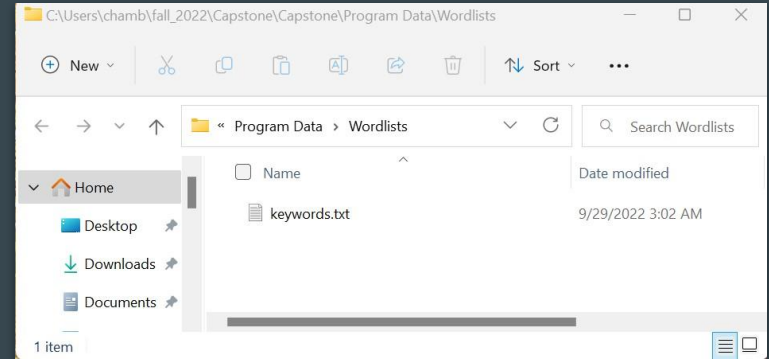
Set Scan Limit: [0]

Open Keywords

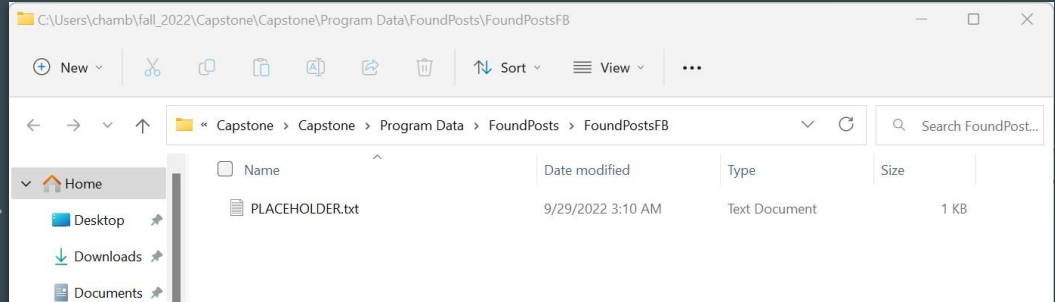
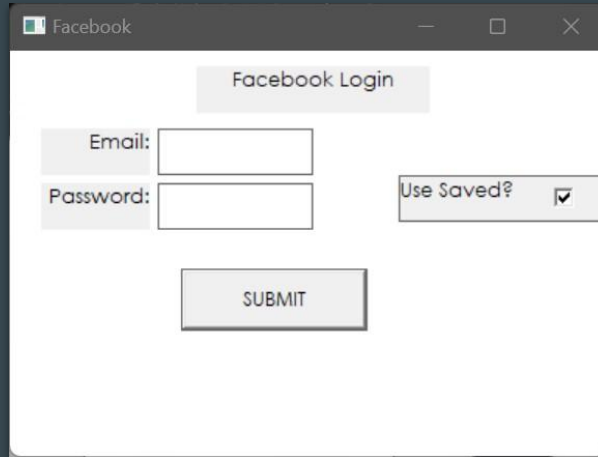
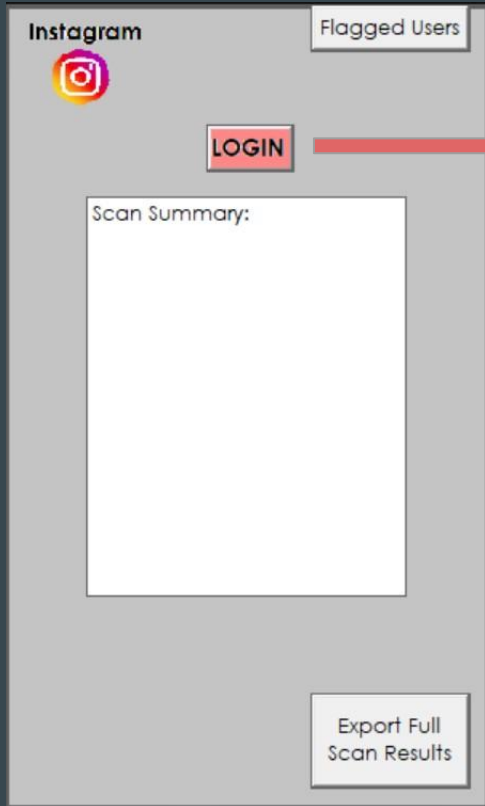
Select Region: Alaska

Output Folder: Default: .\Program Data\FoundPosts\

SUBMIT



# GUI - Login and Export



# Login Authenticator

- This python script authenticates user logins.
  - On success, will write username and password to a configuration file.
  - Saves acquired cookie value for future use by the respective scraper.
- The authenticator uses the Selenium library to access a webpage and submit a username and password.
- Twitter and Facebook require this due to content hashing, while Instagram requires it due to post request expiration.

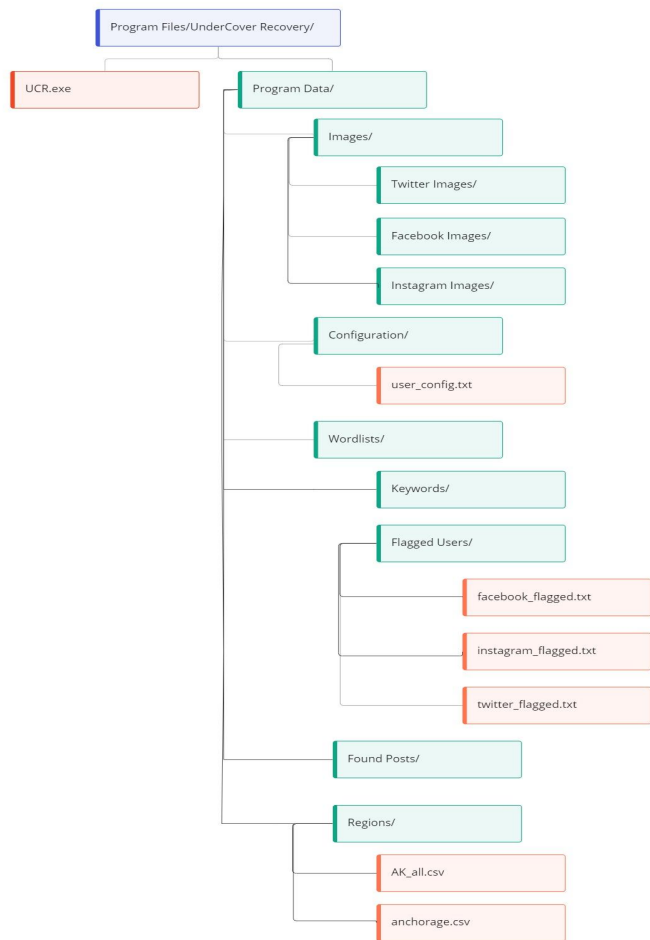
# Instagram Scraper

- This social media scraper, written in python, is the only one to utilize the Requests library with the cookie saved by the authenticator.
- Submits requests for precompiled location id's.
- JSON is returned containing information for all posts tagged to that location.
  - This mass of posts is parsed for keywords and flagged authors using BeautifulSoup.
- Yields the most detailed information of the three scrapers location due to lat/long coordinates included in each post.


# Facebook and Twitter Scrapers

- Web content is hashed in the html source code, cannot be accessed with the Requests library.
- Selenium uses a chrome webdriver to launch an automated browser.
- Location tags not as specific as Instagram, so tailored csv files were created with urls that filter posts by city and keyword.

# Saved Results



IG SCAN REPORT Dec 1 2022 @ 1.33.30

Date/Time	Lat/Long	Username	Full Name	Profile	Caption/Comment	Post	Media
10/11/2022 18:02:25	61.1900720386, -149.8269317766	uaa_coffee	☕ Your campus coffee shop ☕	<a href="#">link</a>	It's spooky season and we have a million themed drinks! I wish I had a cat hiding in my cup.:(	<a href="#">link</a>	

TW KEYWORDS REPORT Dec 1 2022 @ 0.53.43

Post Author	Timestamp	Caption	Post Link	Account Link
ABC News	2022-11-23T10:43:20.000Z	These firefighters got an unusual request for assistance from the Alaska Wildlife Troopers, but it wasn't your mundane cat-stuck-in-a-tree situation.	<a href="#">Tweet Link</a>	<a href="#">Account Link</a>
WCBD News 2	2022-11-23T13:33:03.000Z	Firefighters in Alaska got an unusual request for assistance last weekend from the Alaska Wildlife Troopers, but it wasn't your mundane cat-stuck-in-a-tree situation.	<a href="#">Tweet Link</a>	<a href="#">Account Link</a>
KirkBentonHomelandSecurity	2017-10-06T20:32:35.000Z	<a href="https://youtu.be/h_D3VFhys4">https://youtu.be/h_D3VFhys4</a> RIFF RAFF IS IN CAMP RIPLEY COMRADE EIELSON AFB ALASKA MINNESOTA HOLLYWOOD FAKE HOOD FILMED BY TOMMY CAT WILLIAMSON	<a href="#">Tweet Link</a>	<a href="#">Account Link</a>
John KG4AKV	2018-05-04T03:39:06.000Z	If you are working another full CAT station tuning both bands what you're doing might be fine. I wonder though if you risk walking on fixed uplink stations? If you want to work fixed uplink stations you might want to see if gpredict can be configured to just tune the downlink.	<a href="#">Tweet Link</a>	<a href="#">Account Link</a>

# Results Format

- Evidentiary copies of flagged posts are saved as html documents. This is an accepted format currently in use by the client. Posts with multiple images display the first and save the rest as hyperlinks.
- Images are download, hashed, then only saved if their hash value does not already exist to prevent duplication and save memory.
- Custom directories can be specified as a save location, allowing the use of external drives to address memory concerns.

# Planning and Schedule

- MVP was completed by November 1st, and included the preliminary GUI, authenticator script, and the Instagram scraping script.
- Facebook and Twitter scraping scripts were added afterwards, along with edits to the original methodology of authentication and scraping.
- The GUI redesign occurred in the last half of November.



# Challenges and Solutions

- Requests library and get requests for hashed html content led to the use of the Selenium library as an alternative. Eventually necessary for login authentication of the Instagram scraper.
- Environment setup for the project involved many errors, created difficulty in running application.
- Bitmap resolution defined by Win32, unable to increase image resolutions.
- Within the last week of project development, Instagram changed their permissions for requests, limiting our scraper.

# Defects Log

- Total of 43 defects noted by classmates that have been addressed either by justification or implementation.
- Sampling of most severe defects:

Severity (H M L Q)	Location (line #)	Description	Status	Developer Response
M-H	199,239,258	add more variations? or add in common misspellings	Resolved	Common misspellings is left to the user to define in the keywords file, this would vary for every term and is domain specific. As for the variations, you mentioned the case where the caption is just a keyword without spaces or punctuation, this was our initial implementation but it resulted in many false positives (i.e. keyword 'cat' pulled down posts with 'vacation' in the caption), and to limit review time the client decided to err on the side of caution and avoid false positives at the expense of missing some posts.
H	201	Poor variable name 'word'; already in use on ln198	Resolved	Great catch, thanks! Changed second name to 'w', less descriptive but doesn't compete.
H	53 - 56	Relies on csvs in region resolution table to be formed in a specific way without checks (header, last value is location url)	Resolved	CSV file is prefabricated and provided as part of installation.

# DEMO

## Q/A