Detecting Phishing using Visual Similarity

erin chack
@ErinChack

ME: i'm only afraid of two things: public speaking and ghosts
[later, on stage]
CROWD: BOOOOOOOO
ME: oh no

Josh Pyorre, Security Researcher

November 22, 2024

TALOS

Previously:

zscaler

CISCO Umbrella

OpenDNS

MANDIANT

NASA

Detecting Phishing using Visual Similarity

Outline

- Current Tools & Techniques
- Web Crawling
- Creating Datasets
- Distance & Similarity
- Grouping Images
- Testing Detection
- Beyond Images
    - HTML Similarity
    - Scraping Text from Images
    - Text Classification with LLMs
- Research
- Alerting
- Action

If you see text like this, it's a note for this pdf version of the slides so you don't have to guess what's going on.
If you see my face next to the text, it's just me "saying" the note so you can easily differentiate from other text on a busy slide.

Current Tools & Techniques

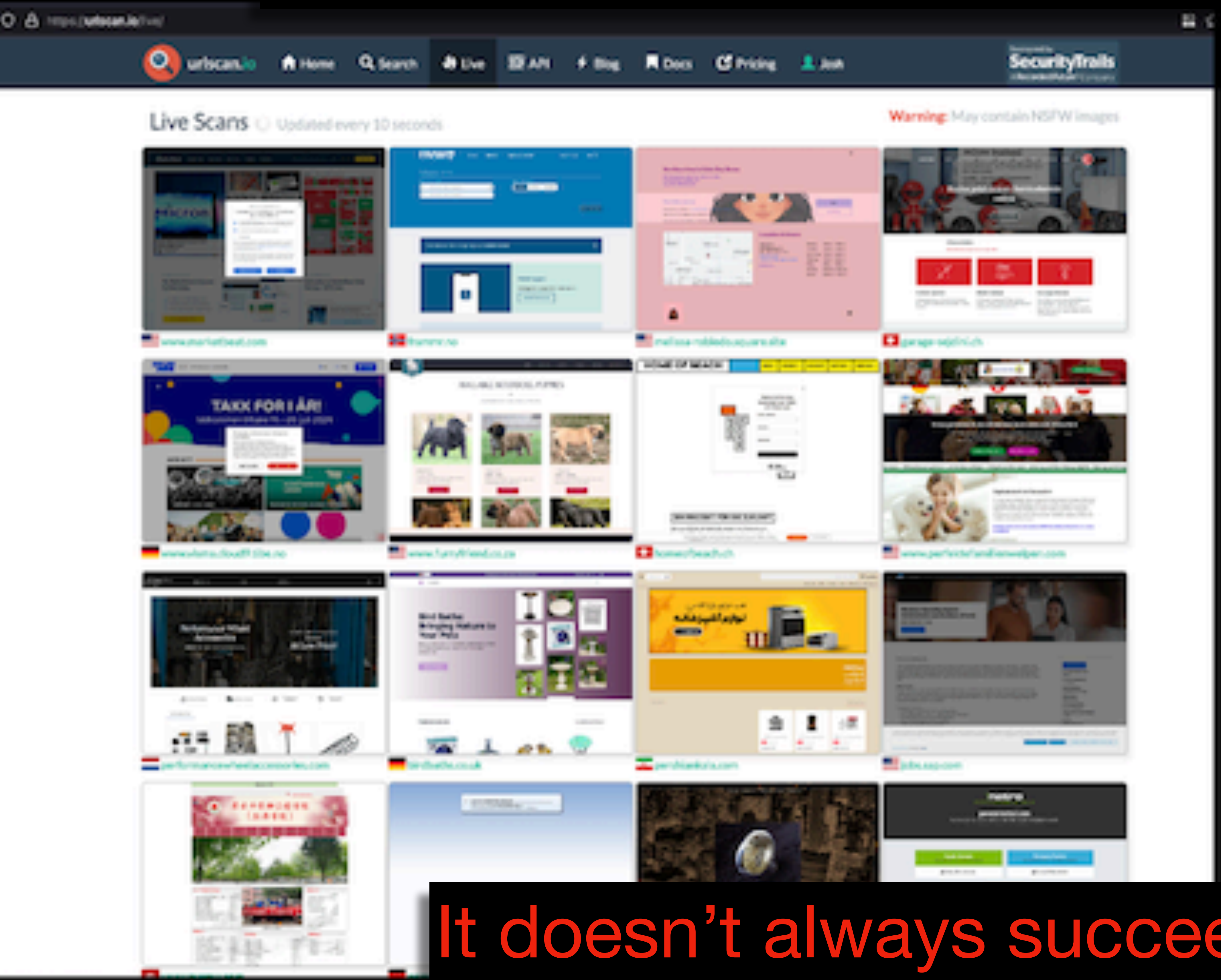
Services

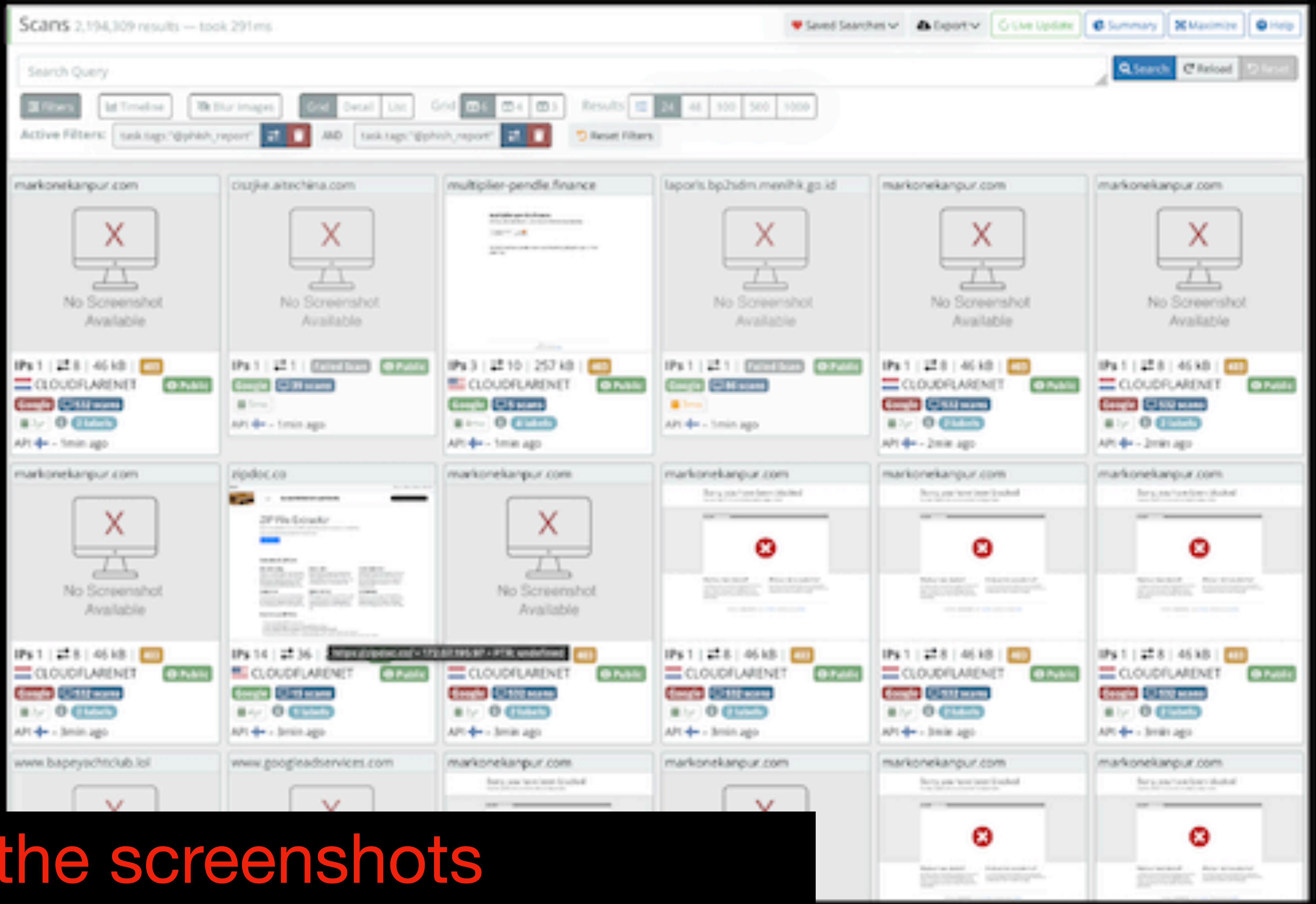
URLScan.io

Crawling

Threat detection

URLScan takes screenshots and provides information on websites.
There is a YARA search that can be setup for certain kinds of website behavior





It doesn’t always succeed at the screenshots
And it doesn’t do extensive analysis on the screenshots

urlscan.io Phishing URL Feed

urlscan.io detects thousands of suspicious, malicious and phishing URLs every day. Our organic phishing URL detection is able to pinpoint these attacks and associate them with one of the 680+ popular brands that we track. The feed of these detected URLs is available to our customers to ingest

- IP address & GeoIP information of the IP hosting the phishing page
- ASN and ASN Name hosting the phishing URL

You can download a static 7-day sample of the feed here:

Phishing URL Feed

> 1500 detected URLs / day
> 680+ brands tracked
✓ Live URL retrieval
✓ Commercial use
✓ JSON & CSV format

Targeted Brand Information
Brand name
Brand country & industry vertical

Phishing Page Information
Phishing page metadata
Phishing domain metadata
Phishing IP metadata

URLScan has a phishing feed that’s useful. In our line of work, we all likely ingest feed from various locations as a quick way to use the intelligence and work of others to improve our security. These feeds can also be used to gather data about the current threat landscape.

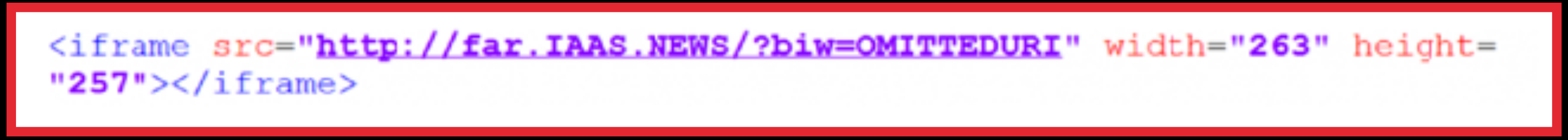
Current Tools & Techniques

Detection Methods

Threat hunting
User reporting
Various Products

Threat Hunting

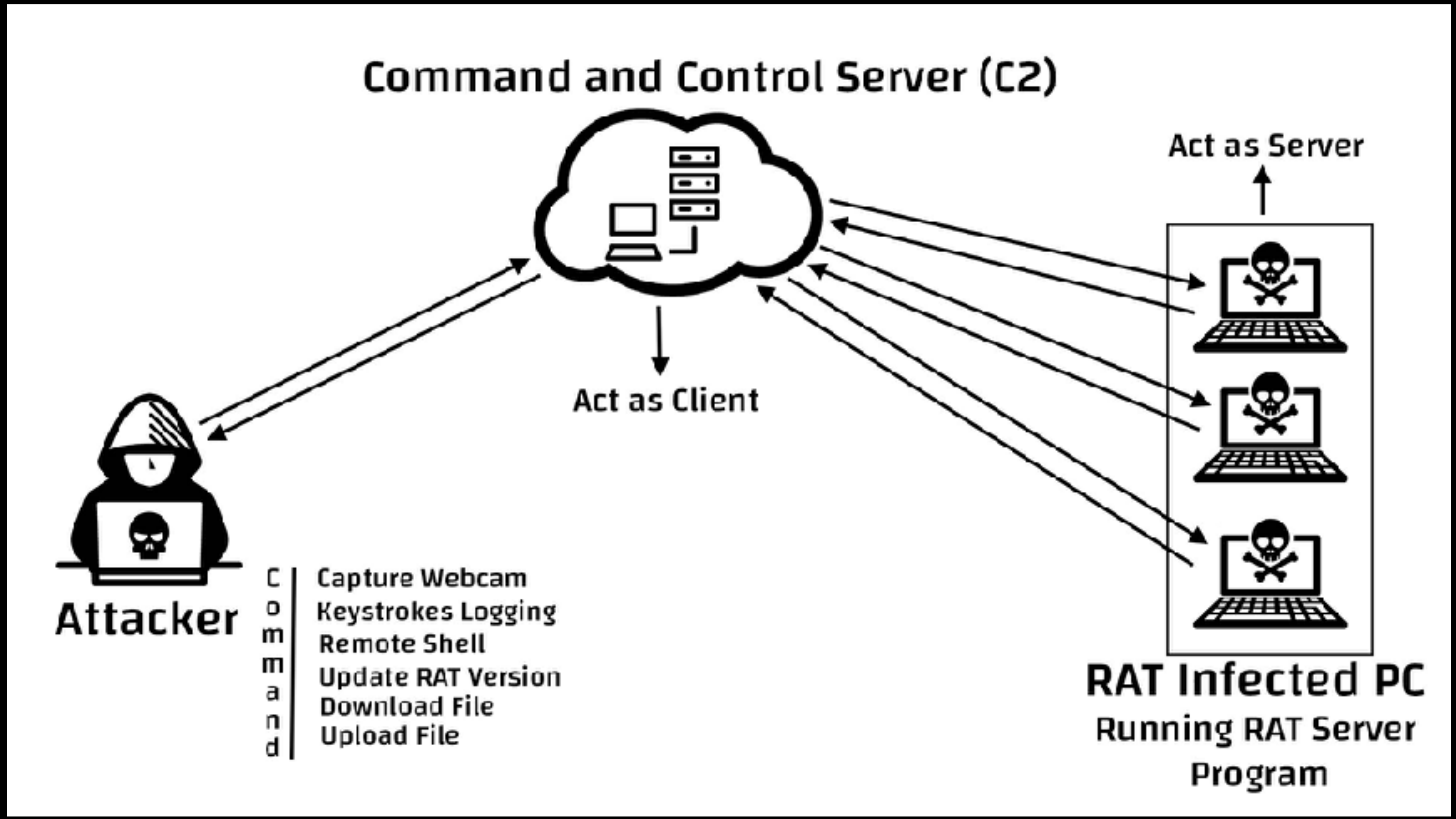
Compromised Infrastructure



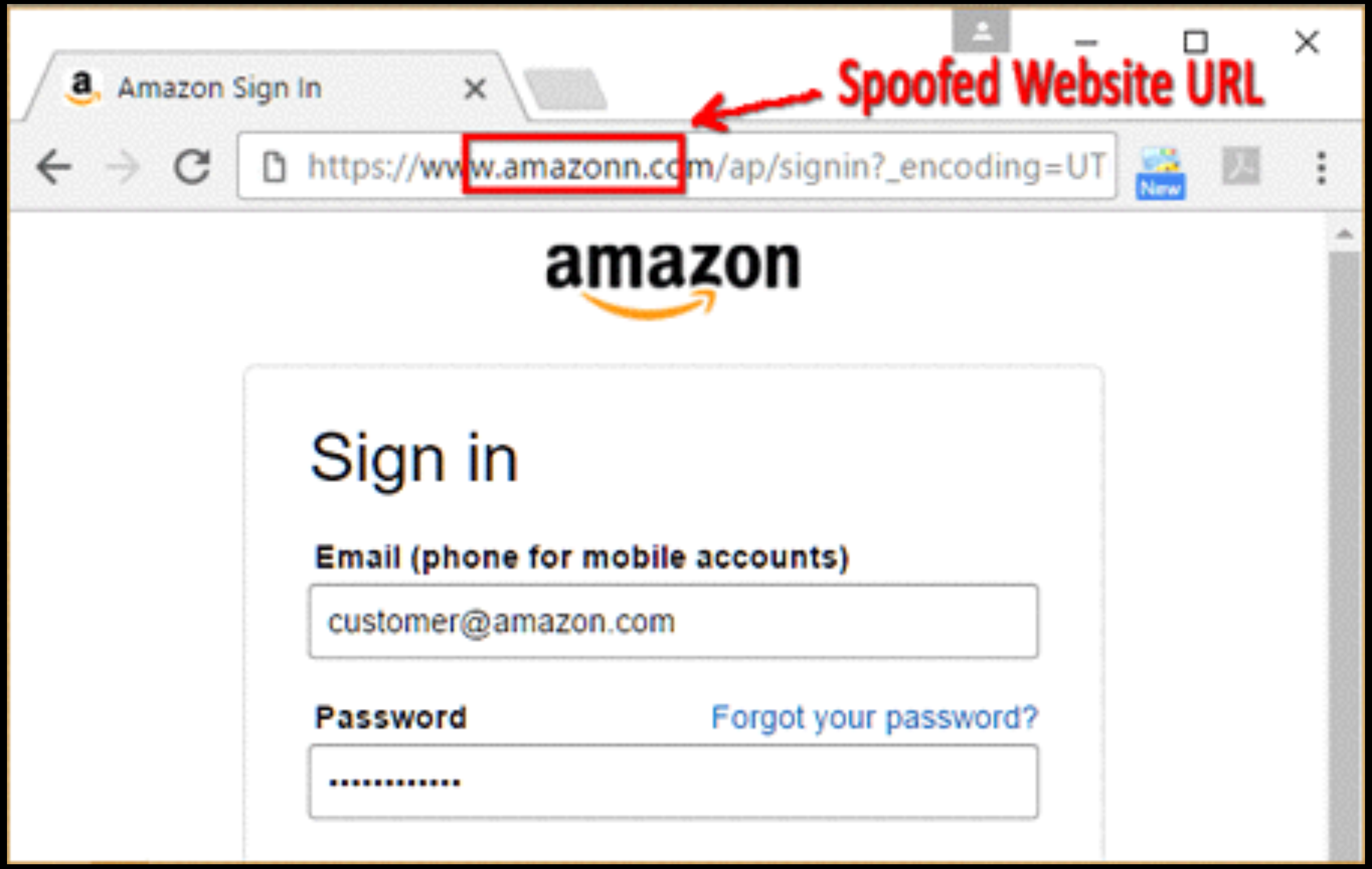
```
<iframe src="http://far.IAAS.NEWS/?biw=OMITTEDURI" width="263" height=
"257"></iframe>
```

```
hXXp://www.fullcircleliterary.com/
hXXp://danielpsheehan.com/areas-of-
expertise/educator/ucsc-2016-rulers-
of-the-realm
hXXp://danielpsheehan.com/
hXXp://www.cafemuseroyaloak.com/
hXXp://kdsross.com/about-us/
hXXp://usdiagnostics.com/index.php/
certification-testing/uscreen-cup
hXXp://psychologywiththal.com/
2015/09/30/life-span-development-
personality/
hXXp://
thefecaltransplantfoundation.org/
what-is-fecal-transplant/
hXXp://optimalwellnessaz.com/about/
hXXp://optimalwellnessaz.com/about/
hXXp://chworks.org/real-estate-
```

C2



Phishing



https://www.amazonn.com/ap/signin?_encoding=UT

Sign in

Email (phone for mobile accounts)

customer@amazon.com

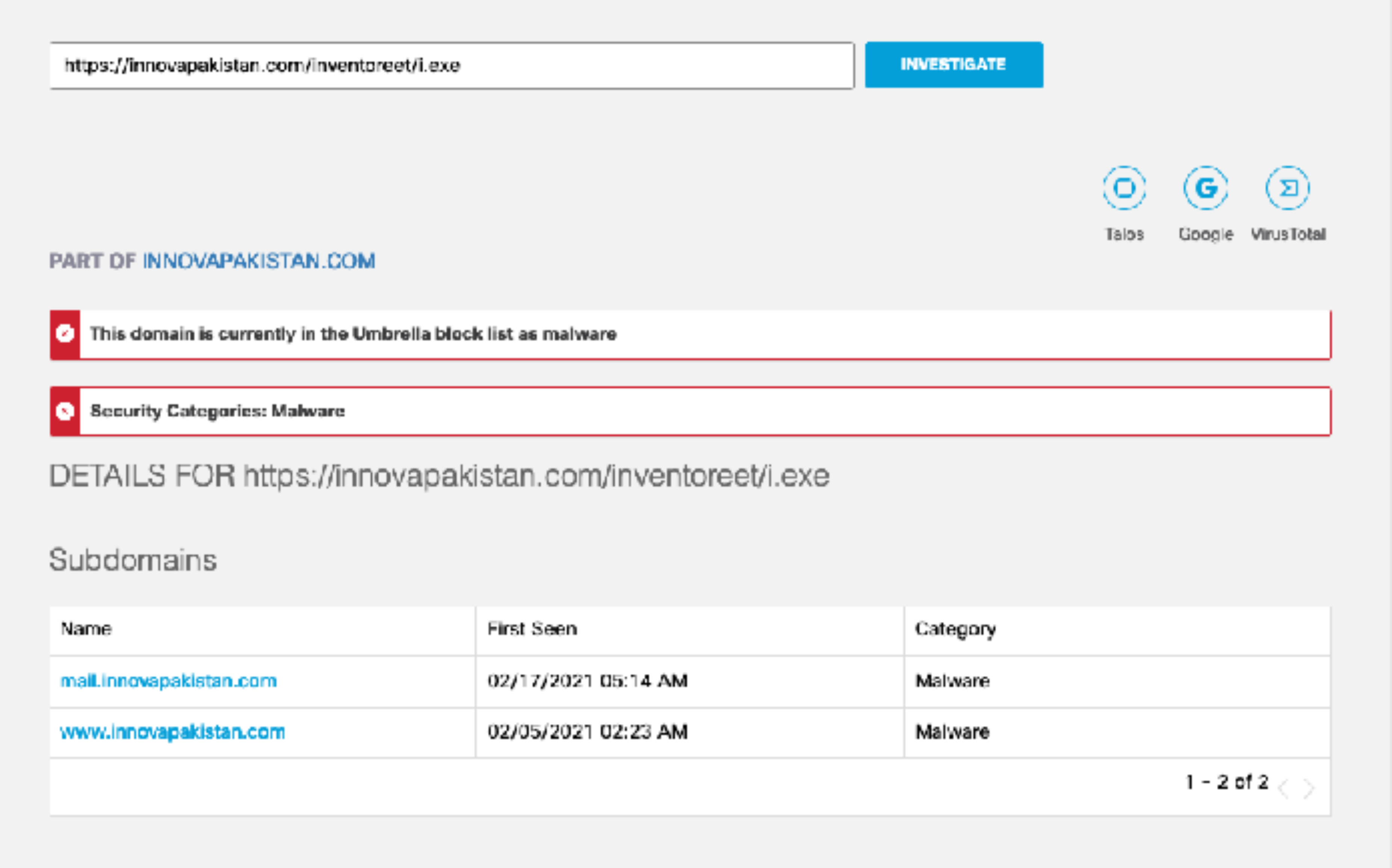
Password

Forgot your password?

••••••••••••

Threat Hunting cont...

DNS



https://innovapakistan.com/inventoreet/i.exe

PART OF INNOVAPAKISTAN.COM

This domain is currently in the Umbrella block list as malware

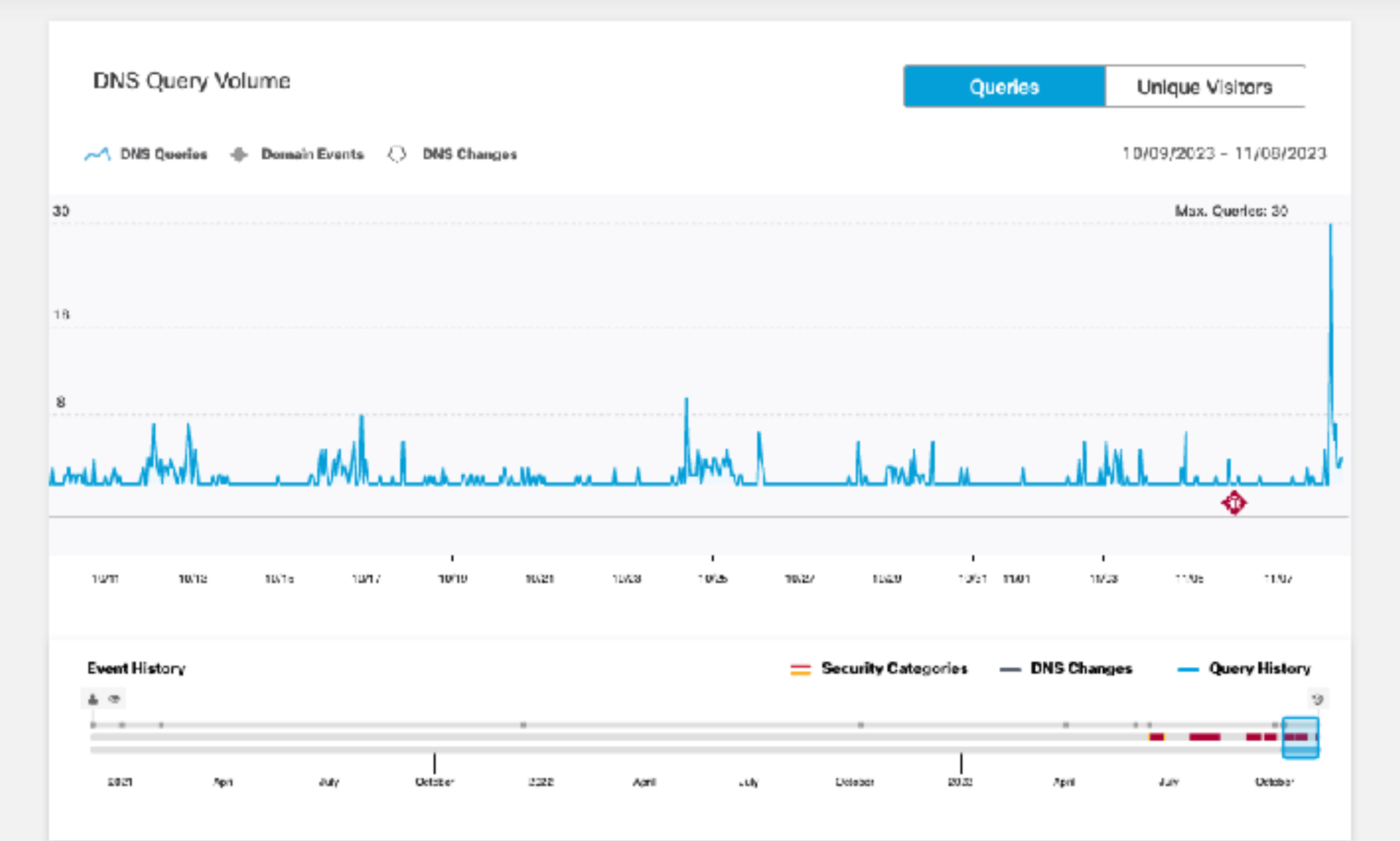
Security Categories: Malware

DETAILS FOR https://innovapakistan.com/inventoreet/i.exe

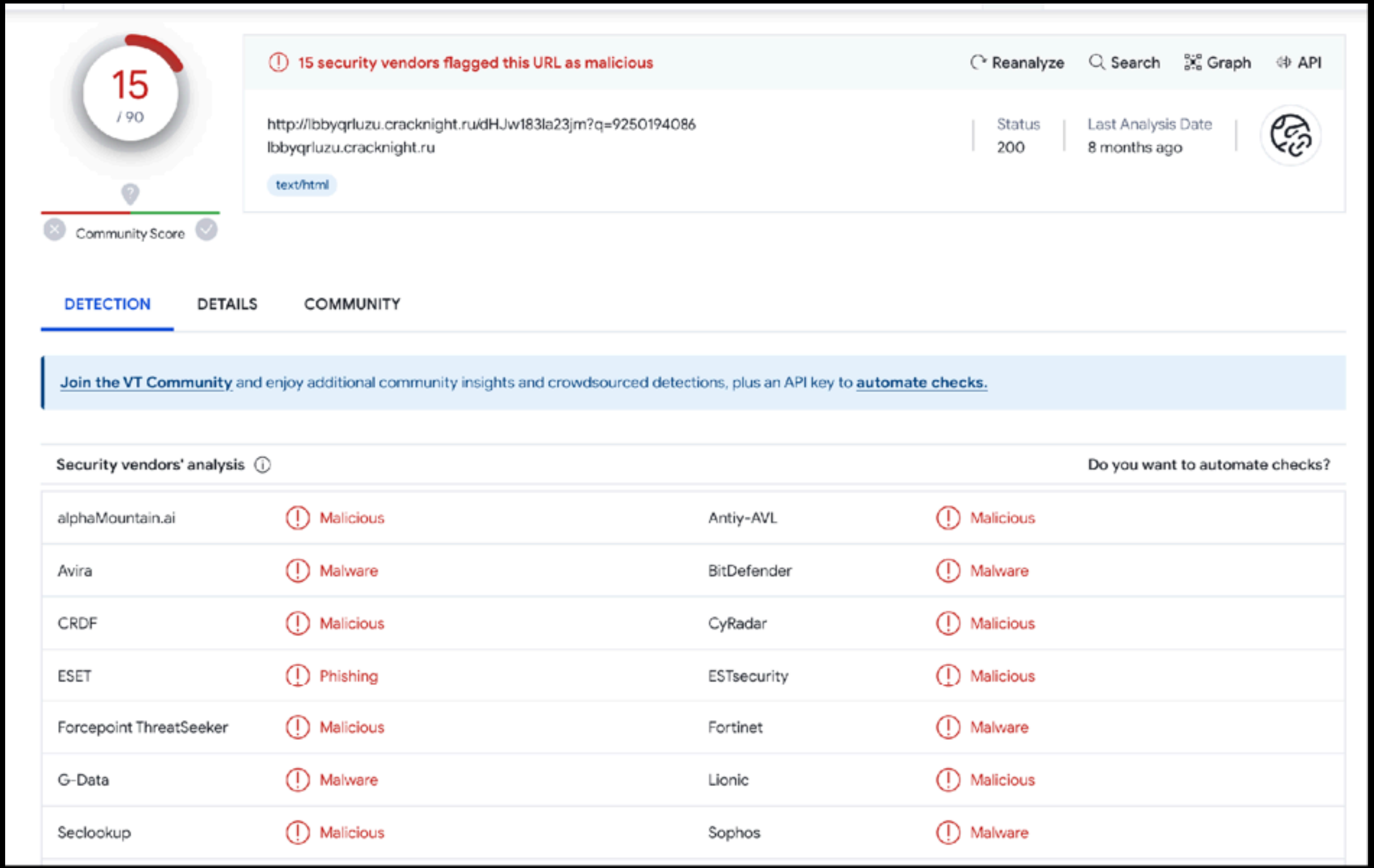
Subdomains

Name	First Seen	Category
mail.innovapakistan.com	02/17/2021 05:14 AM	Malware
www.innovapakistan.com	02/05/2021 02:23 AM	Malware

1 - 2 of 2



URL Analysis



15 / 90

15 security vendors flagged this URL as malicious

http://lbbyqrluzu.cracknight.ru/dHJw183la23jm?q=9250194086
lbbyqrluzu.cracknight.ru
text/html

Status 200

Last Analysis Date 8 months ago

Join the VT Community and enjoy additional community insights and crowdsourced detections, plus an API key to automate checks.

Security vendors' analysis

Do you want to automate checks?

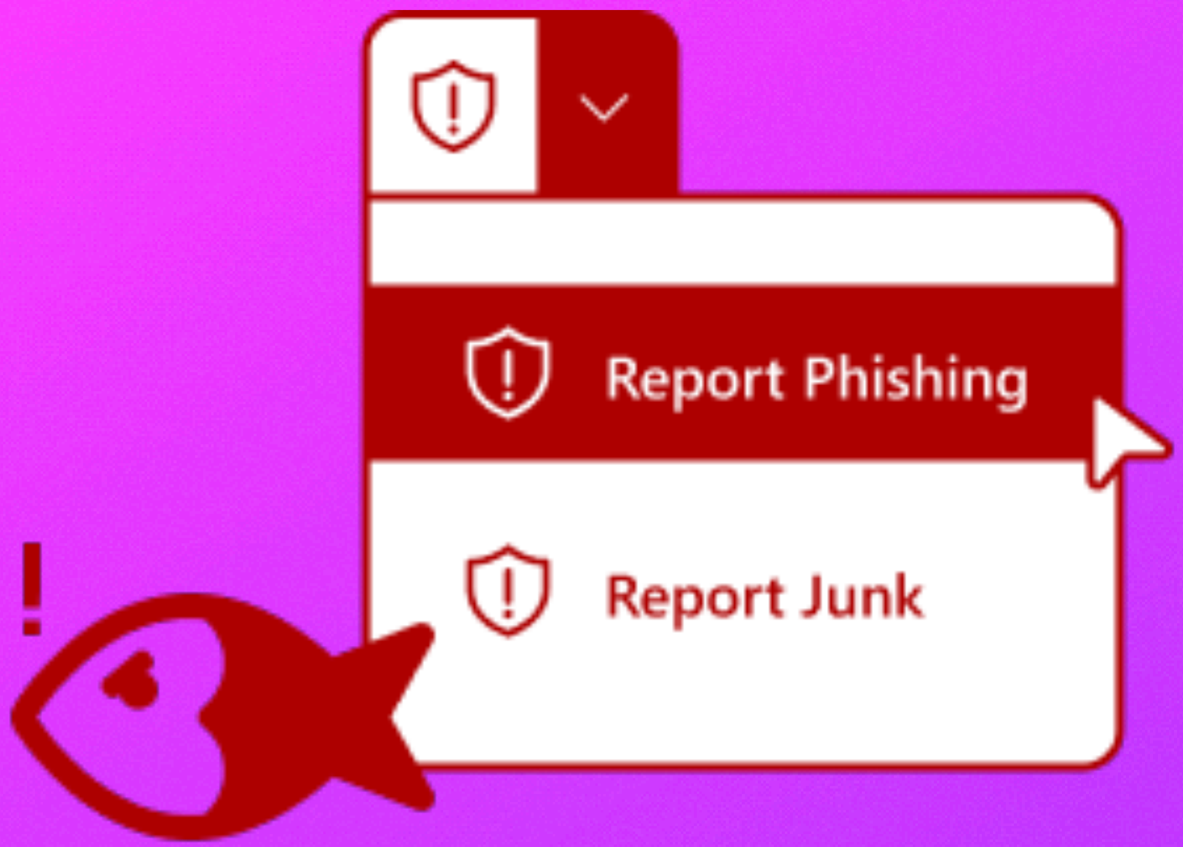
alphaMountain.ai	Malicious	Antiy-AVL	Malicious
Avira	Malware	BitDefender	Malware
CRDF	Malicious	CyRadar	Malicious
ESET	Phishing	ESTsecurity	Malicious
Forcepoint ThreatSeeker	Malicious	Fortinet	Malware
G-Data	Malware	Lionic	Malicious
Seclookup	Malicious	Sophos	Malware

Current Tools & Techniques

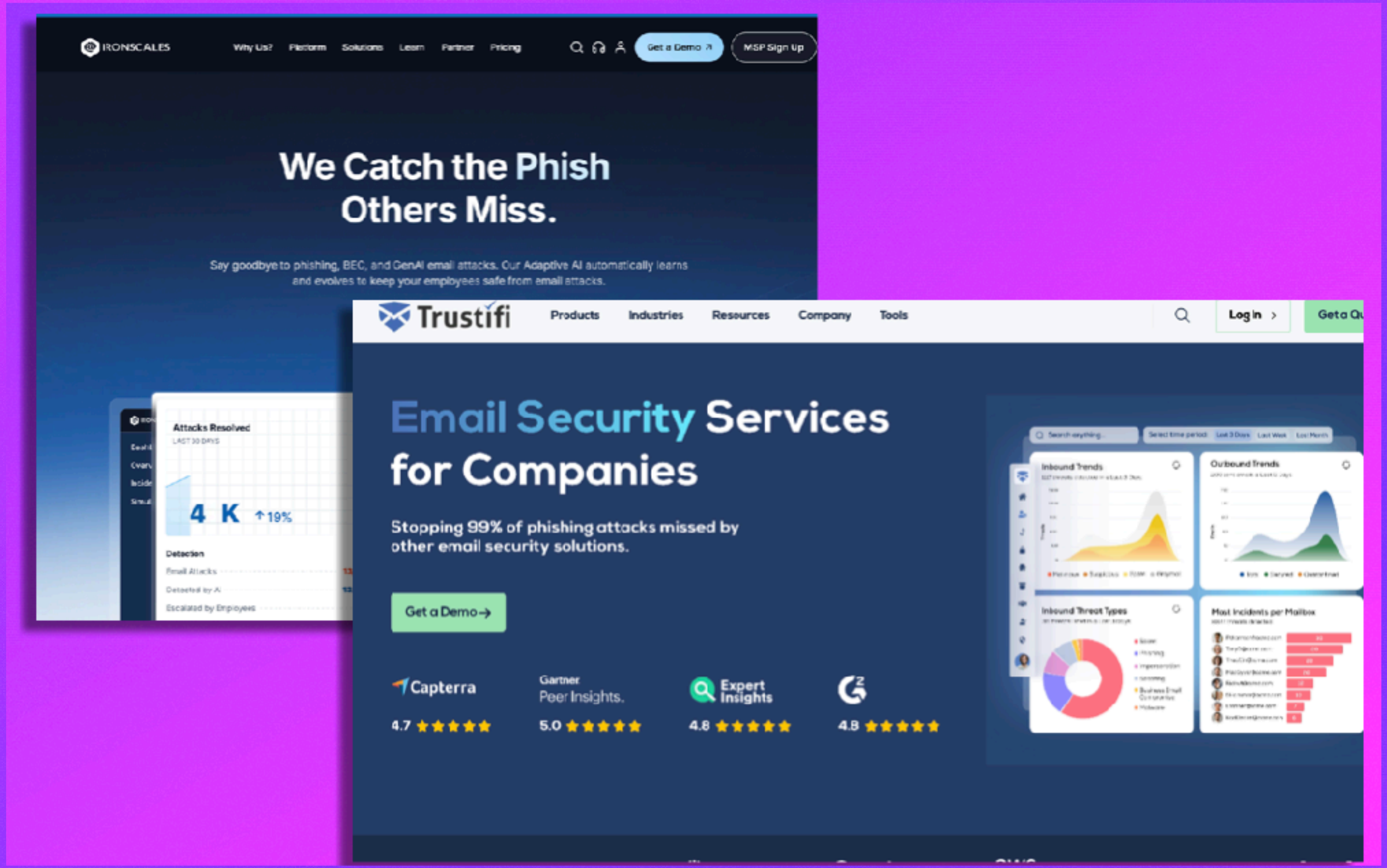
User reporting is another source of malicious activity detection

And there are multiple products you can pay for to increase detection

User Reporting



Products



We Catch the Phish Others Miss.

Say goodbye to phishing, BEC, and GenAI email attacks. Our Adaptive AI automatically learns and evolves to keep your employees safe from email attacks.

Email Security Services for Companies

Stopping 99% of phishing attacks missed by other email security solutions.

Security Is Made of Layers

The work I’m demonstrating are part of a layered approach.

There’s always room for more layers to improve security

What do we have to do?

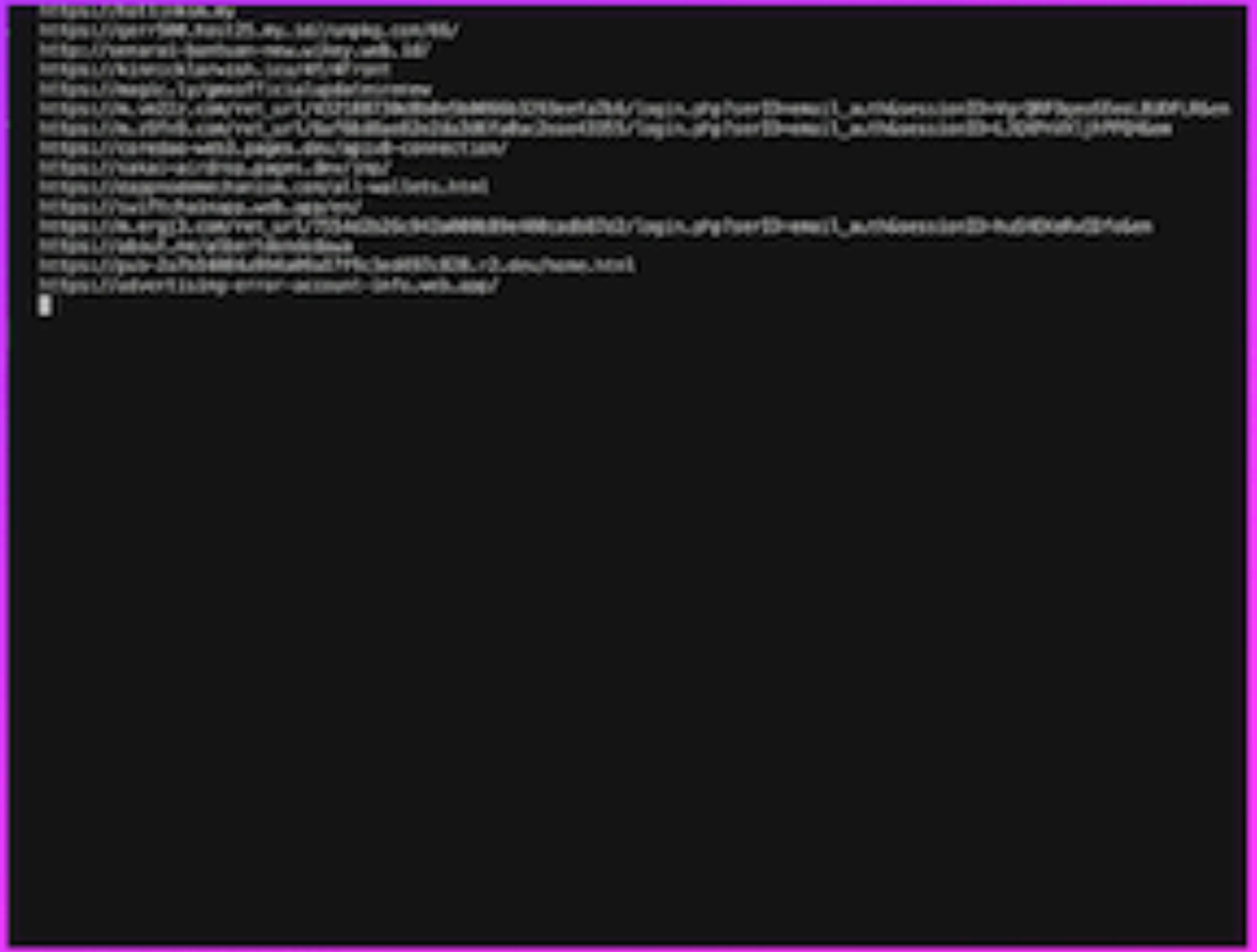
Crawl Websites
Take Screenshots
Build Datasets
Find Similar in…

- Images
- HTML
- Text
- Infrastructure

Web Crawling

Crawler Architecture

Send URLs

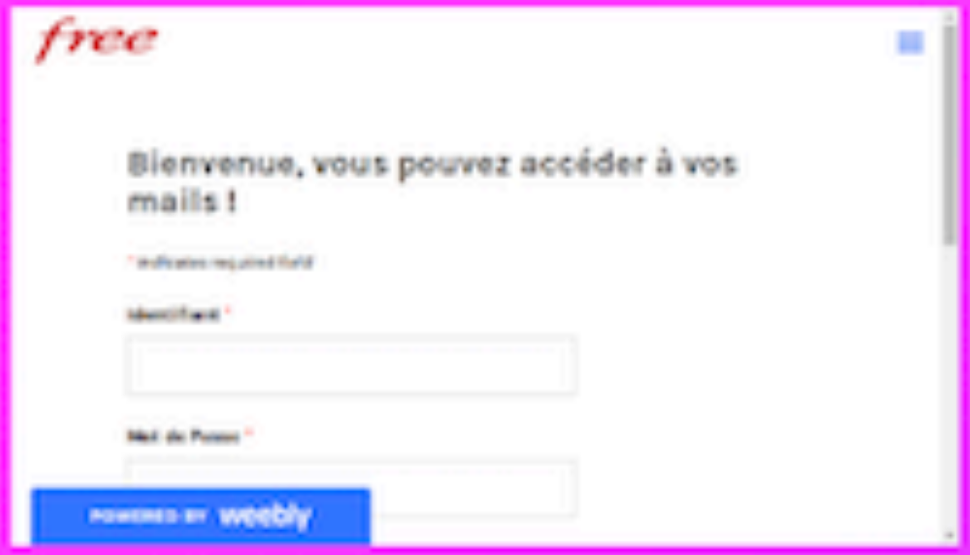


to Selenium

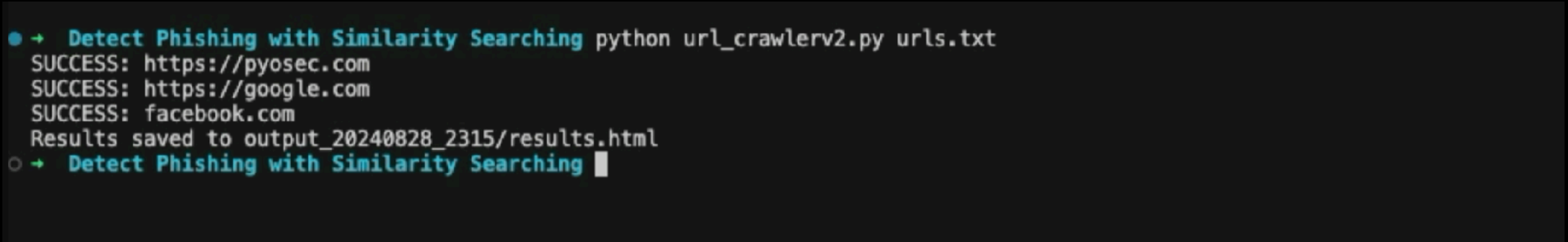
Selenium

Selenium is an umbrella project encapsulating a variety of tools and libraries enabling web browser automation. Selenium specifically provides an infrastructure for the W3C WebDriver specification — a platform and language-neutral coding interface compatible with all major web browsers.

and take screenshots

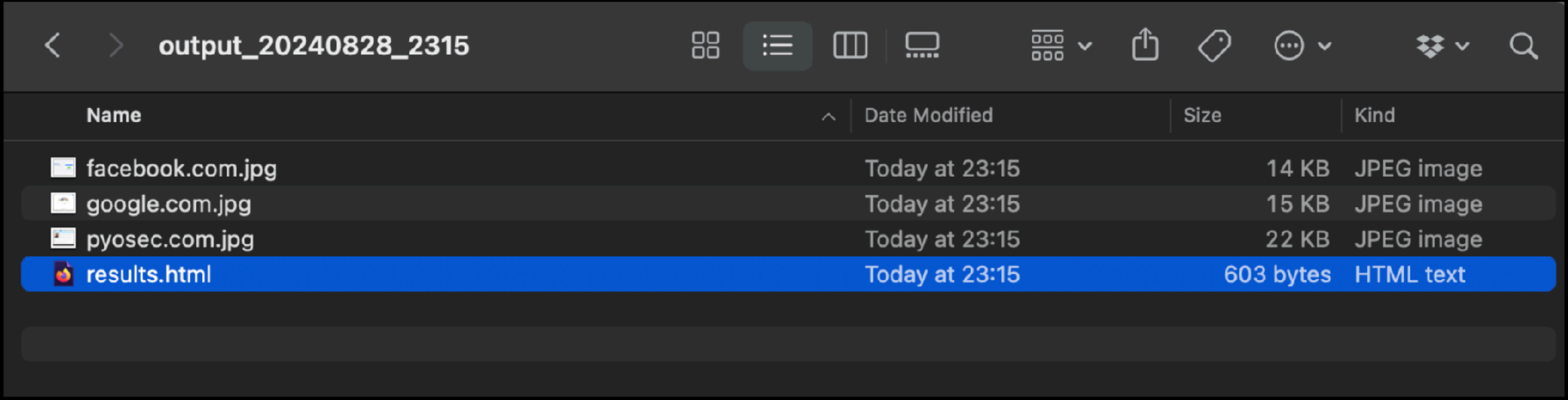


Bienvenue, vous pouvez accéder à vos mails !



```
● → Detect Phishing with Similarity Searching python url_crawlerv2.py urls.txt
  SUCCESS: https://pyosec.com
  SUCCESS: https://google.com
  SUCCESS: facebook.com
  Results saved to output_20240828_2315/results.html
○ → Detect Phishing with Similarity Searching 
```

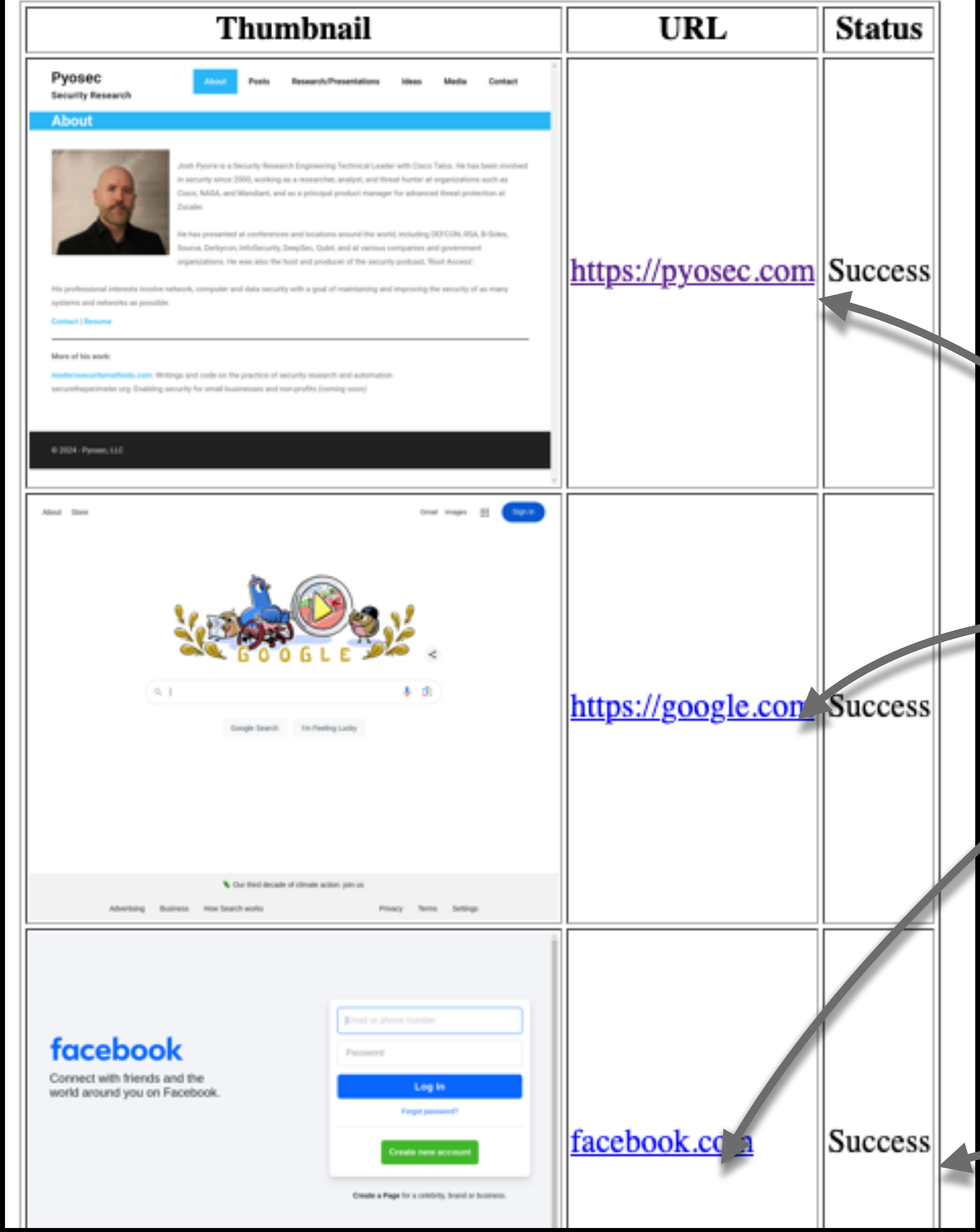
Demo of three websites being crawled, after which, I am left with the screenshots and a results.html file



output_20240828_2315

Name	Date Modified	Size	Kind
facebook.com.jpg	Today at 23:15	14 KB	JPEG image
google.com.jpg	Today at 23:15	15 KB	JPEG image
pyosec.com.jpg	Today at 23:15	22 KB	JPEG image
results.html	Today at 23:15	603 bytes	HTML text

Screenshots

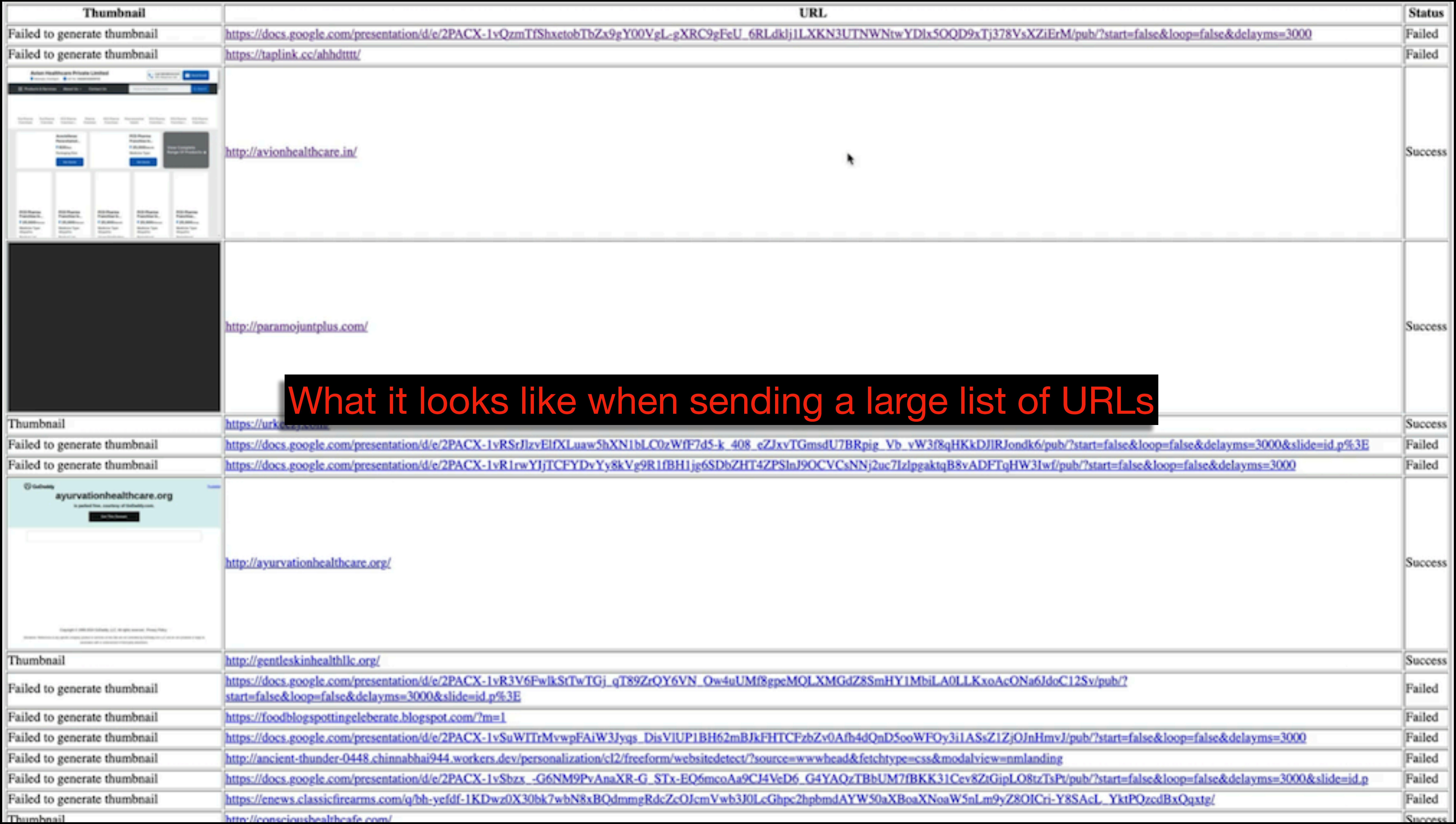


Thumbnail	URL	Status
	https://pyosec.com	Success
	https://google.com	Success
	facebook.com	Success

URLs

Status

This is the start of collecting crawling data, but it will increase/improve through the presentation



Thumbnail	URL	Status
Failed to generate thumbnail	https://docs.google.com/presentation/d/e/2PACX-1vQzmTfShxetobTbZx9gY00VgL-gXRC9gFeU_6RLdkJj1LXKN3UTNWNtwYDlx5OQD9xTj378VsXZiErM/pub/?start=false&loop=false&delayms=3000	Failed
Failed to generate thumbnail	https://taplink.cc/ahhdtttt/	Failed
	http://avionhealthcare.in/	Success
	http://paramojuntplus.com/	Success
Thumbnail	https://urk	Success
Failed to generate thumbnail	https://docs.google.com/presentation/d/e/2PACX-1vRSrJlzvElfXLuaw5hXN1bLC0zWfF7d5-k_408_eZJxvTGmsdU7BRpig_Vb_vW3f8qHKkDJlRJondk6/pub/?start=false&loop=false&delayms=3000&slide=id.p%3E	Failed
Failed to generate thumbnail	https://docs.google.com/presentation/d/e/2PACX-1vR1rwYJjTCFYDvYy8kVg9R1fBH1jg6SDbZHT4ZPSlnJ9OCVCsNNj2uc7IzlpgaktqB8vADFTqHW31wf/pub/?start=false&loop=false&delayms=3000	Failed
	http://ayurvationhealthcare.org/	Success
Thumbnail	http://gentleskinhealthllc.org/	Success
Failed to generate thumbnail	https://docs.google.com/presentation/d/e/2PACX-1vR3V6FwlkStTwTGj_qT89ZrQY6VN_Ow4uUMf8gpeMQLXMGdZ8SmHY1MbiLA0LLKxoAcONa6JdoC12Sv/pub/?start=false&loop=false&delayms=3000&slide=id.p%3E	Failed
Failed to generate thumbnail	https://foodblogspottingeleberate.blogspot.com/?m=1	Failed
Failed to generate thumbnail	https://docs.google.com/presentation/d/e/2PACX-1vSuWITrMvwpFAiW3Jyqs_DisVlUP1BH62mBJkFHTCFzbZv0Afh4dQnD5ooWFOy3i1ASsZ1ZjOJnHmvJ/pub/?start=false&loop=false&delayms=3000	Failed
Failed to generate thumbnail	http://ancient-thunder-0448.chinnabhai944.workers.dev/personalization/cl2/freeform/websitedetect/?source=wwwhead&fetchtype=css&modalview=nmlanding	Failed
Failed to generate thumbnail	https://docs.google.com/presentation/d/e/2PACX-1vSbzx_-G6NM9PvAnaXR-G_STx-EQ6mcoAa9CJ4VeD6_G4YAQzTBbUM7fBKK31Cev8ZtGipLO8tzTsPt/pub/?start=false&loop=false&delayms=3000&slide=id.p	Failed
Failed to generate thumbnail	https://enews.classicfirearms.com/q/bh-yefdf-1KDwz0X30bk7wbN8xBQdmmgRdcZcOJcmVwb3J0LcGhpc2hpbmdAYW50aXBoaXNoaW5nLm9yZ8OICri-Y8SAcL_YktPQzcdBxQgxtg/	Failed
Thumbnail	http://conscioushealthcafe.com/	Success

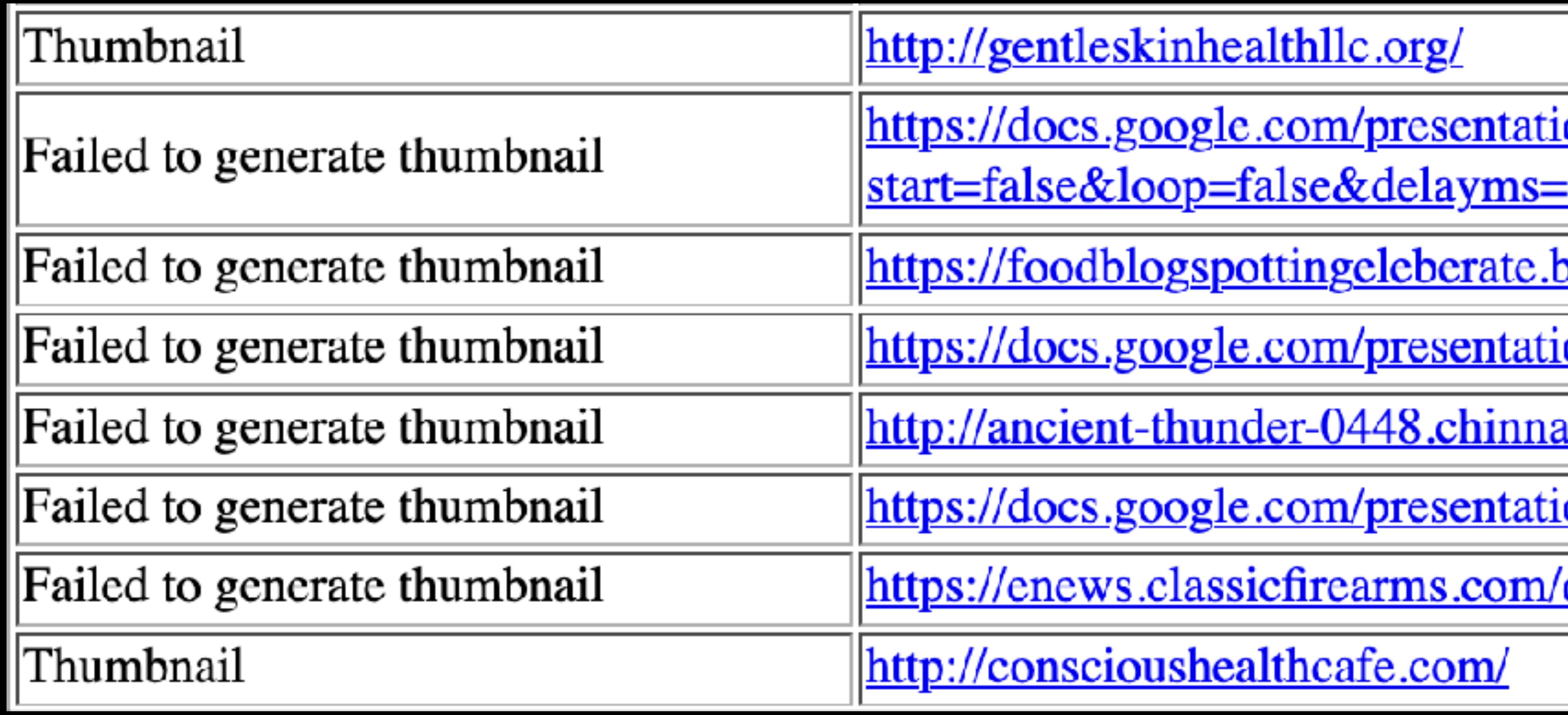
What it looks like when sending a large list of URLs

Issues/Obstacles when Crawling

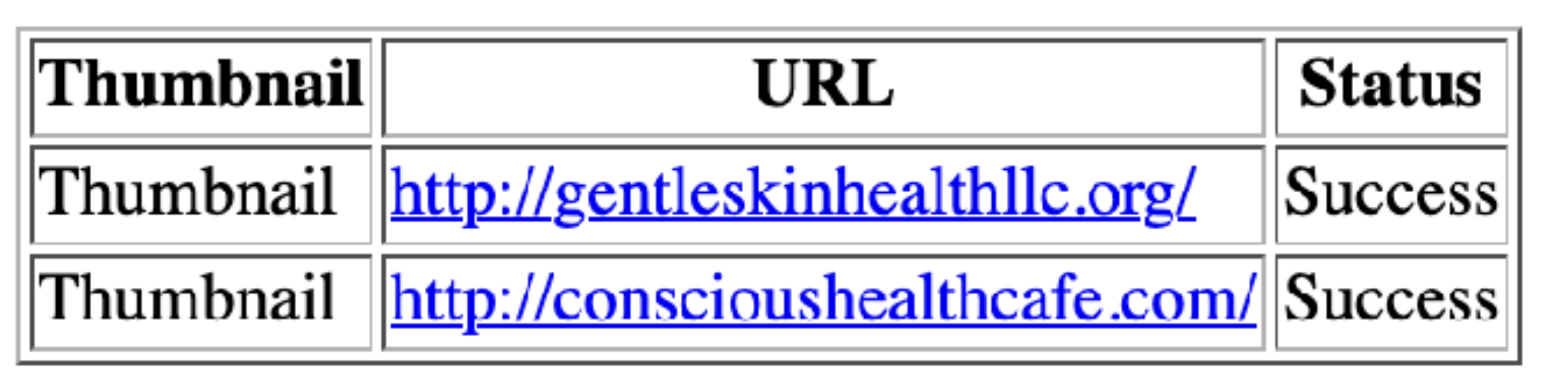
FAIL

Crawl errors

Failed to generate thumbnail errors…



Thumbnail	http://gentleskinhealthllc.org/
Failed to generate thumbnail	https://docs.google.com/presentati start=false&loop=false&delayms=
Failed to generate thumbnail	https://foodblogspottingcelebrate.b
Failed to generate thumbnail	https://docs.google.com/presentati
Failed to generate thumbnail	http://ancient-thunder-0448.chinna
Failed to generate thumbnail	https://docs.google.com/presentati
Failed to generate thumbnail	https://enews.classicfirearms.com/
Thumbnail	http://conscioushealthcafe.com/

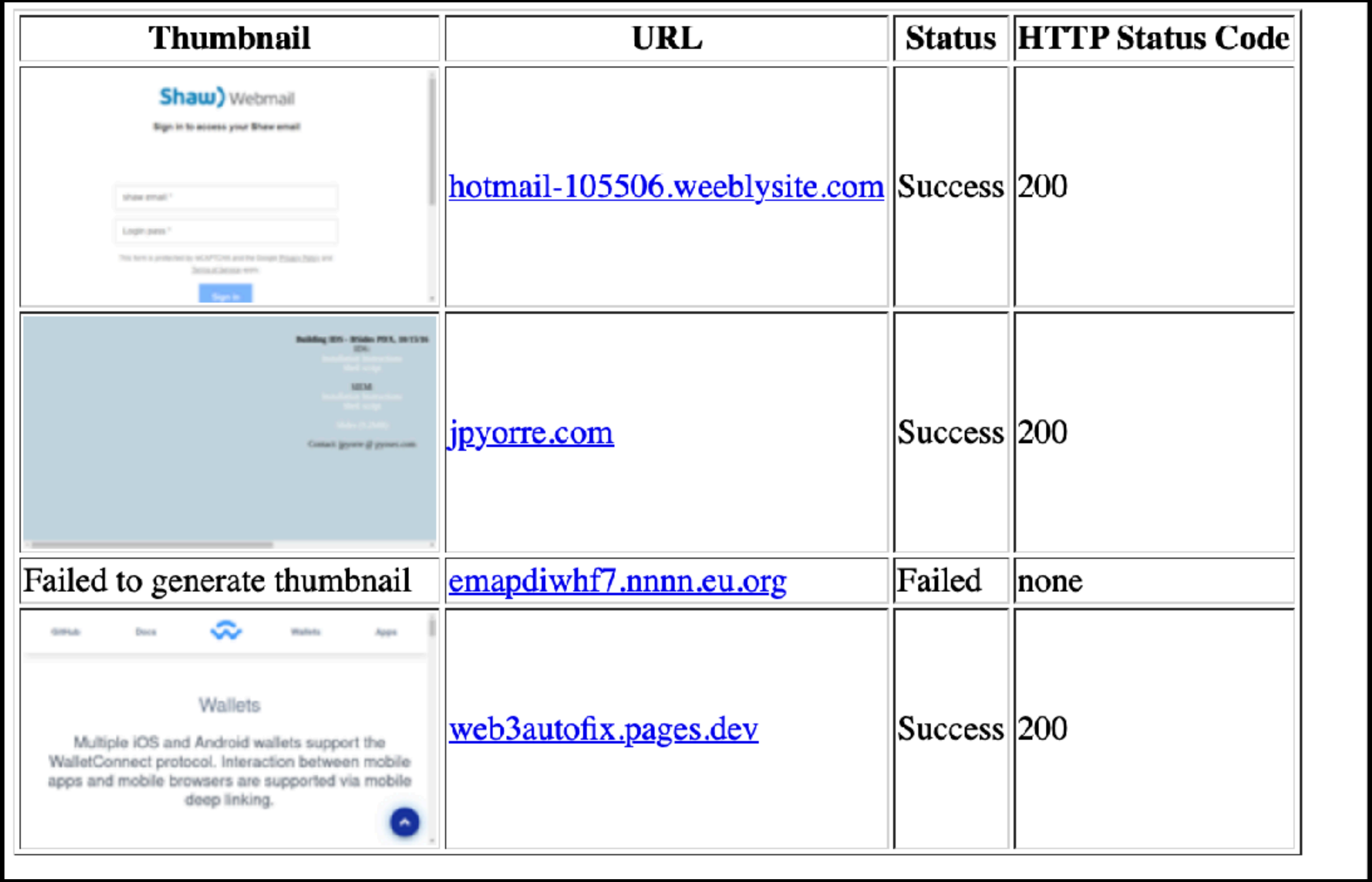


Thumbnail	URL	Status
Thumbnail	http://gentleskinhealthllc.org/	Success
Thumbnail	http://conscioushealthcafe.com/	Success

Was able to crawl, but no screenshot generated…

The crawlers run headless selenium and can be modified easily

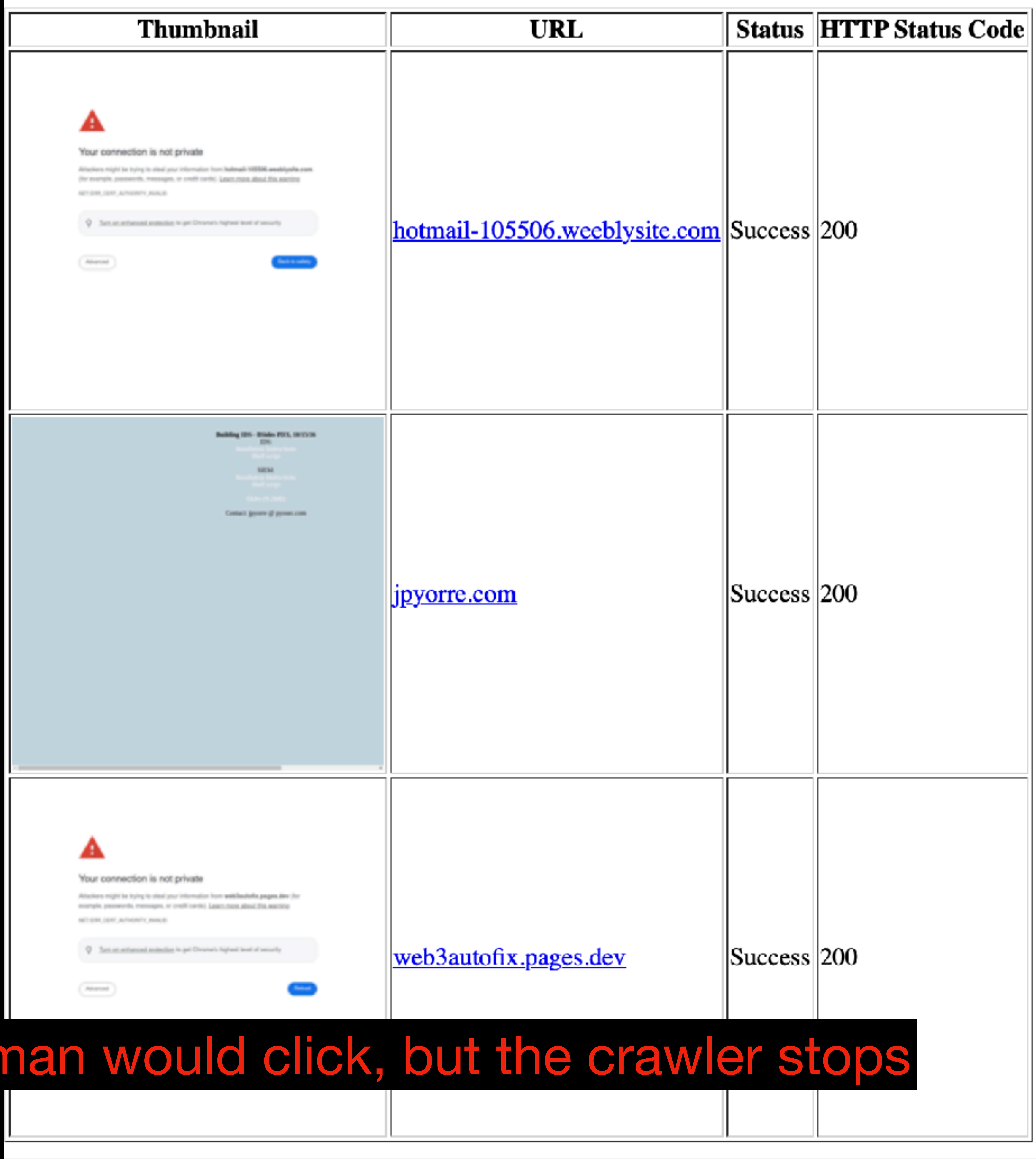
Fixed with more info



Thumbnail	URL	Status	HTTP Status Code
	hotmail-105506.weeblysite.com	Success	200
	jpyorre.com	Success	200
Failed to generate thumbnail	emapdiwhf7.nnnn.eu.org	Failed	none
	web3autofix.pages.dev	Success	200

FAIL

Stopped at Error

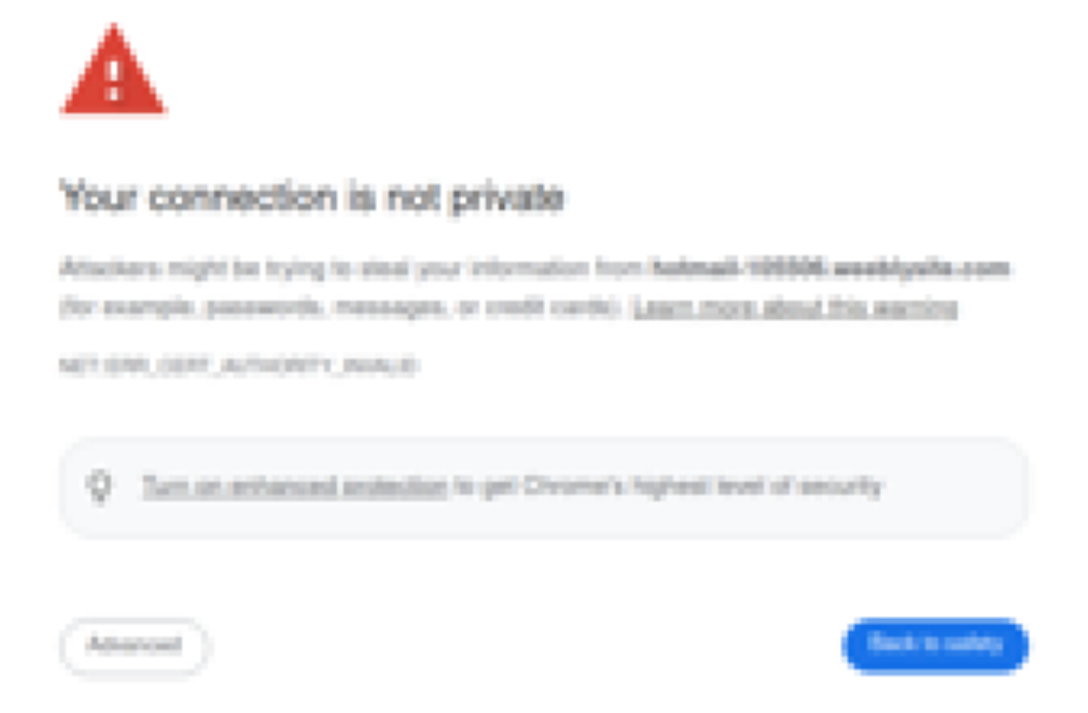


Thumbnail	URL	Status	HTTP Status Code
Your connection is not private	hotmail-105506.weeblysite.com	Success	200
	jpyorre.com	Success	200
Your connection is not private	web3autofix.pages.dev	Success	200

Other page crawling errors. A human would click, but the crawler stops

Use Selenium

Read Text

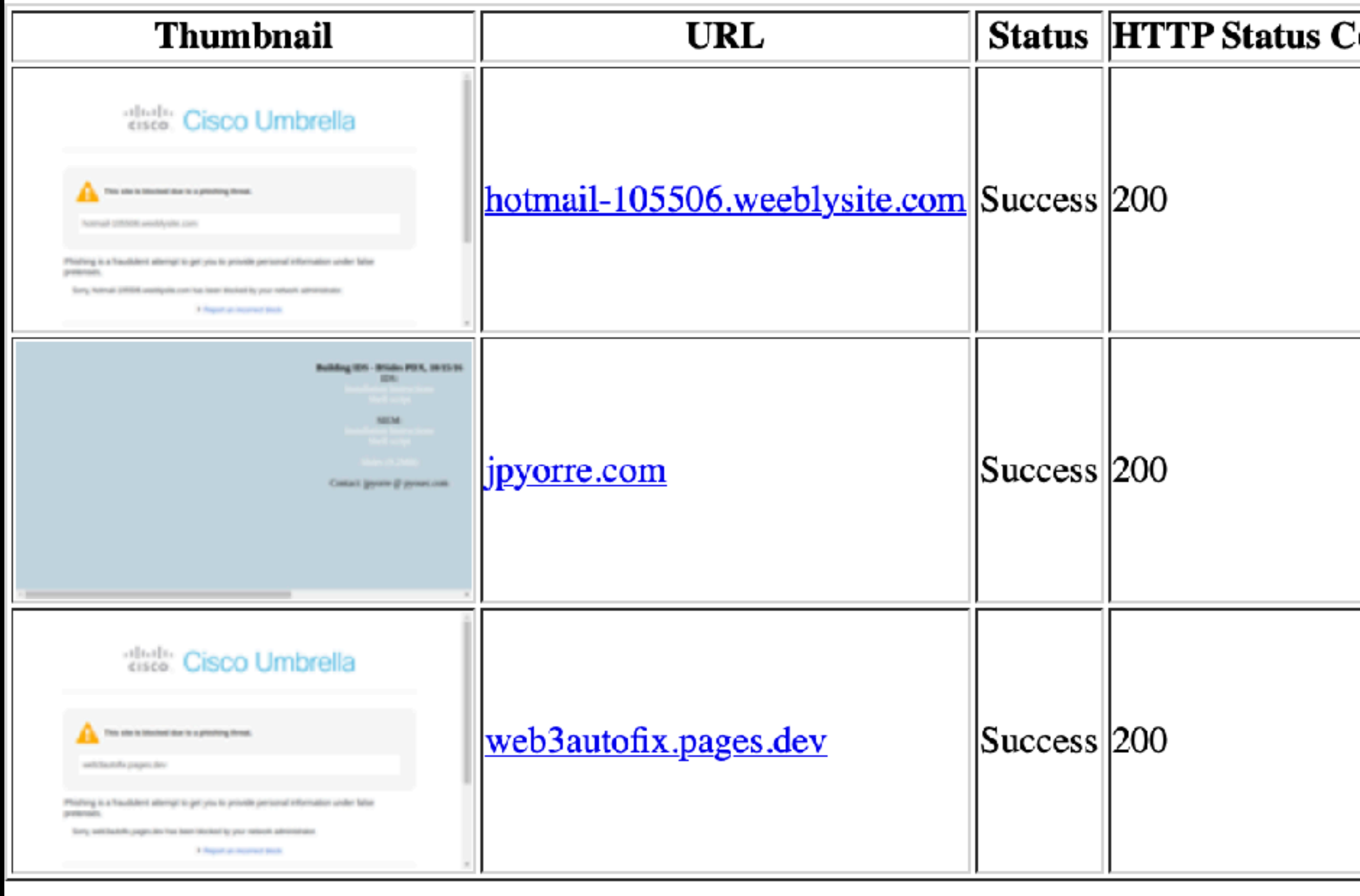


Your connection is not private

Attackers might be trying to steal your information from **hotmail-105506.weeblysite.com** (for example, passwords, messages, or credit cards). Learn more about this warning

NET::ERR_CERT_AUTHORITY_INVALID

Turn on enhanced protection to get Chrome's highest level of security



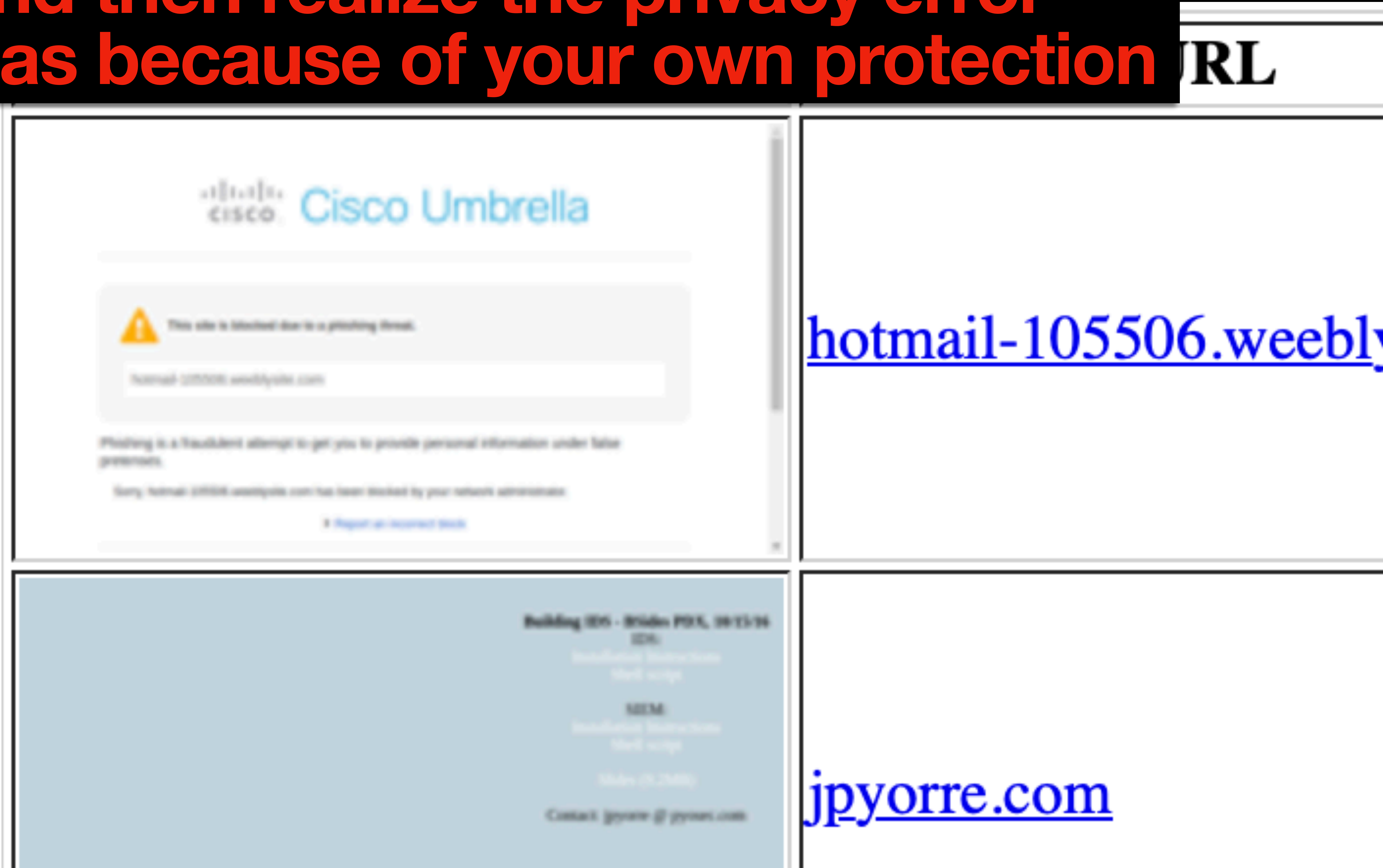
Thumbnail	URL	Status	HTTP Status C
	hotmail-105506.weeblysite.com	Success	200
	jpyorre.com	Success	200
	web3autofix.pages.dev	Success	200

Click through

PROFIT!

Tell Selenium to click as if it were a human…

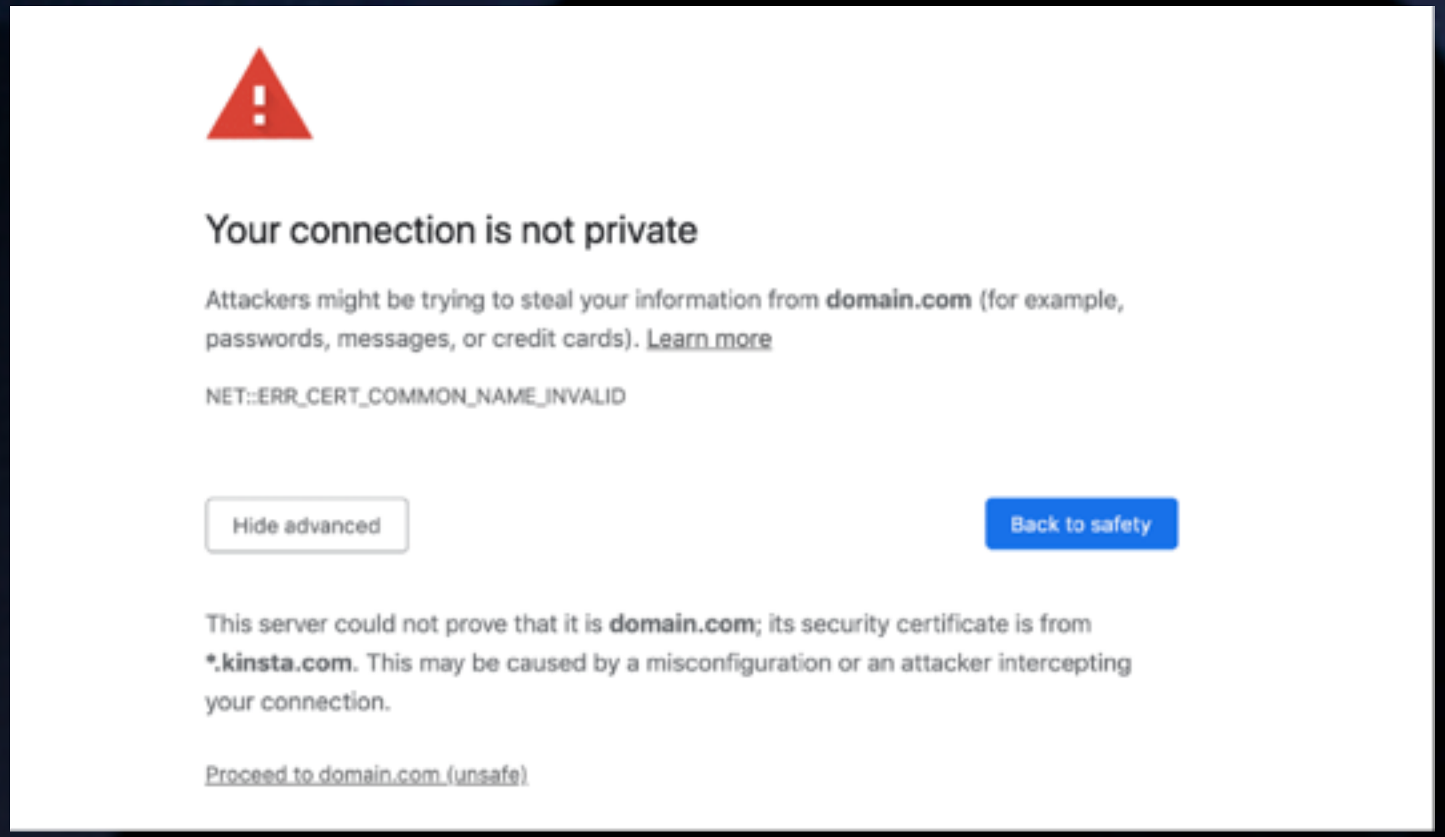
And then realize the privacy error was because of your own protection



hotmail-105506.weebly

jpyorre.com

Bypass Security



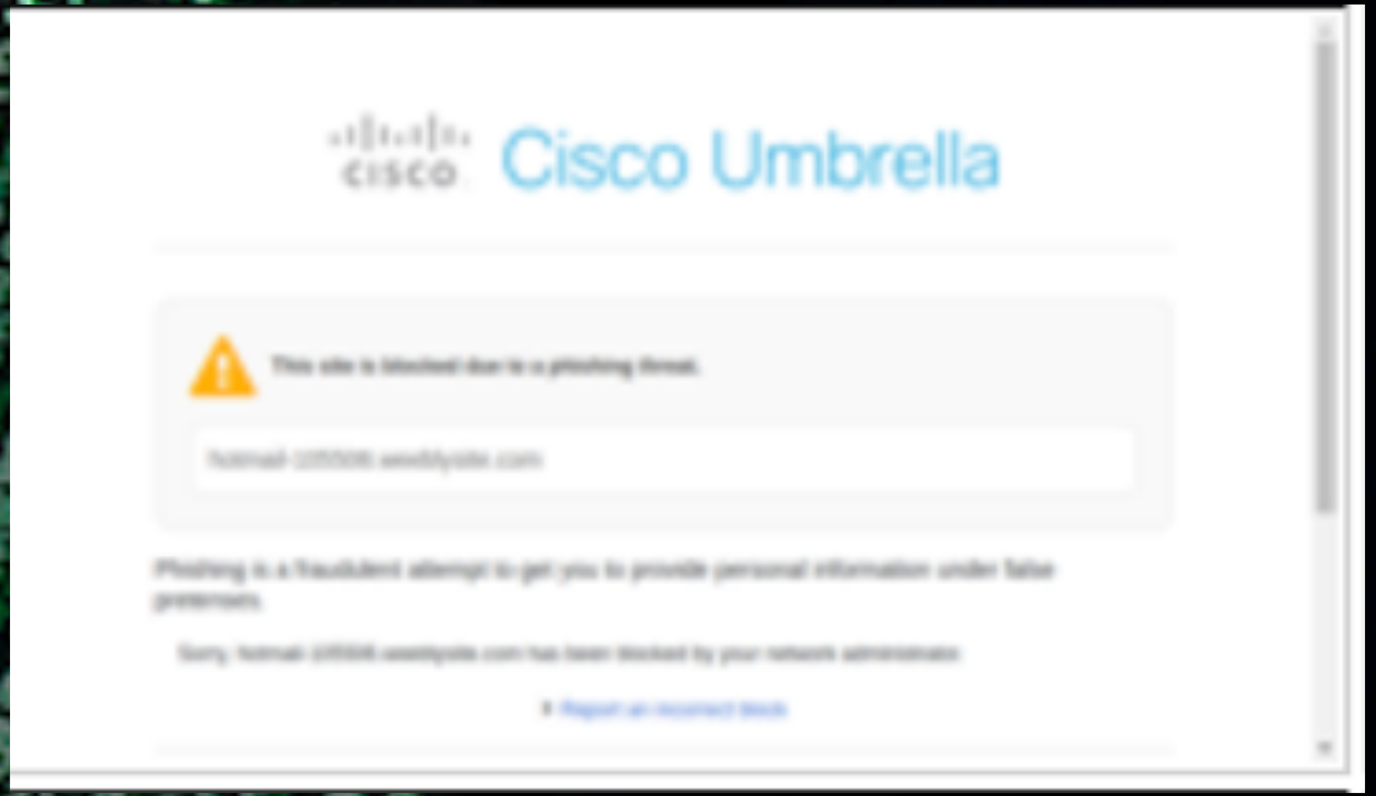
Your connection is not private

Attackers might be trying to steal your information from **domain.com** (for example, passwords, messages, or credit cards). Learn more

NET::ERR_CERT_COMMON_NAME_INVALID

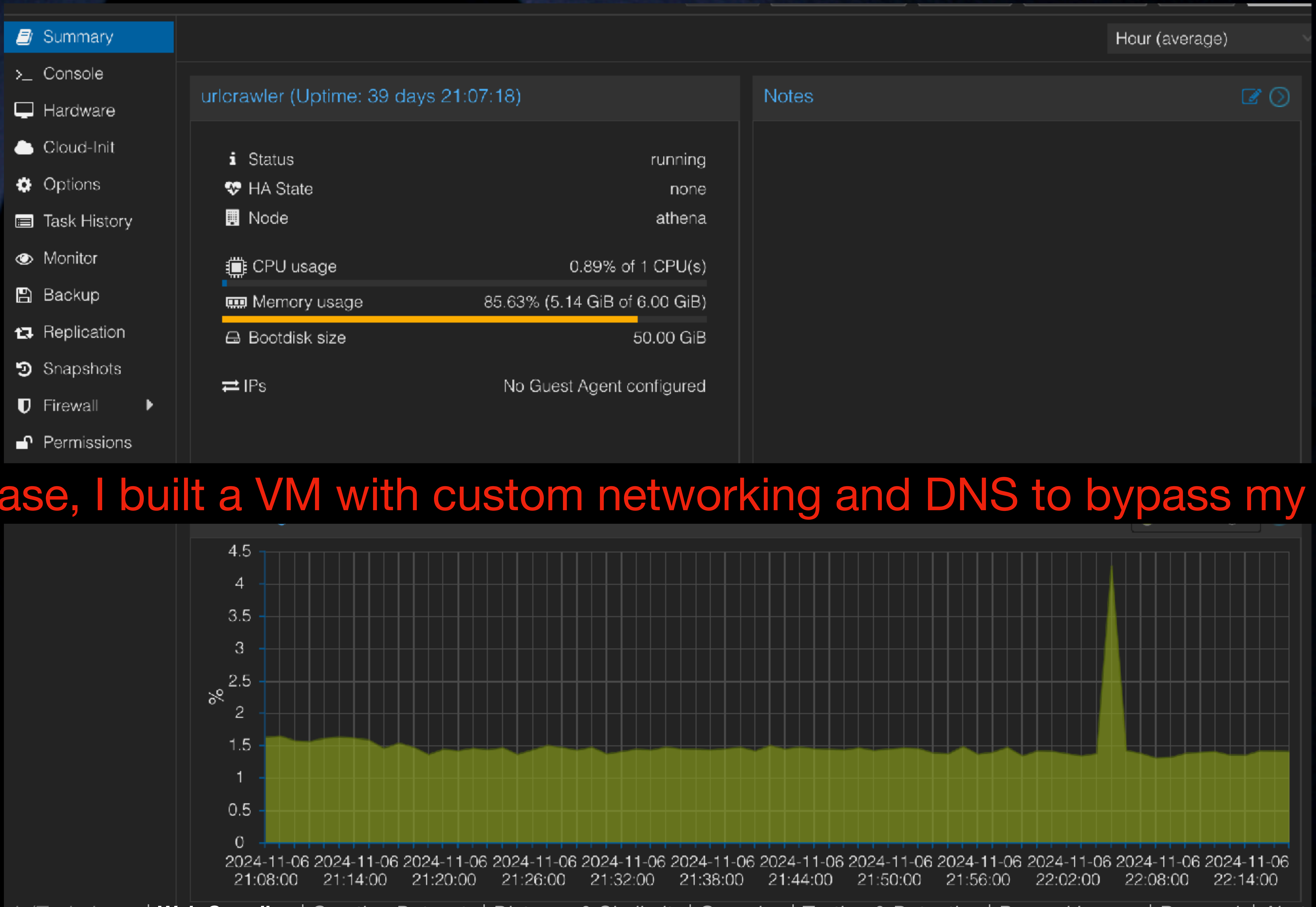
This server could not prove that it is **domain.com**; its security certificate is from ***.kinsta.com**. This may be caused by a misconfiguration or an attacker intercepting your connection.

Proceed to domain.com (unsafe)



You can’t run the crawlers on a network where you are running your threat mitigations

Build a System and use custom DNS



urlcrawler (Uptime: 39 days 21:07:18)

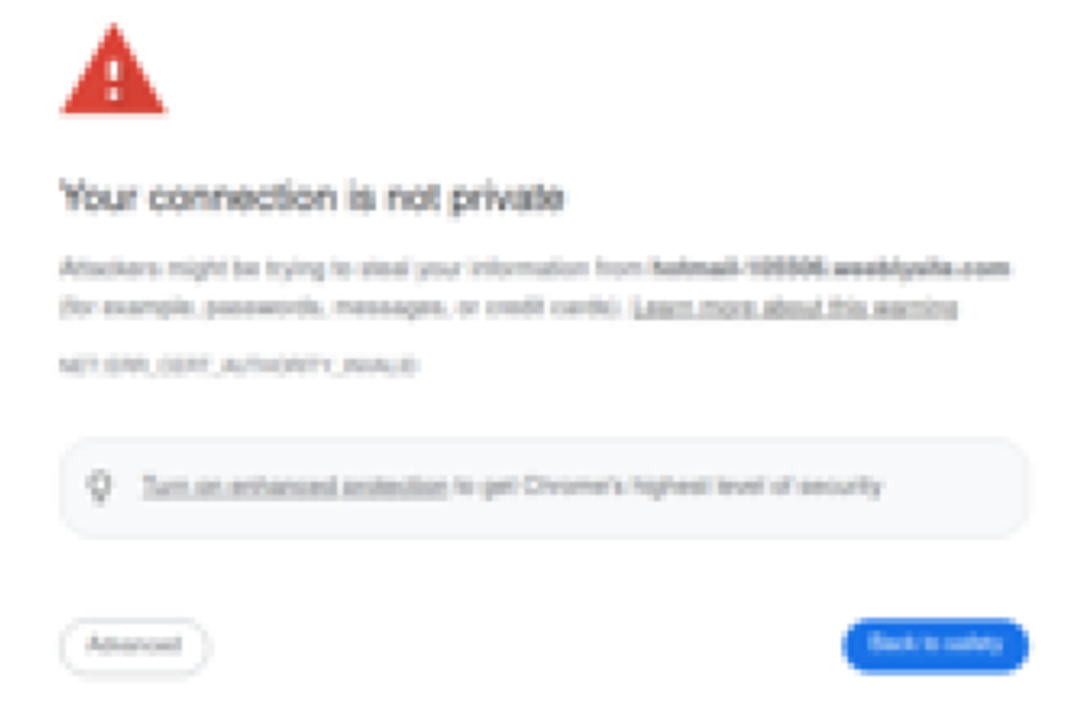
Status	running
HA State	none
Node	athena
CPU usage	0.89% of 1 CPU(s)
Memory usage	85.63% (5.14 GiB of 6.00 GiB)
Bootdisk size	50.00 GiB
IPs	No Guest Agent configured

In my case, I built a VM with custom networking and DNS to bypass my security

Now it’s working…I can visit bad locations

Use Selenium

Read Text

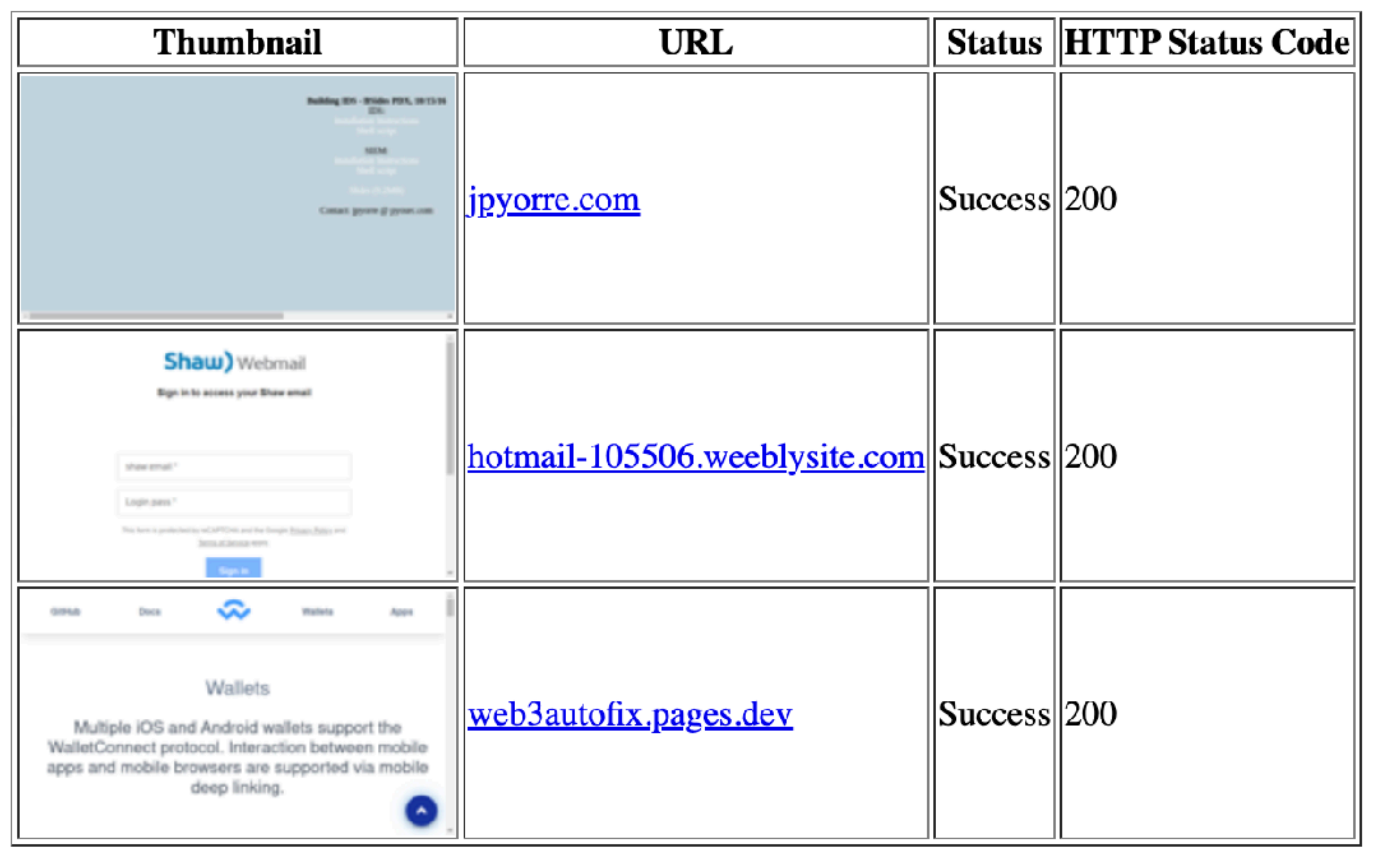


Your connection is not private

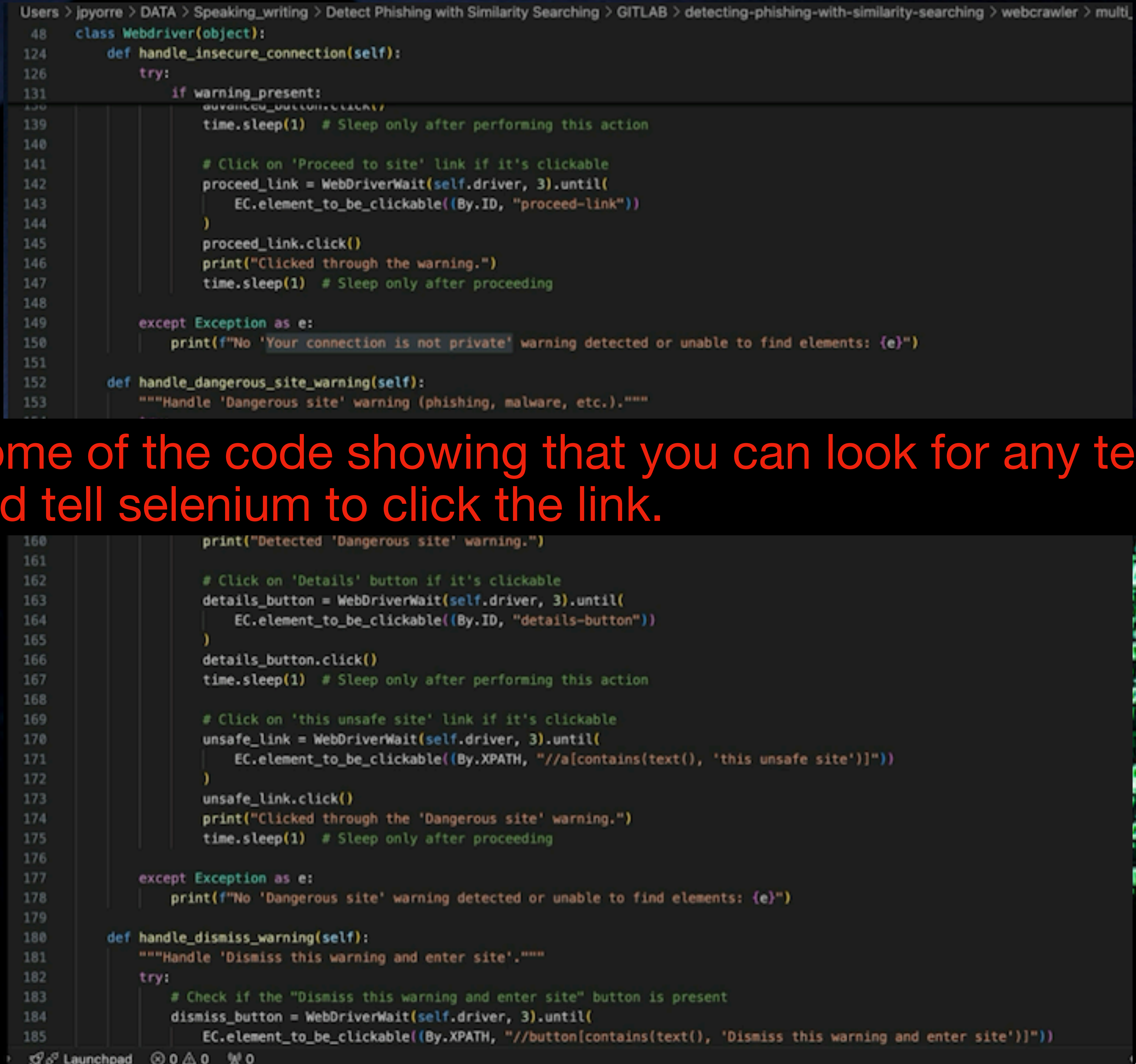
Click through

PROFIT!

…possibly for real this time



Thumbnail	URL	Status	HTTP Status Code
	jpyorrc.com	Success	200
Shaw) Webmail	hotmail-105506.weeblysite.com	Success	200
Wallets Multiple iOS and Android wallets support the WalletConnect protocol. Interaction between mobile apps and mobile browsers are supported via mobile deep linking.	web3autofix.pages.dev	Success	200



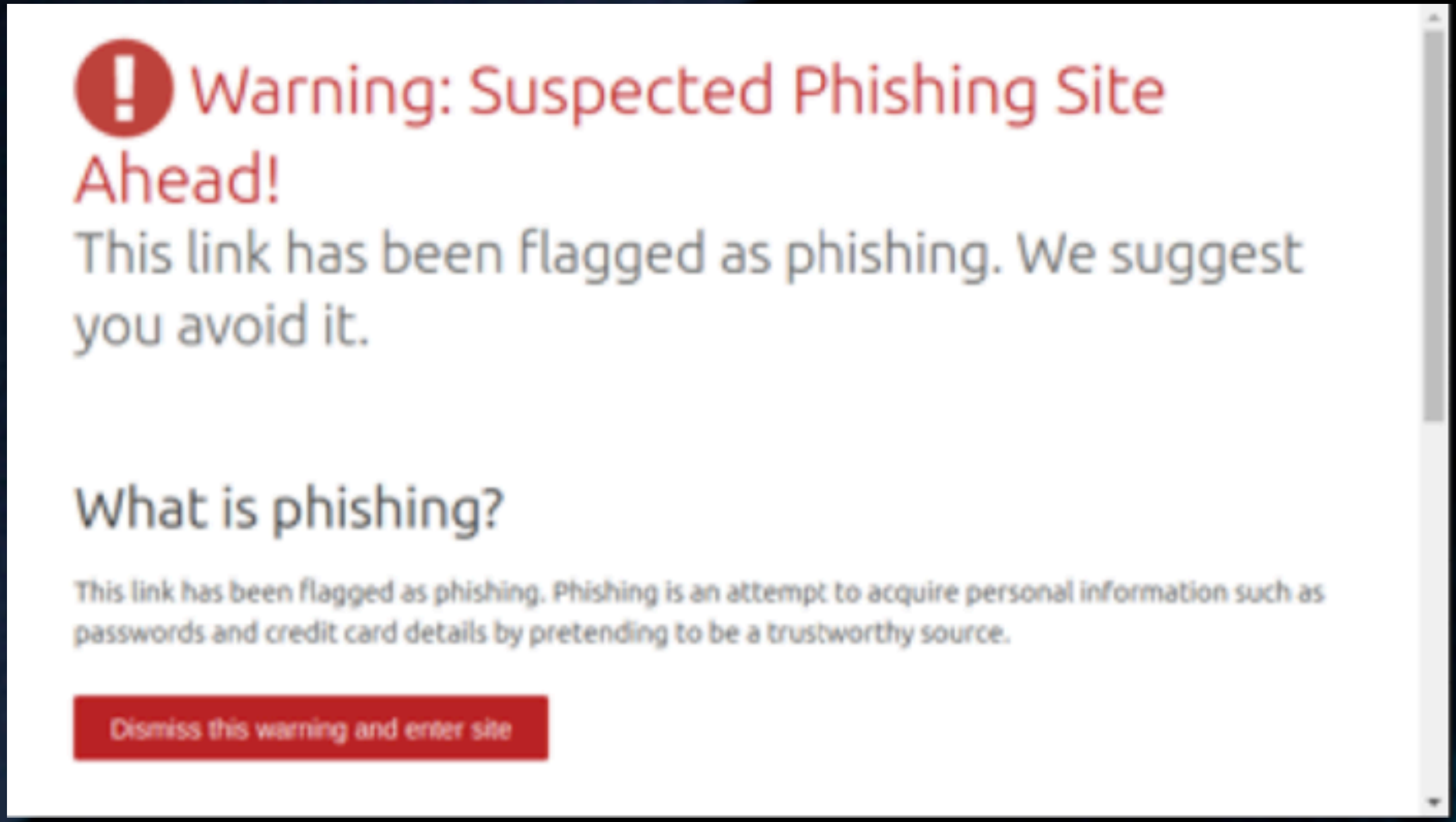
Users > jpyorre > DATA > Speaking_writing > Detect Phishing with Similarity Searching > GITLAB > detecting-phishing-with-similarity-searching > webcrawler > multi_

```
class Webdriver(object):
    def handle_insecure_connection(self):
        try:
            if warning_present:
                advanced_button.click()
                time.sleep(1)  # Sleep only after performing this action

                # Click on 'Proceed to site' link if it's clickable
                proceed_link = WebDriverWait(self.driver, 3).until(
                    EC.element_to_be_clickable((By.ID, "proceed-link"))
                )
                proceed_link.click()
                print("Clicked through the warning.")
                time.sleep(1)  # Sleep only after proceeding

        except Exception as e:
            print(f"No 'Your connection is not private' warning detected or unable to find elements: {e}")

    def handle_dangerous_site_warning(self):
        """Handle 'Dangerous site' warning (phishing, malware, etc.)."""
```

Some of the code showing that you can look for any text and tell selenium to click the link.

```
                print("Detected 'Dangerous site' warning.")

                # Click on 'Details' button if it's clickable
                details_button = WebDriverWait(self.driver, 3).until(
                    EC.element_to_be_clickable((By.ID, "details-button"))
                )
                details_button.click()
                time.sleep(1)  # Sleep only after performing this action

                # Click on 'this unsafe site' link if it's clickable
                unsafe_link = WebDriverWait(self.driver, 3).until(
                    EC.element_to_be_clickable((By.XPATH, "//a[contains(text(), 'this unsafe site')]"))
                )
                unsafe_link.click()
                print("Clicked through the 'Dangerous site' warning.")
                time.sleep(1)  # Sleep only after proceeding

        except Exception as e:
            print(f"No 'Dangerous site' warning detected or unable to find elements: {e}")

    def handle_dismiss_warning(self):
        """Handle 'Dismiss this warning and enter site'."""
        try:
            # Check if the "Dismiss this warning and enter site" button is present
            dismiss_button = WebDriverWait(self.driver, 3).until(
                EC.element_to_be_clickable((By.XPATH, "//button[contains(text(), 'Dismiss this warning and enter site')]"))
```

You end up going from the left side results to the right side results when crawling

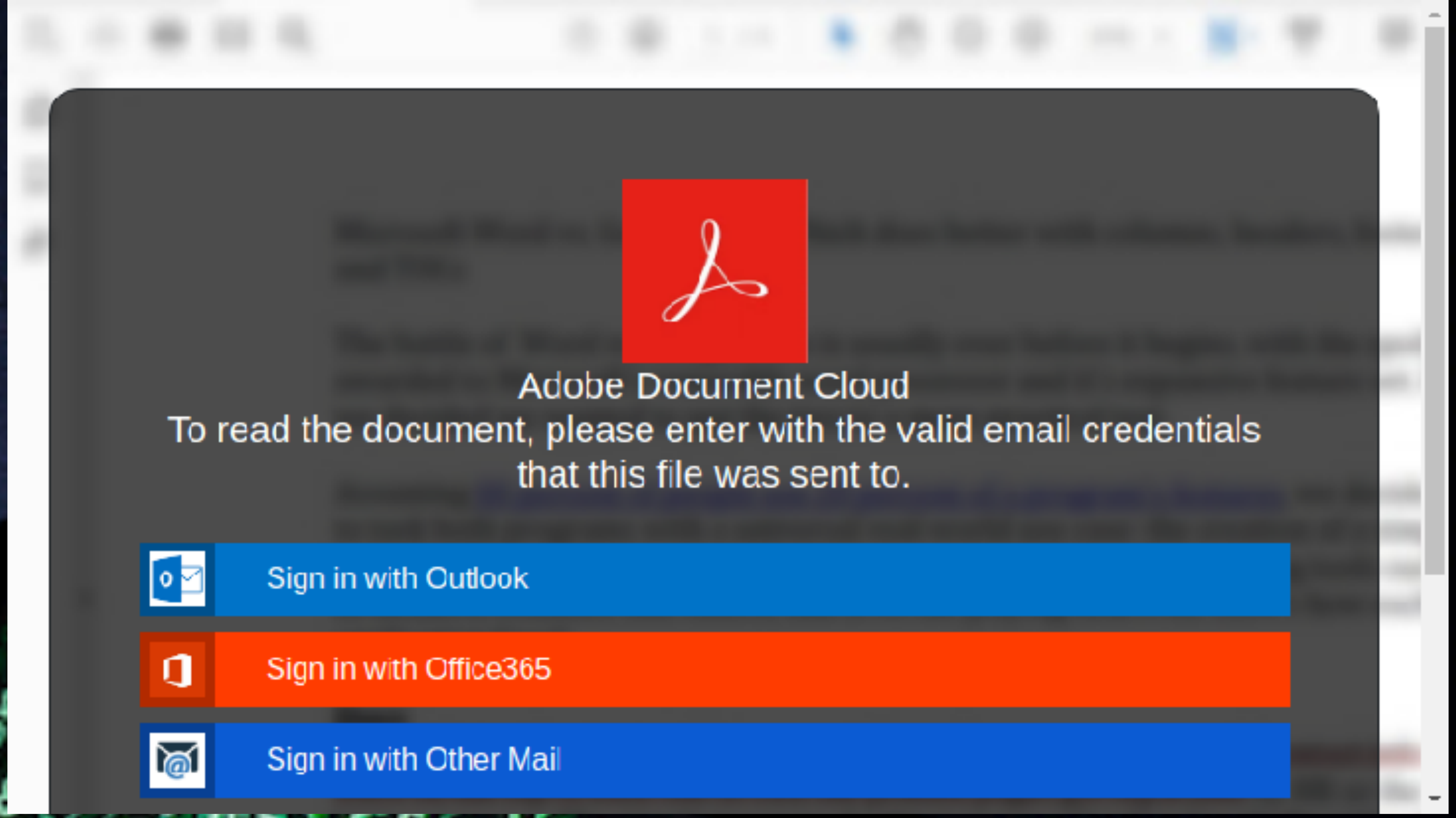


Warning: Suspected Phishing Site Ahead!

This link has been flagged as phishing. We suggest you avoid it.

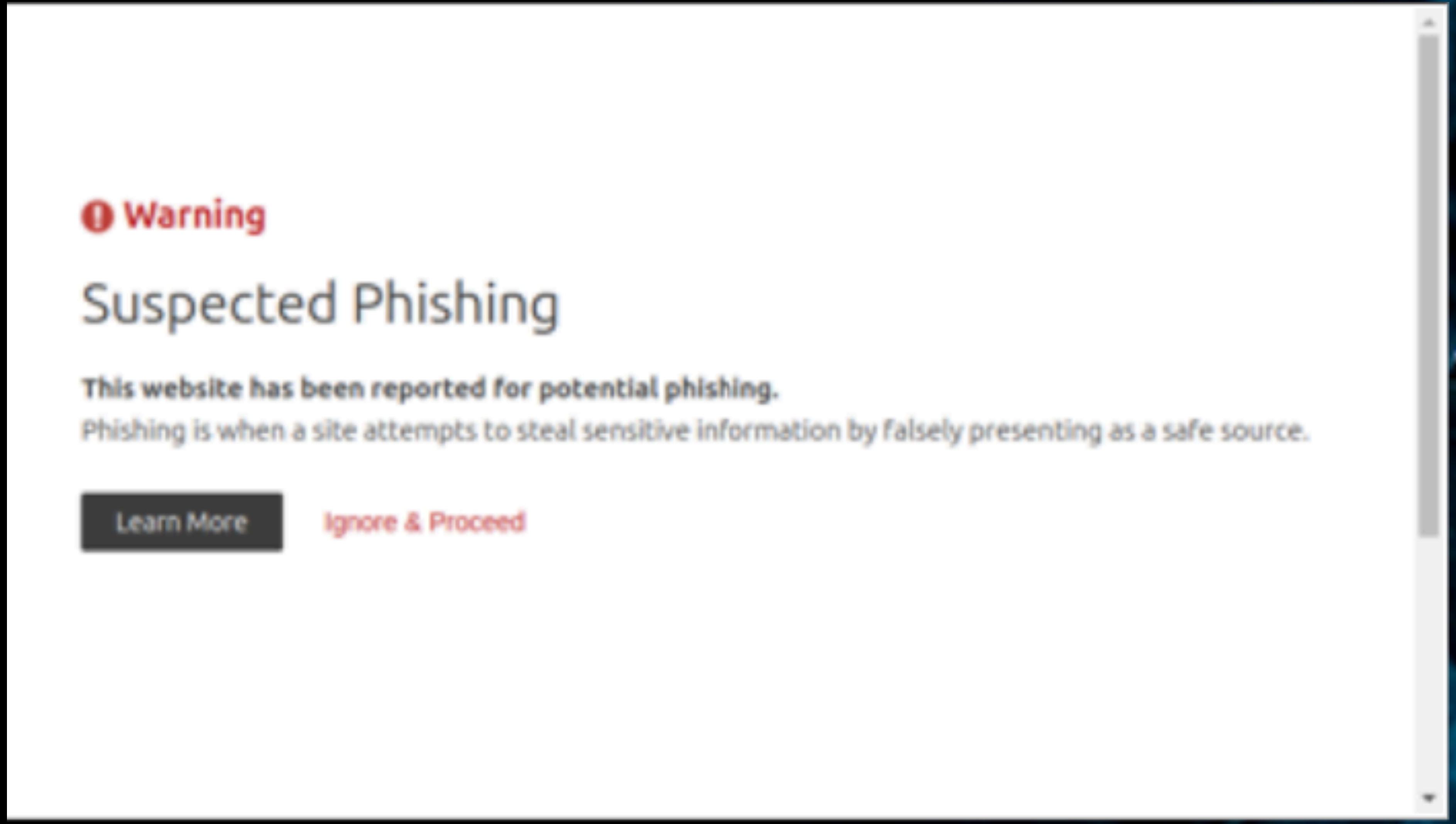
What is phishing?

This link has been flagged as phishing. Phishing is an attempt to acquire personal information such as passwords and credit card details by pretending to be a trustworthy source.



Adobe Document Cloud

To read the document, please enter with the valid email credentials that this file was sent to.

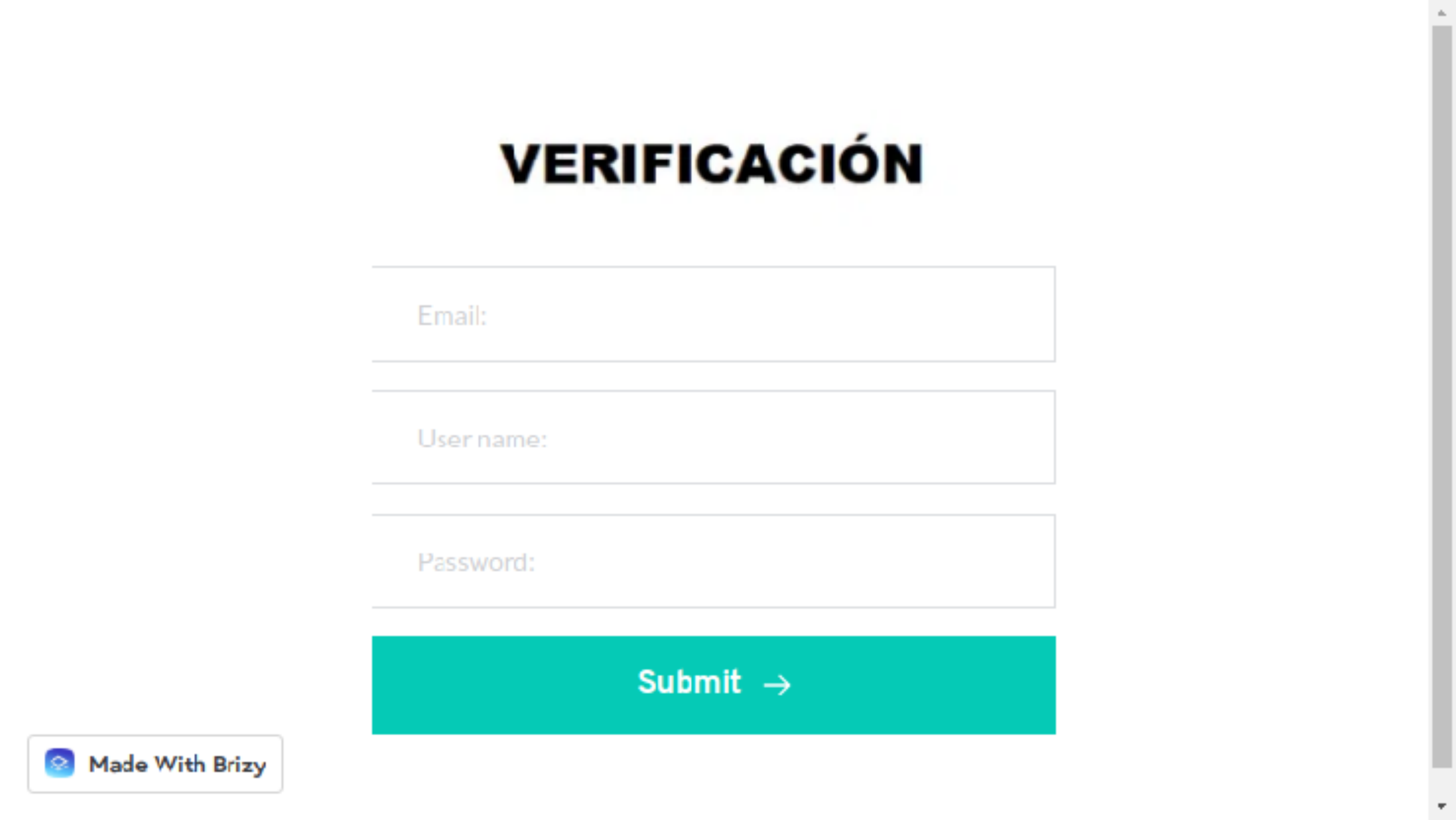


Warning

Suspected Phishing

This website has been reported for potential phishing.

Phishing is when a site attempts to steal sensitive information by falsely presenting as a safe source.



VERIFICACIÓN

Email:

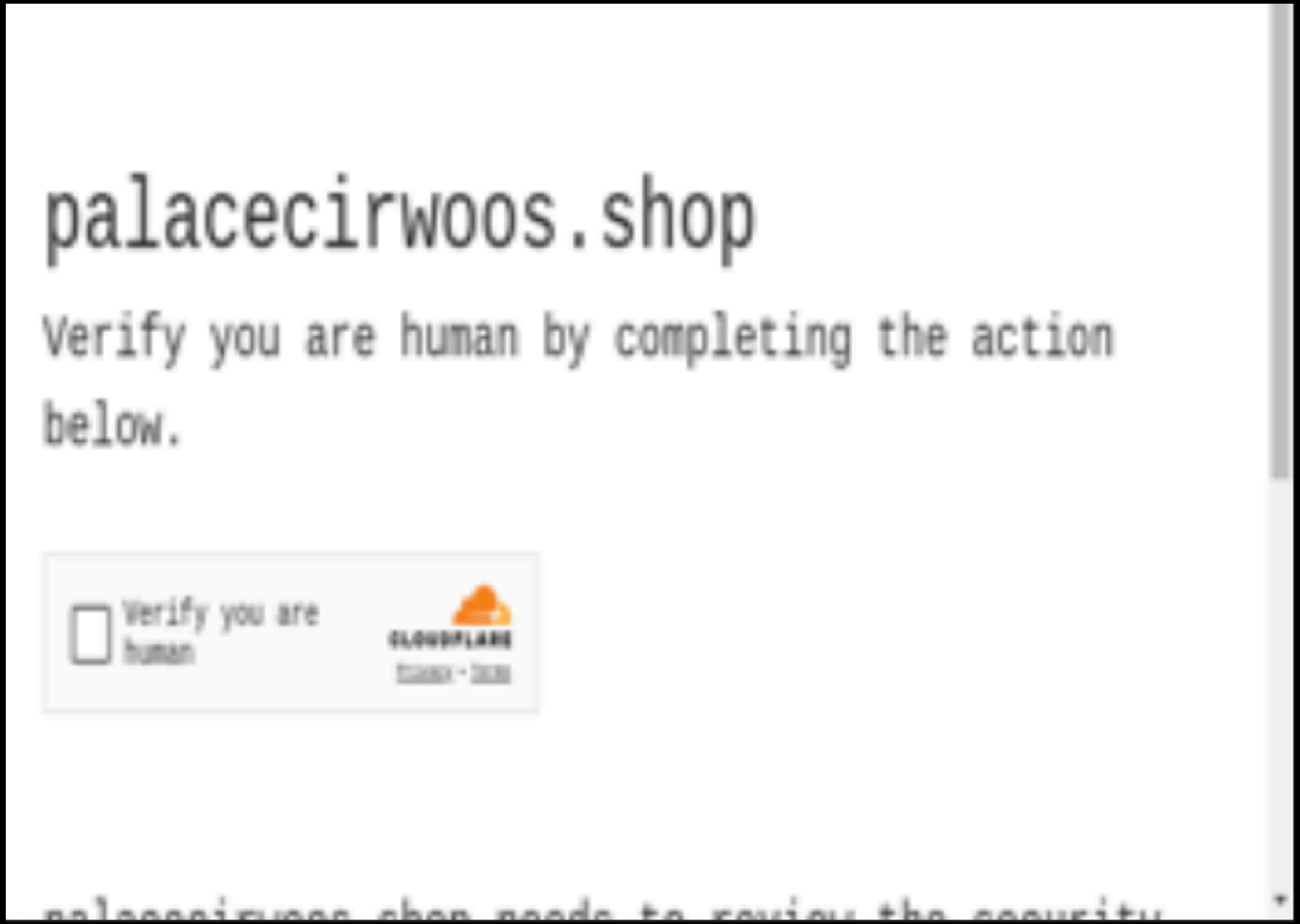
User name:

Password:

FAIL

Another issue is dealing with captchas

Stopped at Captcha

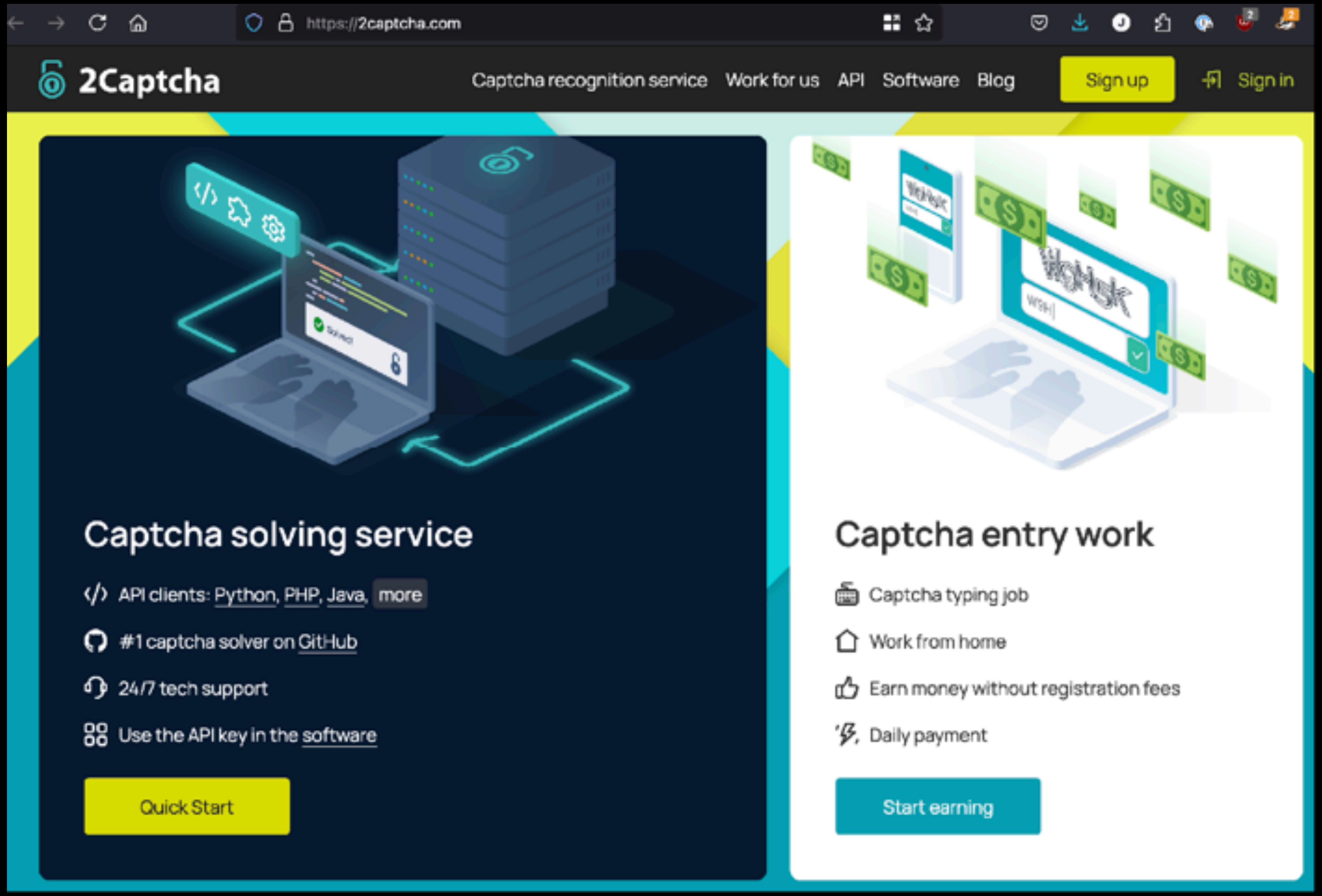

palacecirwoos.shop

Verify you are human by completing the action below.

☐ Verify you are human

palacecirwoos.shop needs to review the security

There are multiple products that make captcha problems easier to solve. I didn't use them in this presentation (they cost money)

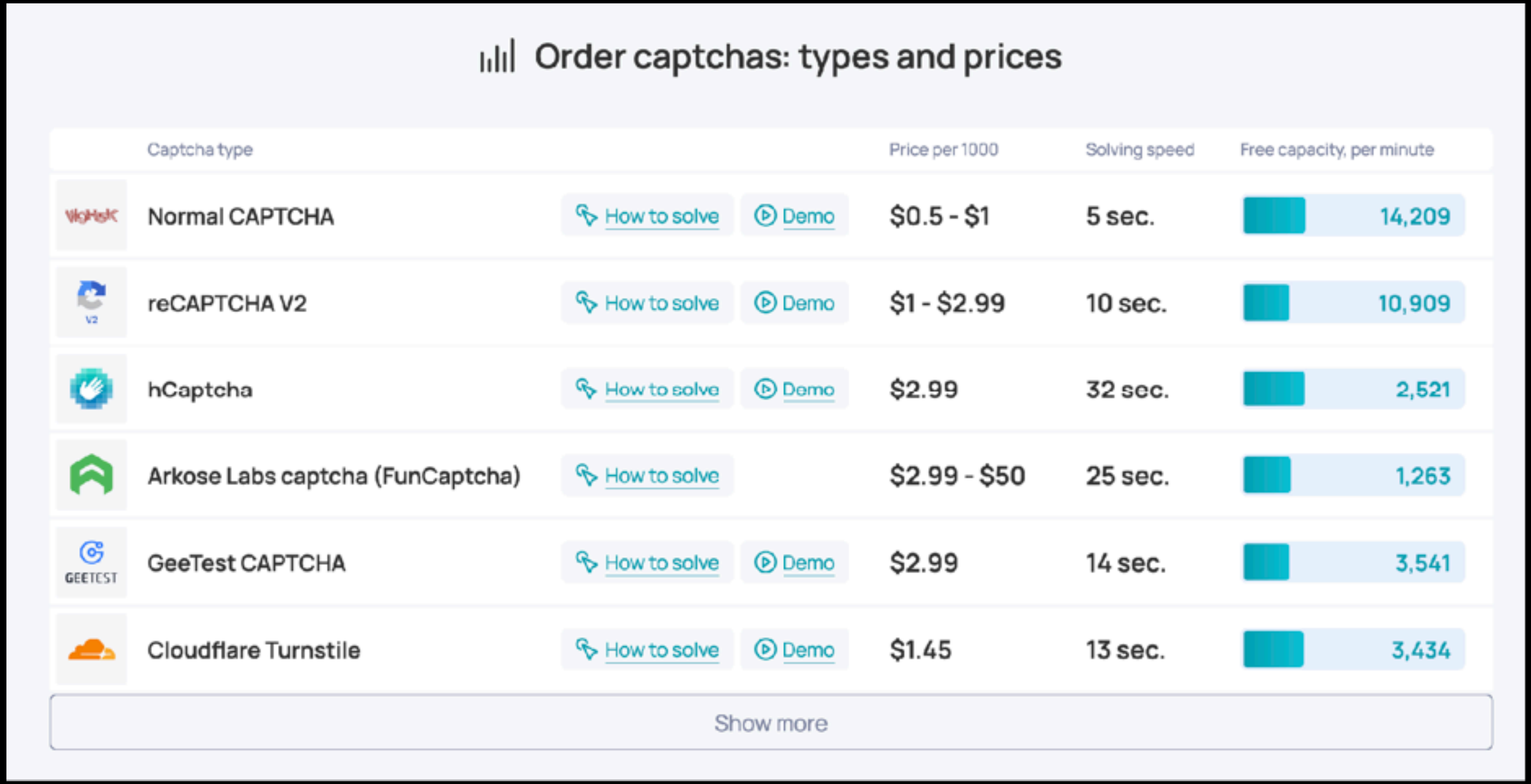


Captcha solving service

- API clients: Python, PHP, Java, more
- #1 captcha solver on GitHub
- 24/7 tech support
- Use the API key in the software

Captcha entry work

- Captcha typing job
- Work from home
- Earn money without registration fees
- Daily payment



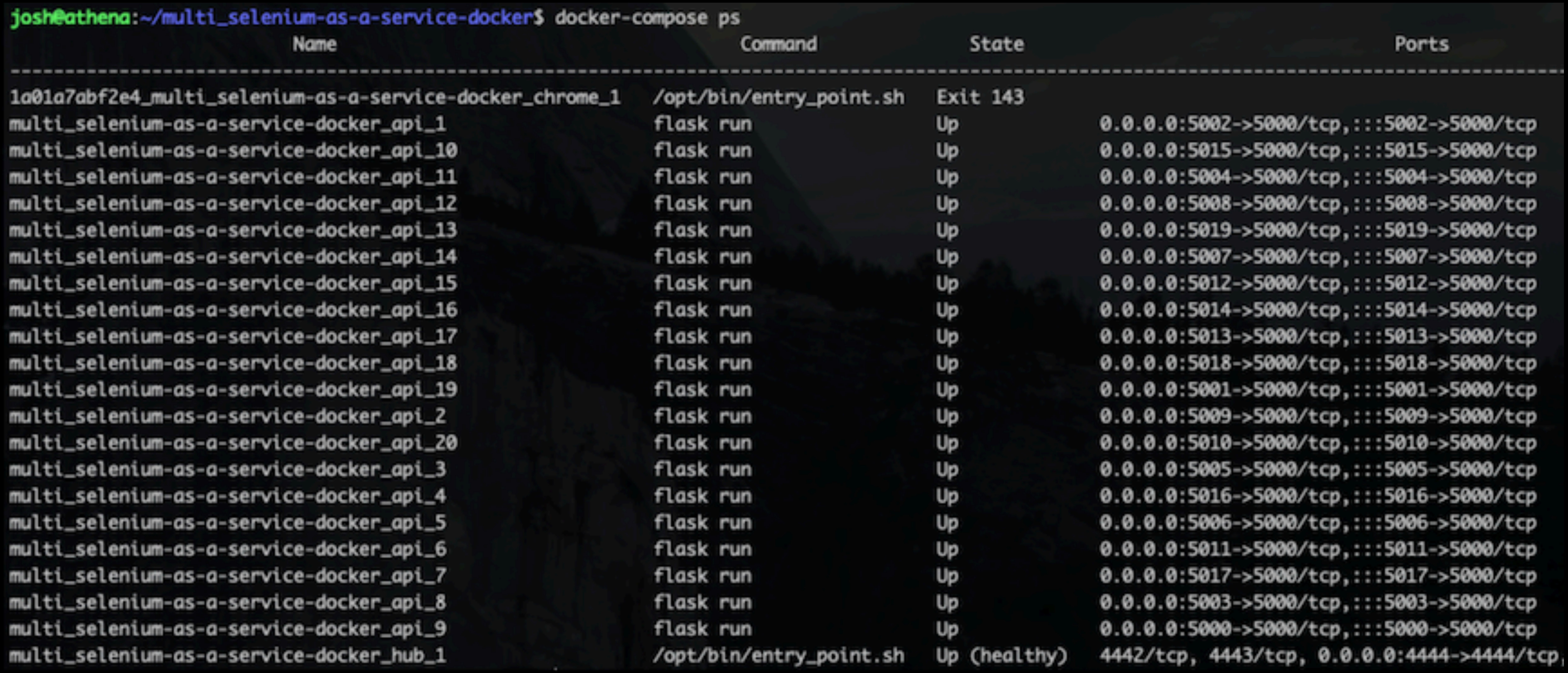
Order captchas: types and prices

Captcha type		Price per 1000	Solving speed	Free capacity, per minute
Normal CAPTCHA	How to solve Demo	\$0.5 - \$1	5 sec.	14,209
reCAPTCHA V2	How to solve Demo	\$1 - \$2.99	10 sec.	10,909
hCaptcha	How to solve Demo	\$2.99	32 sec.	2,521
Arkose Labs captcha (FunCaptcha)	How to solve	\$2.99 - \$50	25 sec.	1,263
GeeTest CAPTCHA	How to solve Demo	\$2.99	14 sec.	3,541
Cloudflare Turnstile	How to solve Demo	\$1.45	13 sec.	3,434

Demonstrating how crawling is slow
You have to go to each page to get the screenshot and other data

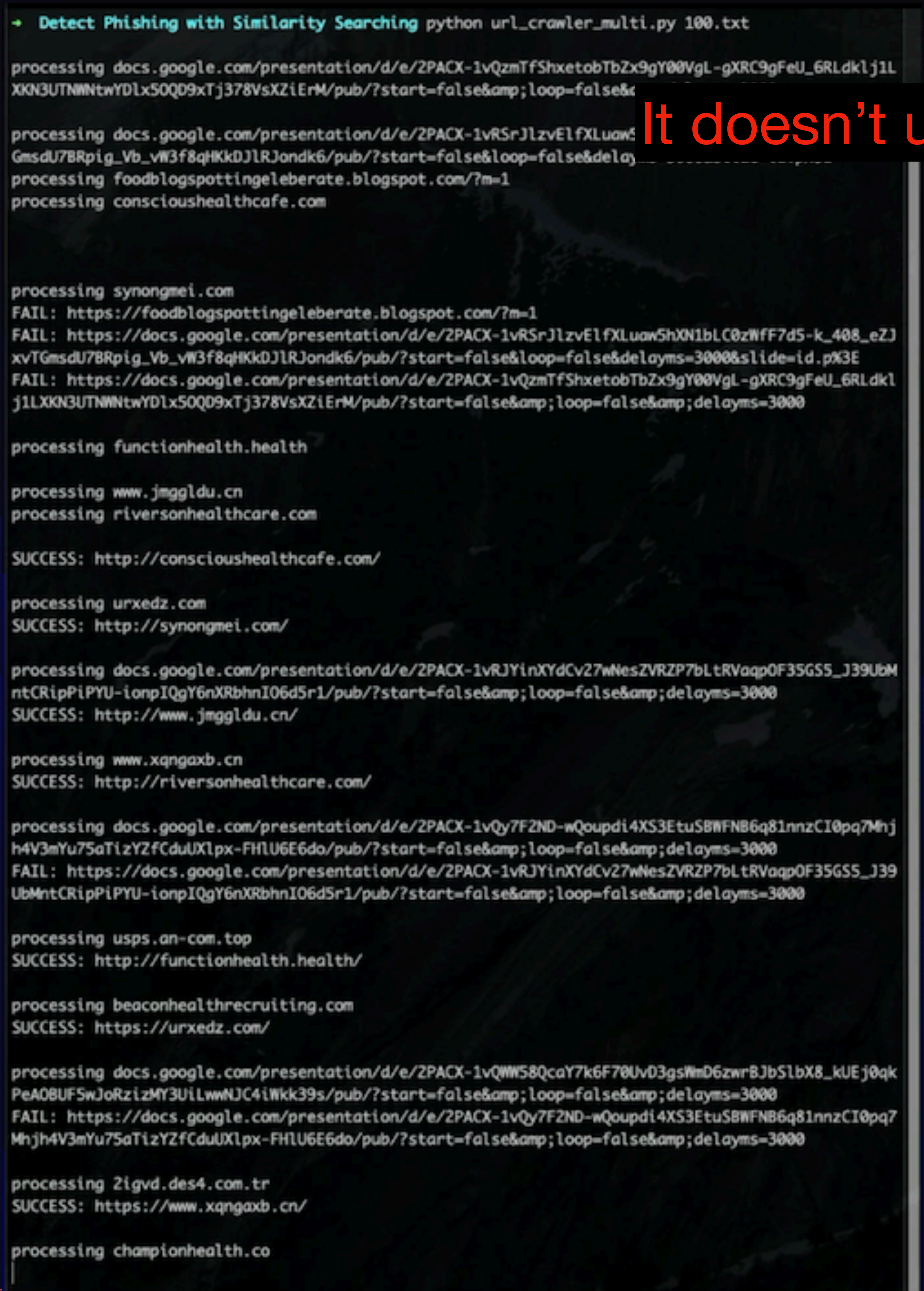
Crawl Speed

To speed things up, I run about 20 docker-based crawlers at once, on one VM

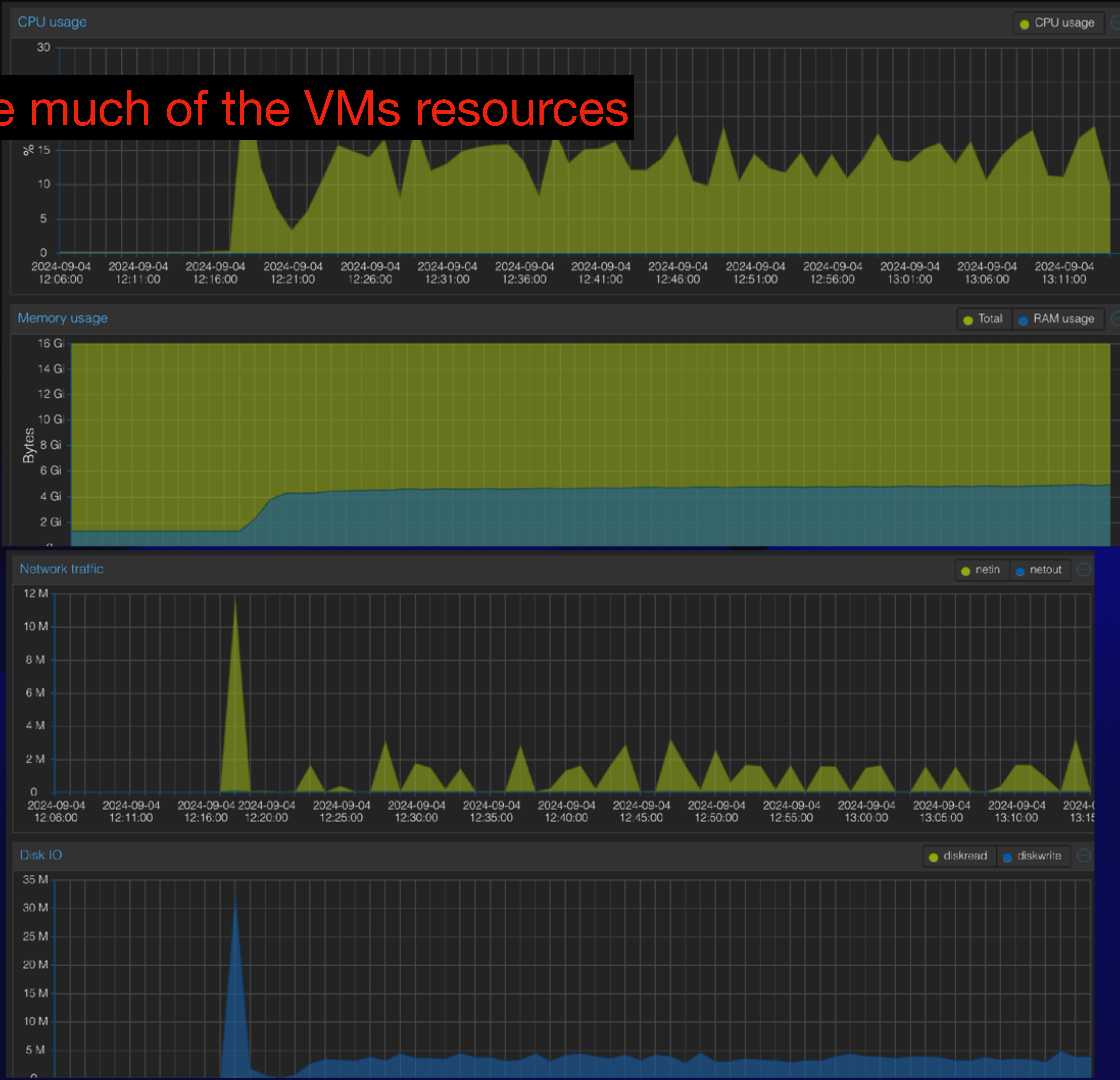


```
josh@athena:~/multi_selenium-as-a-service-docker$ docker-compose ps
                      Name                                   Command              State                         Ports
-------------------------------------------------------------------------------------------------------------------------------------------
1a01a7abf2e4_multi_selenium-as-a-service-docker_chrome_1   /opt/bin/entry_point.sh   Exit 143
multi_selenium-as-a-service-docker_api_1                   flask run                 Up           0.0.0.0:5002->5000/tcp,:::5002->5000/tcp
multi_selenium-as-a-service-docker_api_10                  flask run                 Up           0.0.0.0:5015->5000/tcp,:::5015->5000/tcp
multi_selenium-as-a-service-docker_api_11                  flask run                 Up           0.0.0.0:5004->5000/tcp,:::5004->5000/tcp
multi_selenium-as-a-service-docker_api_12                  flask run                 Up           0.0.0.0:5008->5000/tcp,:::5008->5000/tcp
multi_selenium-as-a-service-docker_api_13                  flask run                 Up           0.0.0.0:5019->5000/tcp,:::5019->5000/tcp
multi_selenium-as-a-service-docker_api_14                  flask run                 Up           0.0.0.0:5007->5000/tcp,:::5007->5000/tcp
multi_selenium-as-a-service-docker_api_15                  flask run                 Up           0.0.0.0:5012->5000/tcp,:::5012->5000/tcp
multi_selenium-as-a-service-docker_api_16                  flask run                 Up           0.0.0.0:5014->5000/tcp,:::5014->5000/tcp
multi_selenium-as-a-service-docker_api_17                  flask run                 Up           0.0.0.0:5013->5000/tcp,:::5013->5000/tcp
multi_selenium-as-a-service-docker_api_18                  flask run                 Up           0.0.0.0:5018->5000/tcp,:::5018->5000/tcp
multi_selenium-as-a-service-docker_api_19                  flask run                 Up           0.0.0.0:5001->5000/tcp,:::5001->5000/tcp
multi_selenium-as-a-service-docker_api_2                   flask run                 Up           0.0.0.0:5009->5000/tcp,:::5009->5000/tcp
multi_selenium-as-a-service-docker_api_20                  flask run                 Up           0.0.0.0:5010->5000/tcp,:::5010->5000/tcp
multi_selenium-as-a-service-docker_api_3                   flask run                 Up           0.0.0.0:5005->5000/tcp,:::5005->5000/tcp
multi_selenium-as-a-service-docker_api_4                   flask run                 Up           0.0.0.0:5016->5000/tcp,:::5016->5000/tcp
multi_selenium-as-a-service-docker_api_5                   flask run                 Up           0.0.0.0:5006->5000/tcp,:::5006->5000/tcp
multi_selenium-as-a-service-docker_api_6                   flask run                 Up           0.0.0.0:5011->5000/tcp,:::5011->5000/tcp
multi_selenium-as-a-service-docker_api_7                   flask run                 Up           0.0.0.0:5017->5000/tcp,:::5017->5000/tcp
multi_selenium-as-a-service-docker_api_8                   flask run                 Up           0.0.0.0:5003->5000/tcp,:::5003->5000/tcp
multi_selenium-as-a-service-docker_api_9                   flask run                 Up           0.0.0.0:5000->5000/tcp,:::5000->5000/tcp
multi_selenium-as-a-service-docker_hub_1                   /opt/bin/entry_point.sh   Up (healthy) 4442/tcp, 4443/tcp, 0.0.0.0:4444->4444/tcp,
```

It doesn’t use much of the VMs resources



```
→ Detect Phishing with Similarity Searching python url_crawler_multi.py 100.txt

processing docs.google.com/presentation/d/e/2PACX-1vQzmTfShxetobTbZx9gY00VgL-gXRC9gFeU_6RLdklj1L
XKN3UTNWNtwYDlx50QD9xTj378VsXZiErM/pub/?start=false&amp;loop=false&

processing docs.google.com/presentation/d/e/2PACX-1vRSrJlzvElfXLuaw
GmsdU7BRpig_Vb_vW3f8qHKkDJlRJondk6/pub/?start=false&loop=false&delay
processing foodblogspottingeleberate.blogspot.com/?m=1
processing conscioushealthcafe.com



processing synongmei.com
FAIL: https://foodblogspottingeleberate.blogspot.com/?m=1
FAIL: https://docs.google.com/presentation/d/e/2PACX-1vRSrJlzvElfXLuaw5hXN1bLC0zWfF7d5-k_408_eZJ
xvTGmsdU7BRpig_Vb_vW3f8qHKkDJlRJondk6/pub/?start=false&loop=false&delayms=3000&slide=id.pX3E
FAIL: https://docs.google.com/presentation/d/e/2PACX-1vQzmTfShxetobTbZx9gY00VgL-gXRC9gFeU_6RLdkl
j1LXKN3UTNWNtwYDlx50QD9xTj378VsXZiErM/pub/?start=false&amp;loop=false&amp;delayms=3000

processing functionhealth.health

processing www.jmggldu.cn
processing riversonhealthcare.com

SUCCESS: http://conscioushealthcafe.com/

processing urxedz.com
SUCCESS: http://synongmei.com/

processing docs.google.com/presentation/d/e/2PACX-1vRJYinXYdCv27wNesZVRZP7bLtRVaqpOF35GS5_J39UbM
ntCRipPiPYU-ionpIQgY6nXRbhnIO6d5r1/pub/?start=false&amp;loop=false&amp;delayms=3000
SUCCESS: http://www.jmggldu.cn/

processing www.xqngaxb.cn
SUCCESS: http://riversonhealthcare.com/

processing docs.google.com/presentation/d/e/2PACX-1vQy7F2ND-wQoupdi4XS3EtuSBWFNB6q81nnzCI0pq7Mhj
h4V3mYu75aTizYZfCduUXlpx-FHlU6E6do/pub/?start=false&amp;loop=false&amp;delayms=3000
FAIL: https://docs.google.com/presentation/d/e/2PACX-1vRJYinXYdCv27wNesZVRZP7bLtRVaqpOF35GS5_J39
UbMntCRipPiPYU-ionpIQgY6nXRbhnIO6d5r1/pub/?start=false&amp;loop=false&amp;delayms=3000

processing usps.an-com.top
SUCCESS: http://functionhealth.health/

processing beaconhealthrecruiting.com
SUCCESS: https://urxedz.com/

processing docs.google.com/presentation/d/e/2PACX-1vQWW58QcaY7k6F70UvD3gsWmD6zwrBJbSlbX8_kUEj0qk
PeAOBUF5wJoRzizMY3UiLwwNJC4iWkk39s/pub/?start=false&amp;loop=false&amp;delayms=3000
FAIL: https://docs.google.com/presentation/d/e/2PACX-1vQy7F2ND-wQoupdi4XS3EtuSBWFNB6q81nnzCI0pq7
Mhjh4V3mYu75aTizYZfCduUXlpx-FHlU6E6do/pub/?start=false&amp;loop=false&amp;delayms=3000

processing 2igvd.des4.com.tr
SUCCESS: https://www.xqngaxb.cn/

processing championhealth.co
```



1 VM with 20 crawlers is not enough. I want them all over, but where?

But it’s not enough. I want

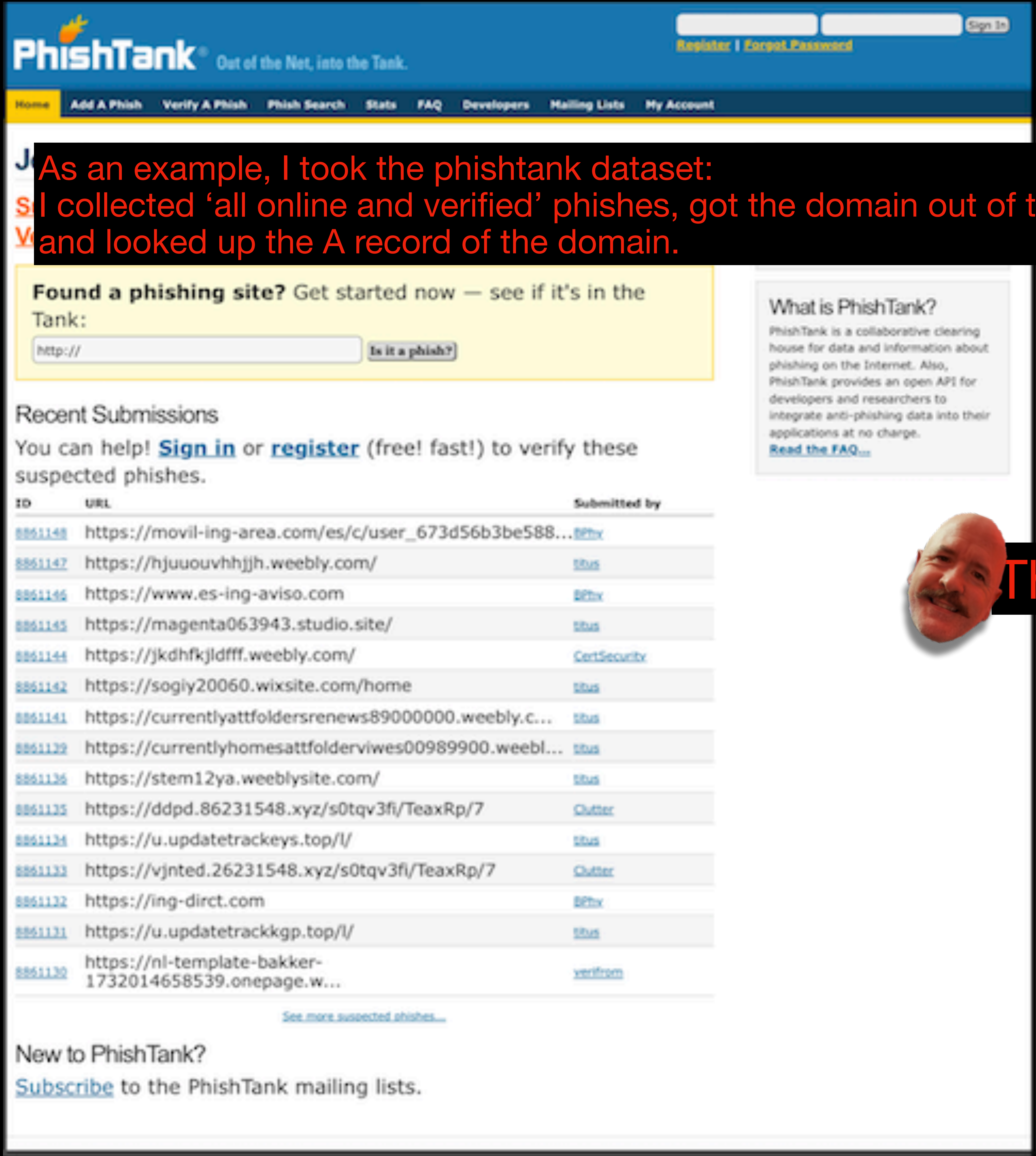
Even More Crawlers!

In the right places

Where is the best place to put my crawlers?

Crawler Distribution

As an example, I took the phishtank dataset:
I collected ‘all online and verified’ phishes, got the domain out of the URL, and looked up the A record of the domain.



Found a phishing site? Get started now — see if it's in the Tank:

http:// Is it a phish?

What is PhishTank?

PhishTank is a collaborative clearing house for data and information about phishing on the Internet. Also, PhishTank provides an open API for developers and researchers to integrate anti-phishing data into their applications at no charge.
Read the FAQ...

Recent Submissions

You can help! Sign in or register (free! fast!) to verify these suspected phishes.

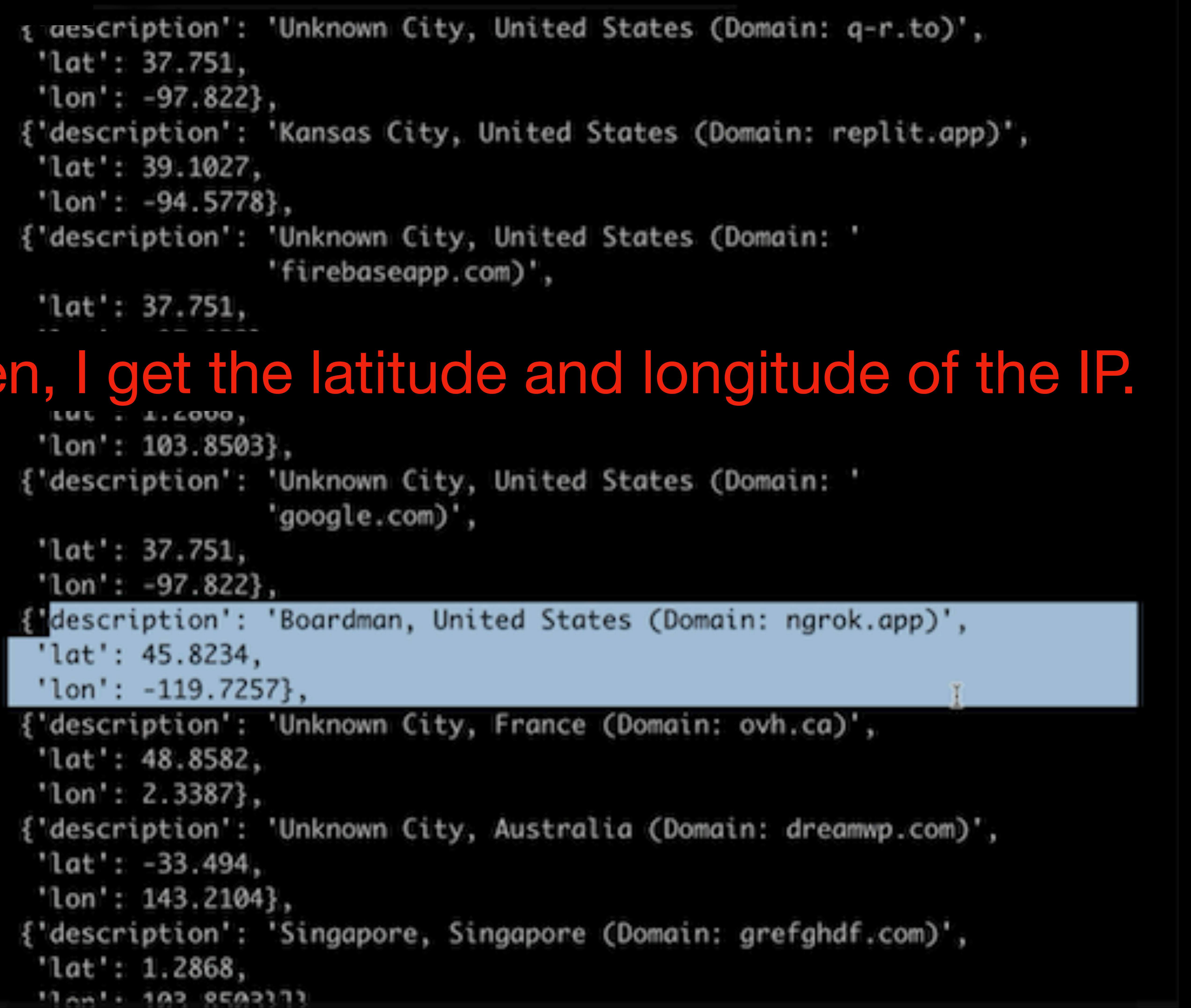
ID	URL	Submitted by
8861148	https://movil-ing-area.com/es/c/user_673d56b3be588...	BPhy
8861147	https://hjuuouvhhjjh.weebly.com/	titus
8861146	https://www.es-ing-aviso.com	BPhy
8861145	https://magenta063943.studio.site/	titus
8861144	https://jkdhfkjldfff.weebly.com/	CertSecurity
8861142	https://sogiy20060.wixsite.com/home	titus
8861141	https://currentlyattfoldersrenews89000000.weebly.c...	titus
8861139	https://currentlyhomesattfolderviwes00989900.weebl...	titus
8861136	https://stem12ya.weeblysite.com/	titus
8861135	https://ddpd.86231548.xyz/s0tqv3fi/TeaxRp/7	Clutter
8861134	https://u.updatetrackeys.top/l/	titus
8861133	https://vjnted.26231548.xyz/s0tqv3fi/TeaxRp/7	Clutter
8861132	https://ing-dirct.com	BPhy
8861131	https://u.updatetrackkgp.top/l/	titus
8861130	https://nl-template-bakker-1732014658539.onepage.w...	verifrom

See more suspected phishes...

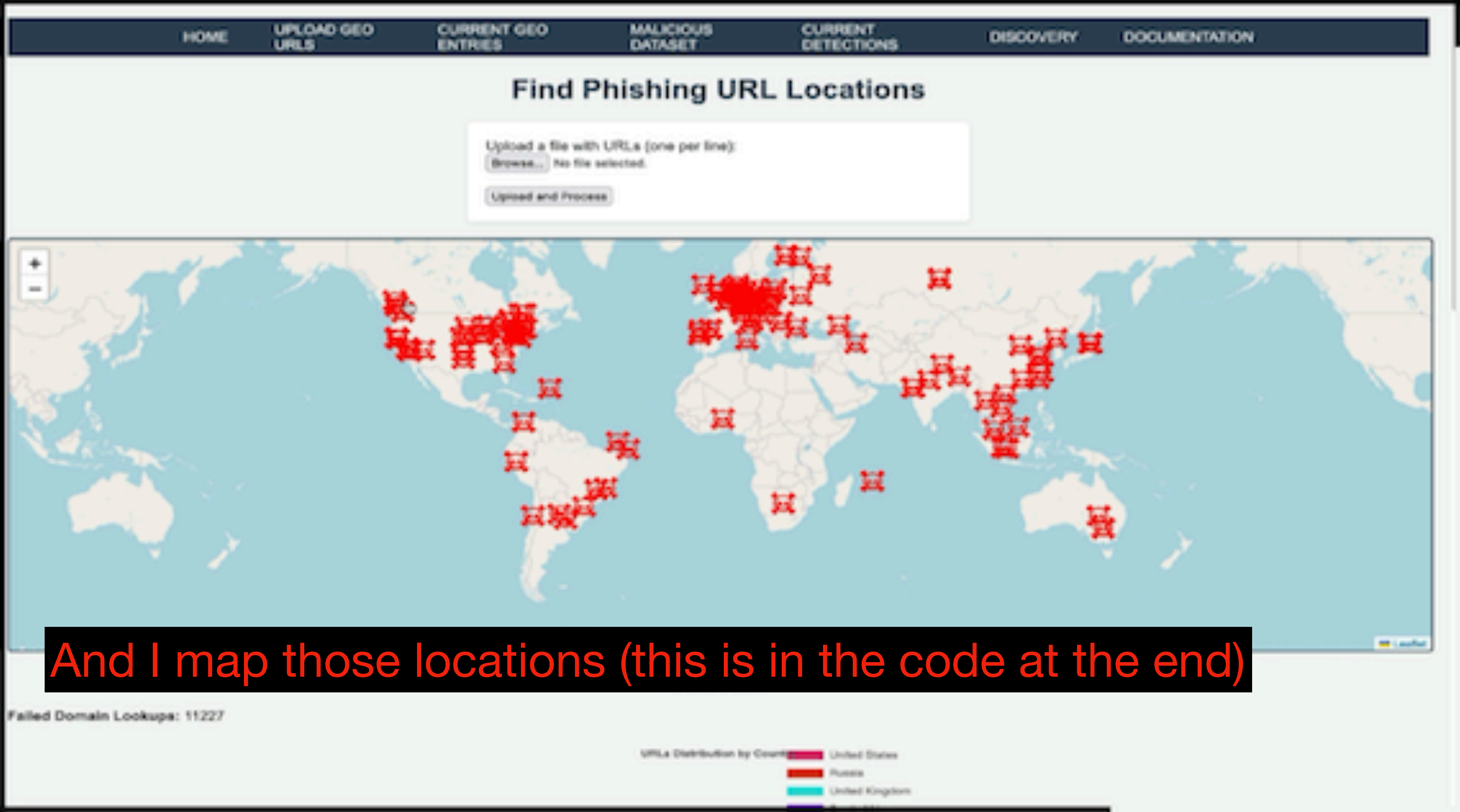
New to PhishTank?

Subscribe to the PhishTank mailing lists.

Then, I get the latitude and longitude of the IP.



```
{'description': 'Unknown City, United States (Domain: q-r.to)',
 'lat': 37.751,
 'lon': -97.822},
{'description': 'Kansas City, United States (Domain: replit.app)',
 'lat': 39.1027,
 'lon': -94.5778},
{'description': 'Unknown City, United States (Domain: '
                'firebaseapp.com)',
 'lat': 37.751,
 'lat': 1.2868,
 'lon': 103.8503},
{'description': 'Unknown City, United States (Domain: '
                'google.com)',
 'lat': 37.751,
 'lon': -97.822},
{'description': 'Boardman, United States (Domain: ngrok.app)',
 'lat': 45.8234,
 'lon': -119.7257},
{'description': 'Unknown City, France (Domain: ovh.ca)',
 'lat': 48.8582,
 'lon': 2.3387},
{'description': 'Unknown City, Australia (Domain: dreamwp.com)',
 'lat': -33.494,
 'lon': 143.2104},
{'description': 'Singapore, Singapore (Domain: grefghdf.com)',
 'lat': 1.2868,
 'lon': 103.8503}]
```

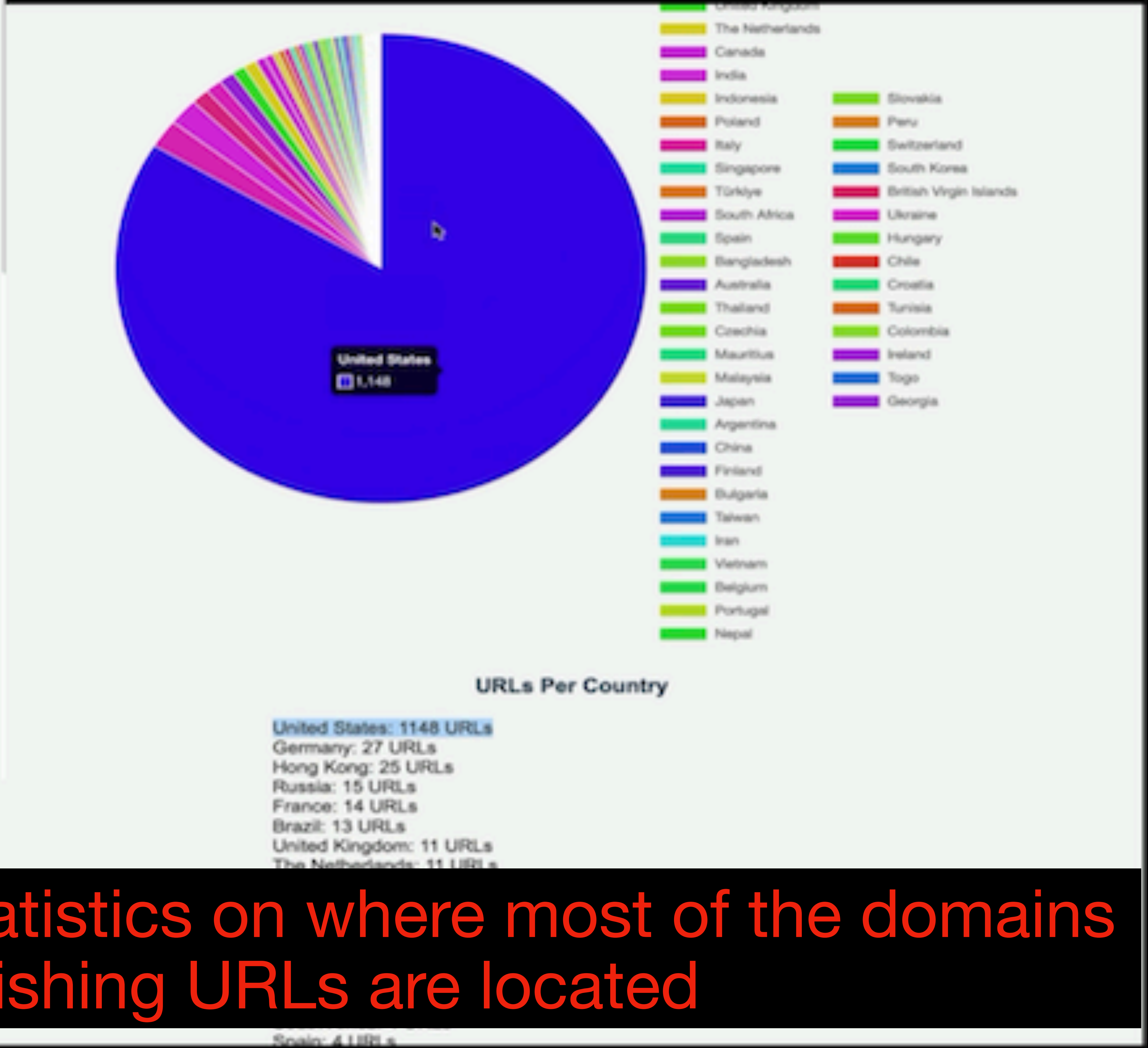


Find Phishing URL Locations

Upload a file with URLs (one per line):

Failed Domain Lookups: 11227

And I map those locations (this is in the code at the end)

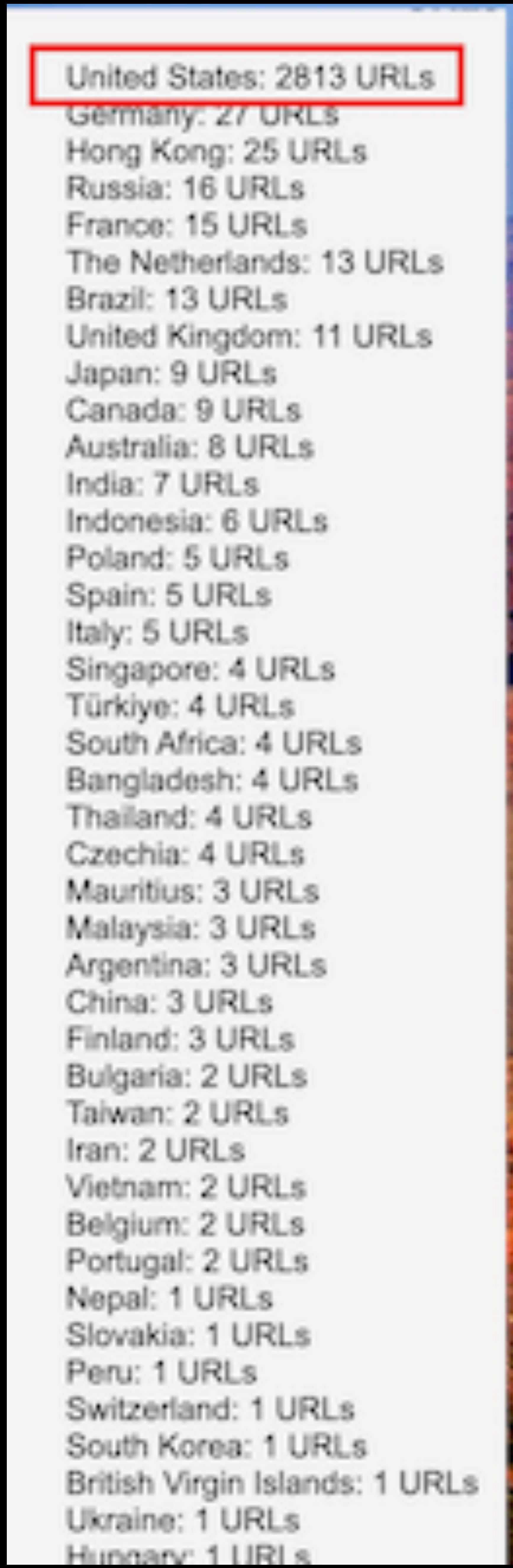


URLs Per Country

United States: 1148 URLs
Germany: 27 URLs
Hong Kong: 25 URLs
Russia: 15 URLs
France: 14 URLs
Brazil: 13 URLs
United Kingdom: 11 URLs
The Netherlands: 11 URLs
Spain: 4 URLs

And get statistics on where most of the domains hosting phishing URLs are located

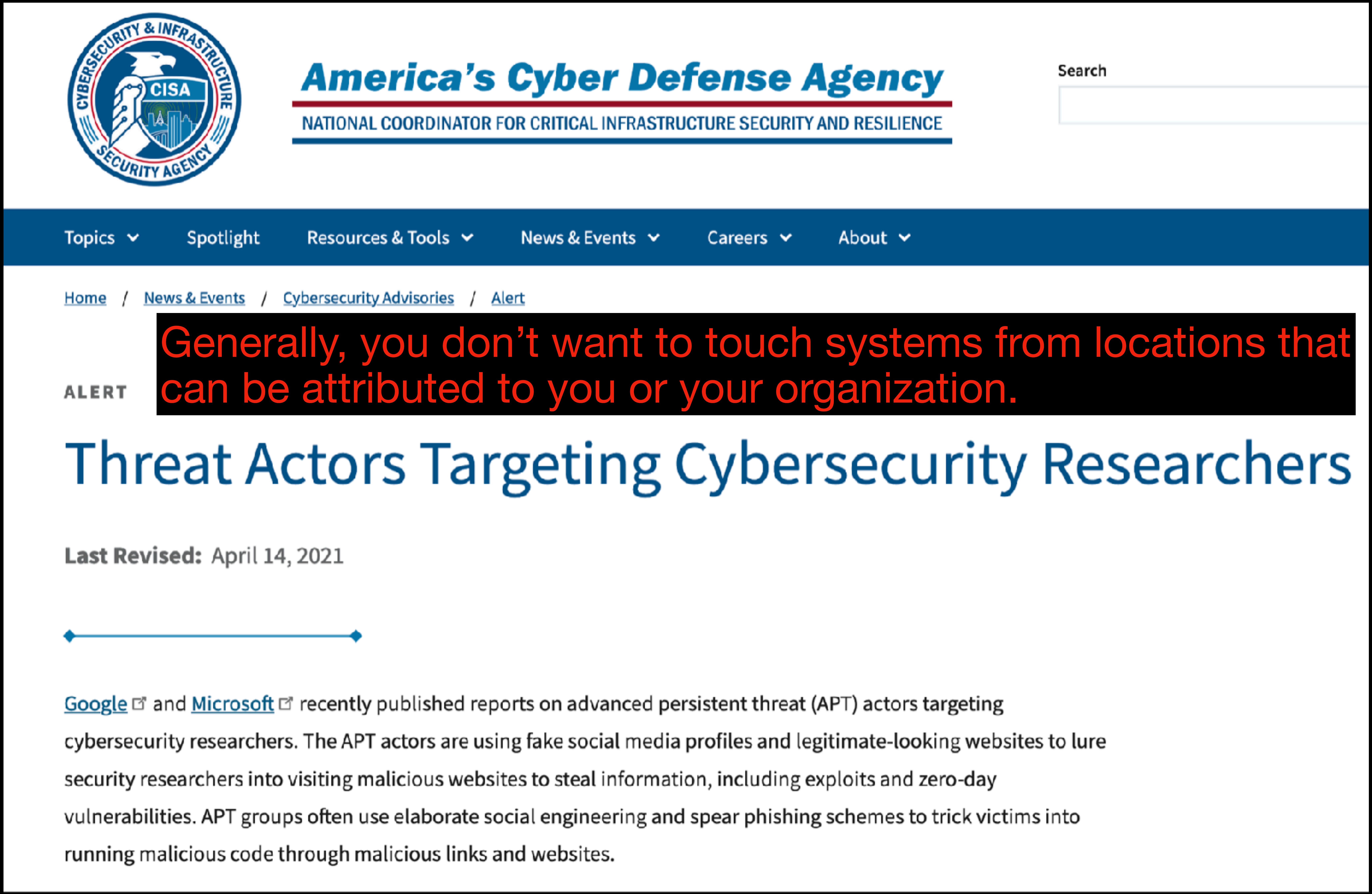
USA in the lead… *(cries.jpg)*



United States: 2813 URLs
Germany: 27 URLs
Hong Kong: 25 URLs
Russia: 16 URLs
France: 15 URLs
The Netherlands: 13 URLs
Brazil: 13 URLs
United Kingdom: 11 URLs
Japan: 9 URLs
Canada: 9 URLs
Australia: 8 URLs
India: 7 URLs
Indonesia: 6 URLs
Poland: 5 URLs
Spain: 5 URLs
Italy: 5 URLs
Singapore: 4 URLs
Türkiye: 4 URLs
South Africa: 4 URLs
Bangladesh: 4 URLs
Thailand: 4 URLs
Czechia: 4 URLs
Mauritius: 3 URLs
Malaysia: 3 URLs
Argentina: 3 URLs
China: 3 URLs
Finland: 3 URLs
Bulgaria: 2 URLs
Taiwan: 2 URLs
Iran: 2 URLs
Vietnam: 2 URLs
Belgium: 2 URLs
Portugal: 2 URLs
Nepal: 1 URLs
Slovakia: 1 URLs
Peru: 1 URLs
Switzerland: 1 URLs
South Korea: 1 URLs
British Virgin Islands: 1 URLs
Ukraine: 1 URLs
Hungary: 1 URLs

Quick sidebar about staying hidden as a threat hunter

Staying hidden?



America's Cyber Defense Agency

NATIONAL COORDINATOR FOR CRITICAL INFRASTRUCTURE SECURITY AND RESILIENCE

ALERT

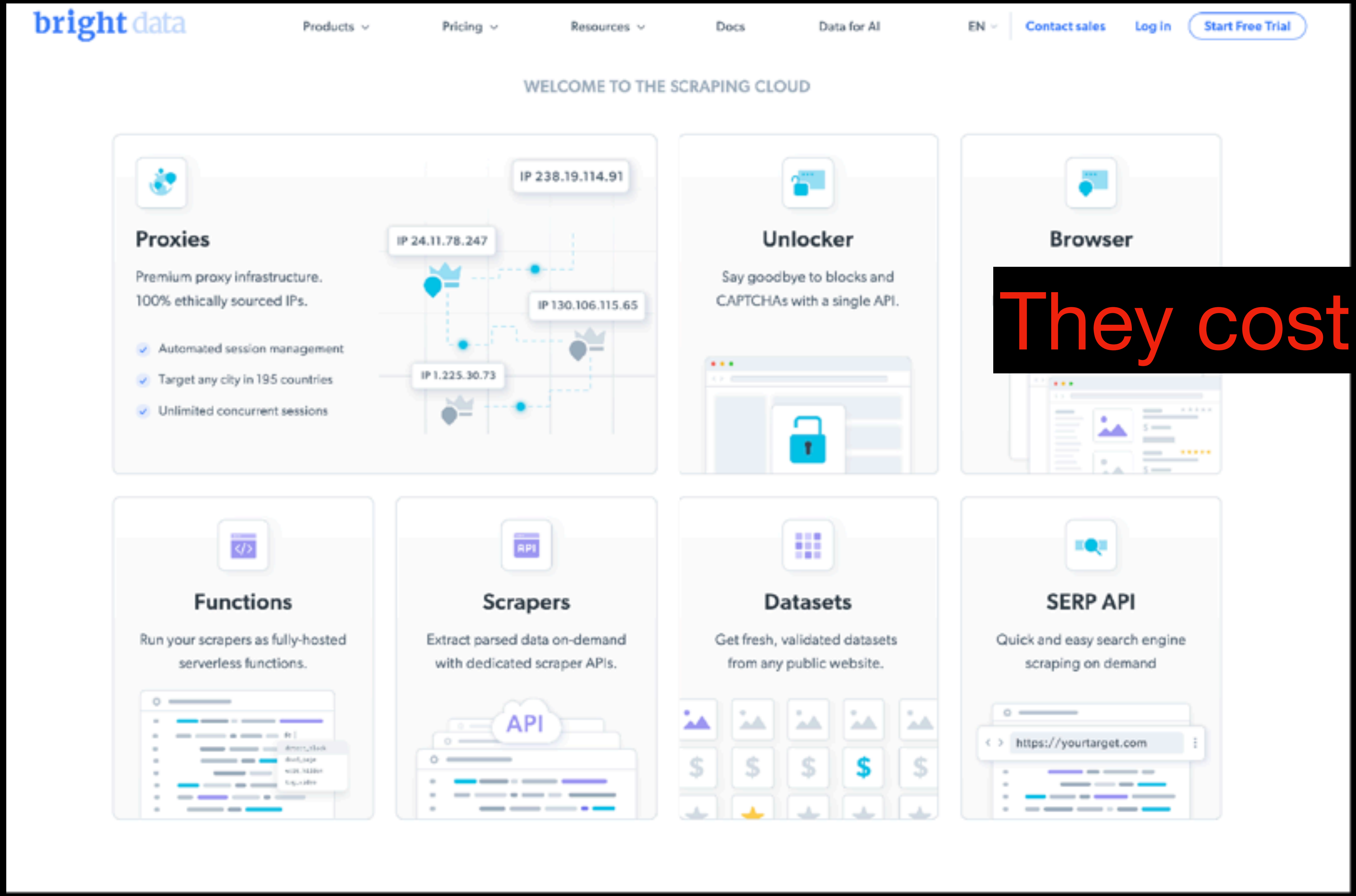
Threat Actors Targeting Cybersecurity Researchers

Last Revised: April 14, 2021

Google and Microsoft recently published reports on advanced persistent threat (APT) actors targeting cybersecurity researchers. The APT actors are using fake social media profiles and legitimate-looking websites to lure security researchers into visiting malicious websites to steal information, including exploits and zero-day vulnerabilities. APT groups often use elaborate social engineering and spear phishing schemes to trick victims into running malicious code through malicious links and websites.

Generally, you don't want to touch systems from locations that can be attributed to you or your organization.

There are services that provide access to distributed proxies that you can use for crawling



bright data

Products ⌄ Pricing ⌄ Resources ⌄ Docs Data for AI EN ⌄ Contact sales Log in Start Free Trial

WELCOME TO THE SCRAPING CLOUD

Proxies

Premium proxy infrastructure. 100% ethically sourced IPs.

- ✓ Automated session management
- ✓ Target any city in 195 countries
- ✓ Unlimited concurrent sessions

Unlocker

Say goodbye to blocks and CAPTCHAs with a single API.

Browser

Functions

Run your scrapers as fully-hosted serverless functions.

Scrapers

Extract parsed data on-demand with dedicated scraper APIs.

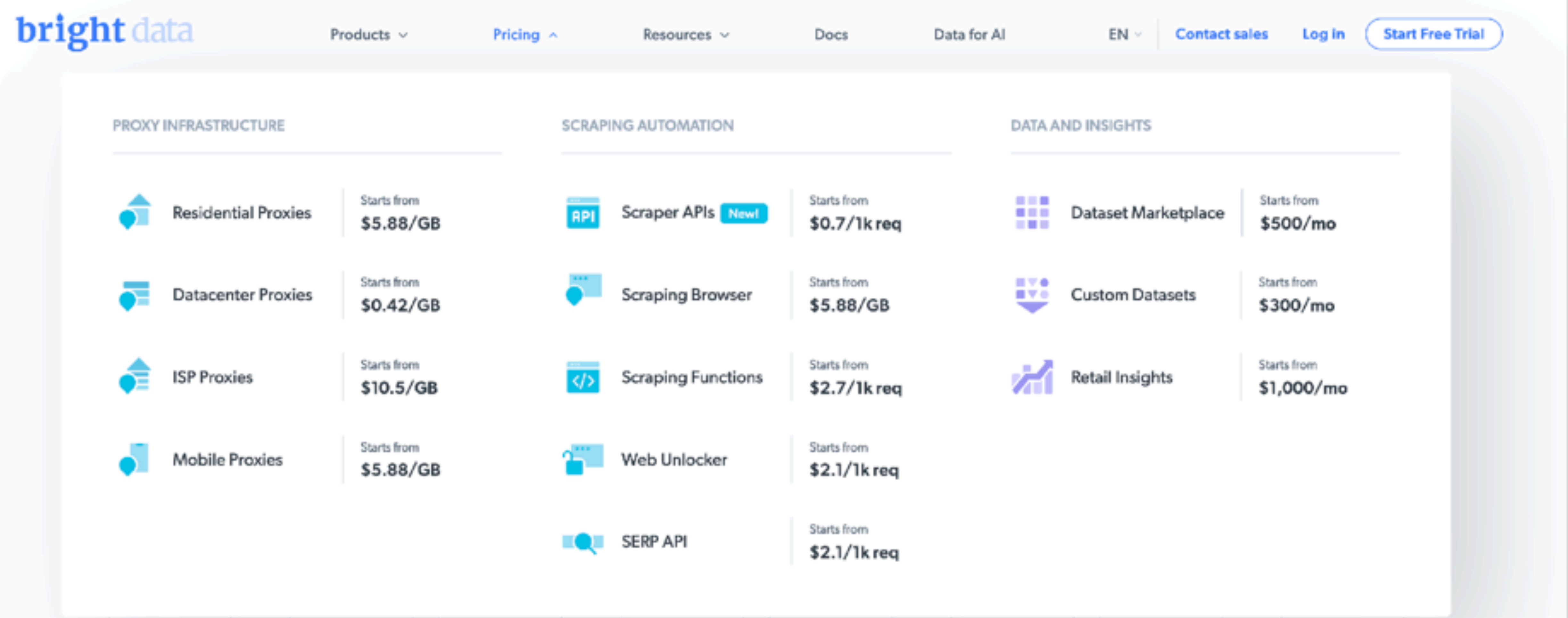
Datasets

Get fresh, validated datasets from any public website.

SERP API

Quick and easy search engine scraping on demand

They cost money (I didn't use them for this presentation)



bright data

Products ⌄ Pricing ⌃ Resources ⌄ Docs Data for AI EN ⌄ Contact sales Log in Start Free Trial

PROXY INFRASTRUCTURE

Residential Proxies	Starts from $5.88/GB
Datacenter Proxies	Starts from $0.42/GB
ISP Proxies	Starts from $10.5/GB
Mobile Proxies	Starts from $5.88/GB

SCRAPING AUTOMATION

Scraper APIs New!	Starts from $0.7/1k req
Scraping Browser	Starts from $5.88/GB
Scraping Functions	Starts from $2.7/1k req
Web Unlocker	Starts from $2.1/1k req
SERP API	Starts from $2.1/1k req

DATA AND INSIGHTS

Dataset Marketplace	Starts from $500/mo
Custom Datasets	Starts from $300/mo
Retail Insights	Starts from $1,000/mo

Your own Proxy

Privoxy - Home Page

Privoxy is a non-caching web proxy with advanced fil

controlling access, and removing ads and other obnoxious Internet junk. Privoxy has a flexible configuration and can be customized to suit individual needs and tastes. It has application for both stand-alone systems and multi-user networks.

Privoxy is Free Software and licensed under the GNU GPLv2 or later.

Privoxy is an associated project of Software in the Public Interest (SPI).

Helping hands and donations are welcome:

- https://www.privoxy.org/participate
- https://www.privoxy.org/donate

The most recent release is 3.0.34 (stable).

But I did set up my own proxies in various networks

Docker-Compose

```
privoxy:
  build:
    context: ./privoxydocker
  container_name: privoxy
  restart: always
  volumes:
    - ./privoxydocker/logs/privoxy:/var/log/privoxy
    - ./privoxydocker/privoxy.conf:/etc/privoxy/config:ro
  ports:
    - "1080:1080"
  dns:
    - 192.168.1.5
    - 192.168.1.3
```

Now that crawling is more or less sorted,
let’s move onto creating the malicious dataset
that will be used to compare unknown URLs against.

Creating Datasets

Building a Malicious Dataset

Data Sources

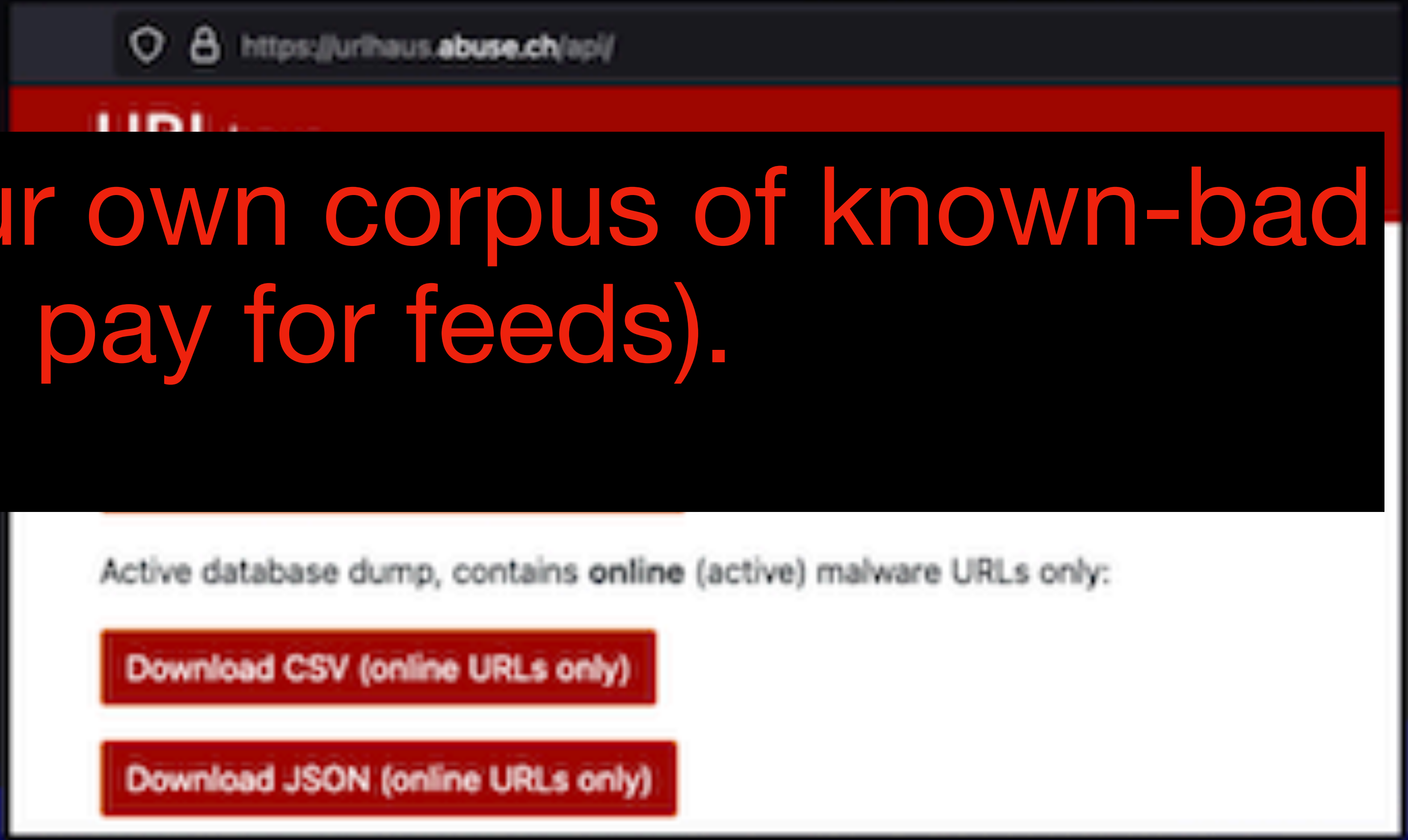
Some are better than others

URLhaus

Community Sourced

Malware Activity

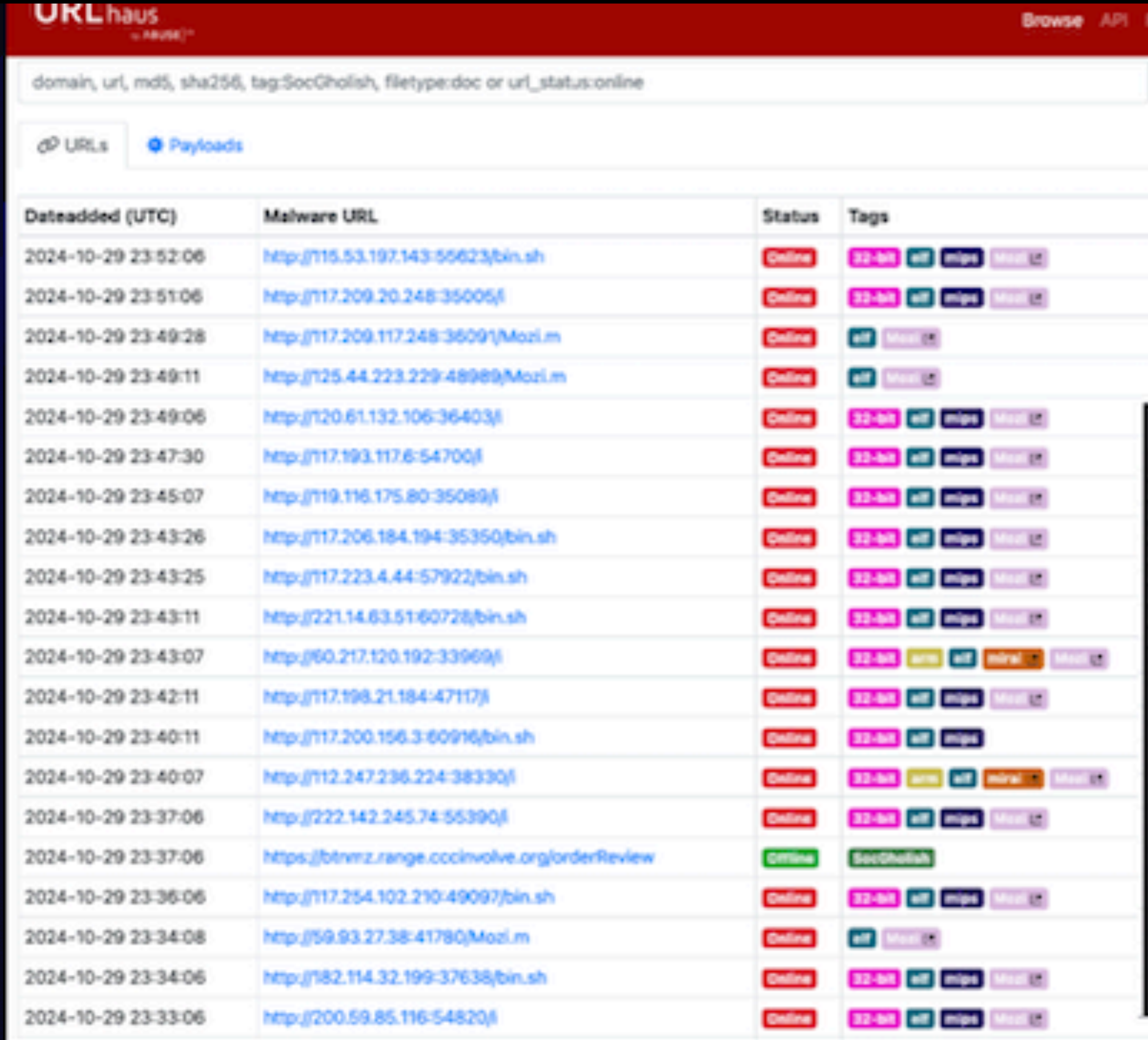
There are lots of datasources. You can use your own corpus of known-bad URLs, or you can use free services (or you can pay for feeds).
This example shows using URLHaus



https://urlhaus.abuse.ch/api/

Active database dump, contains **online** (active) malware URLs only:

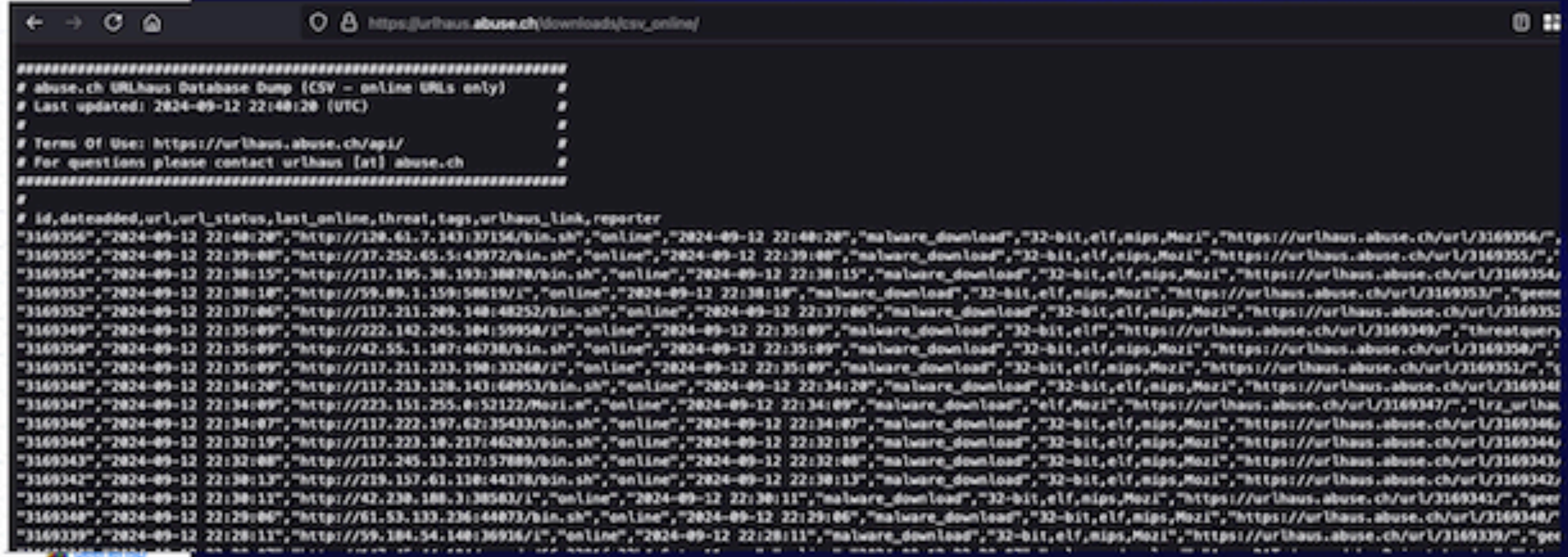
Download CSV (online URLs only)
Download JSON (online URLs only)



domain, url, md5, sha256, tag:SocGholish, filetype:doc or url_status:online

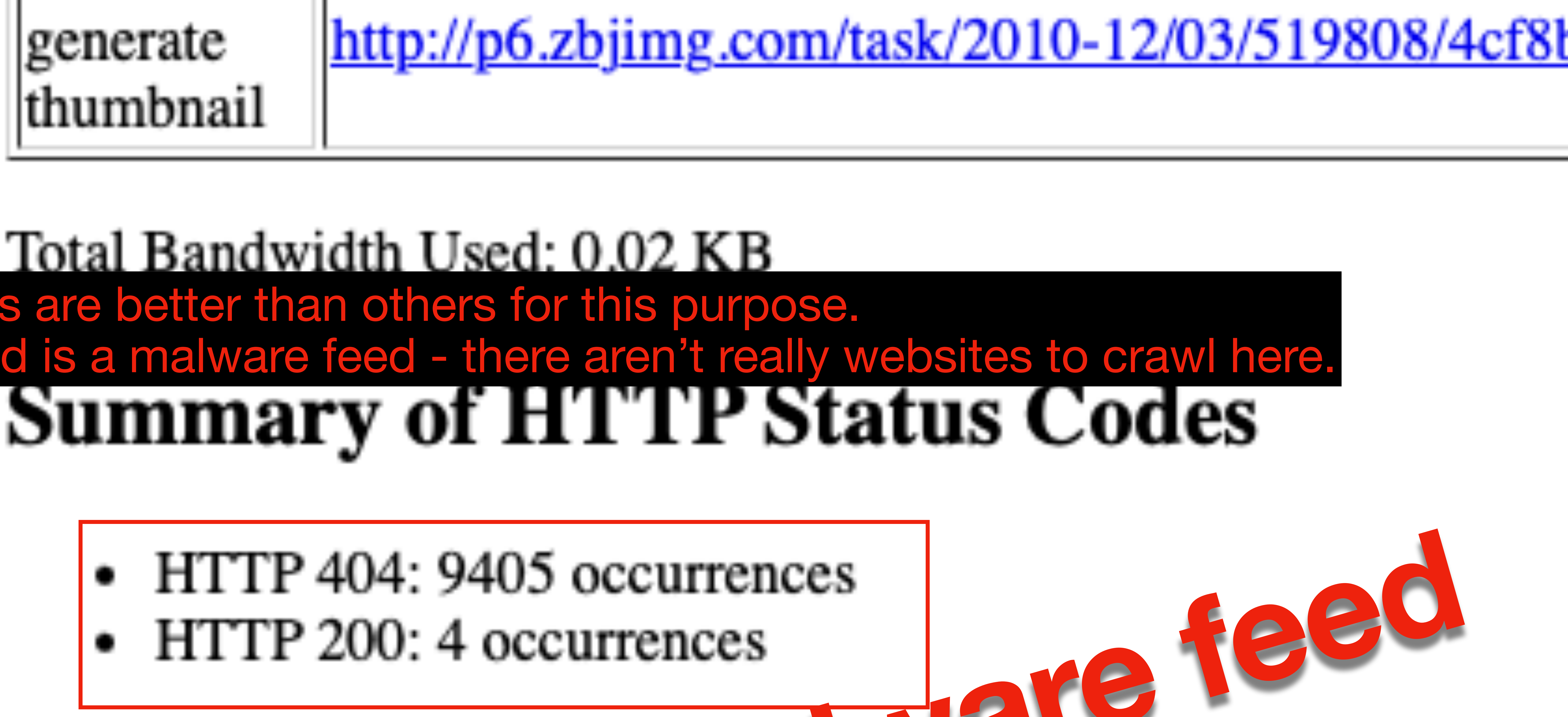
Dateadded (UTC)	Malware URL	Status	Tags	Reporter
2024-10-29 23:52:06	http://115.53.197.143:55623/bin.sh	Online	32-bit elf mips Mozi	
2024-10-29 23:51:06	http://117.209.20.248:35006/i	Online	32-bit elf mips Mozi	
2024-10-29 23:49:28	http://117.209.117.248:36091/Mozi.m	Online	elf Mozi	lrz_urlhaus
2024-10-29 23:49:11	http://125.44.223.229:48989/Mozi.m	Online	elf Mozi	lrz_urlhaus
2024-10-29 23:49:06	http://120.61.132.106:36403/i	Online	32-bit elf mips Mozi	
2024-10-29 23:47:30	http://117.193.117.6:54700/i	Online	32-bit elf mips Mozi	
2024-10-29 23:45:07	http://119.116.175.80:35089/i	Online	32-bit elf mips Mozi	
2024-10-29 23:43:26	http://117.206.184.194:35350/bin.sh	Online	32-bit elf mips Mozi	
2024-10-29 23:43:25	http://117.223.4.44:57922/bin.sh	Online	32-bit elf mips Mozi	
2024-10-29 23:43:11	http://221.14.63.51:60728/bin.sh	Online	32-bit elf mips Mozi	
2024-10-29 23:43:07	http://60.217.120.192:33969/i	Online	32-bit arm elf mirai Mozi	
2024-10-29 23:42:11	http://117.198.21.184:47117/i	Online	32-bit elf mips Mozi	
2024-10-29 23:40:11	http://117.200.156.3:60916/bin.sh	Online	32-bit elf mips	
2024-10-29 23:40:07	http://112.247.236.224:38330/i	Online	32-bit arm elf mirai Mozi	
2024-10-29 23:37:06	http://222.142.245.74:55390/i	Online	32-bit elf mips Mozi	
2024-10-29 23:37:06	https://btnmz.range.cccinvolve.org/orderReview	Offline	SocGholish	
2024-10-29 23:36:06	http://117.254.102.210:49097/bin.sh	Online	32-bit elf mips Mozi	
2024-10-29 23:34:08	http://59.93.27.38:41780/Mozi.m	Online	elf Mozi	
2024-10-29 23:34:06	http://182.114.32.199:37638/bin.sh	Online	32-bit elf mips Mozi	
2024-10-29 23:33:06	http://200.59.85.116:54820/i	Online	32-bit elf mips Mozi	

You can often grab a CSV, json, or other file with URLs



https://urlhaus.abuse.ch/downloads/csv_online/

```
################################################################
# abuse.ch URLhaus Database Dump (CSV - online URLs only)      #
# Last updated: 2024-09-12 22:40:20 (UTC)                      #
#                                                              #
# Terms Of Use: https://urlhaus.abuse.ch/api/                  #
# For questions please contact urlhaus [at] abuse.ch           #
################################################################
#
# id,dateadded,url,url_status,last_online,threat,tags,urlhaus_link,reporter
"3169356","2024-09-12 22:40:20","http://120.61.7.143:37156/bin.sh","online","2024-09-12 22:40:20","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169356/",
"3169355","2024-09-12 22:39:00","http://37.252.65.5:43972/bin.sh","online","2024-09-12 22:39:00","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169355/","
"3169354","2024-09-12 22:38:15","http://117.195.38.193:38070/bin.sh","online","2024-09-12 22:38:15","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169354/
"3169353","2024-09-12 22:38:10","http://59.89.1.159:58619/i","online","2024-09-12 22:38:10","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169353/","geene
"3169352","2024-09-12 22:37:06","http://117.211.209.140:48252/bin.sh","online","2024-09-12 22:37:06","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/316935
"3169349","2024-09-12 22:35:09","http://222.142.245.104:59950/i","online","2024-09-12 22:35:09","malware_download","32-bit,elf","https://urlhaus.abuse.ch/url/3169349/","threatquery
"3169350","2024-09-12 22:35:09","http://42.55.1.107:46738/bin.sh","online","2024-09-12 22:35:09","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169350/","
"3169351","2024-09-12 22:35:09","http://117.211.233.190:33260/i","online","2024-09-12 22:35:09","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169351/","
"3169348","2024-09-12 22:34:20","http://117.213.120.143:60953/bin.sh","online","2024-09-12 22:34:20","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/316934
"3169347","2024-09-12 22:34:09","http://223.151.255.0:52122/Mozi.m","online","2024-09-12 22:34:09","malware_download","elf,Mozi","https://urlhaus.abuse.ch/url/3169347/","lrz_urlhau
"3169346","2024-09-12 22:34:07","http://117.222.197.62:35433/bin.sh","online","2024-09-12 22:34:07","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169346/
"3169344","2024-09-12 22:32:19","http://117.223.10.217:46203/bin.sh","online","2024-09-12 22:32:19","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169344/
"3169343","2024-09-12 22:32:08","http://117.245.13.217:57889/bin.sh","online","2024-09-12 22:32:08","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169343/
"3169342","2024-09-12 22:30:13","http://219.157.61.110:44178/bin.sh","online","2024-09-12 22:30:13","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169342/
"3169341","2024-09-12 22:30:11","http://42.230.188.3:38583/i","online","2024-09-12 22:30:11","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169341/","geer
"3169340","2024-09-12 22:29:06","http://61.53.133.236:44073/bin.sh","online","2024-09-12 22:29:06","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169340/
"3169339","2024-09-12 22:28:11","http://59.104.54.140:36916/i","online","2024-09-12 22:28:11","malware_download","32-bit,elf,mips,Mozi","https://urlhaus.abuse.ch/url/3169339/","gee
```



generate thumbnail	http://p6.zbjimg.com/task/2010-12/03/519808/4cf8

Total Bandwidth Used: 0.02 KB

Some feeds are better than others for this purpose.
urlhaus feed is a malware feed - there aren’t really websites to crawl here.

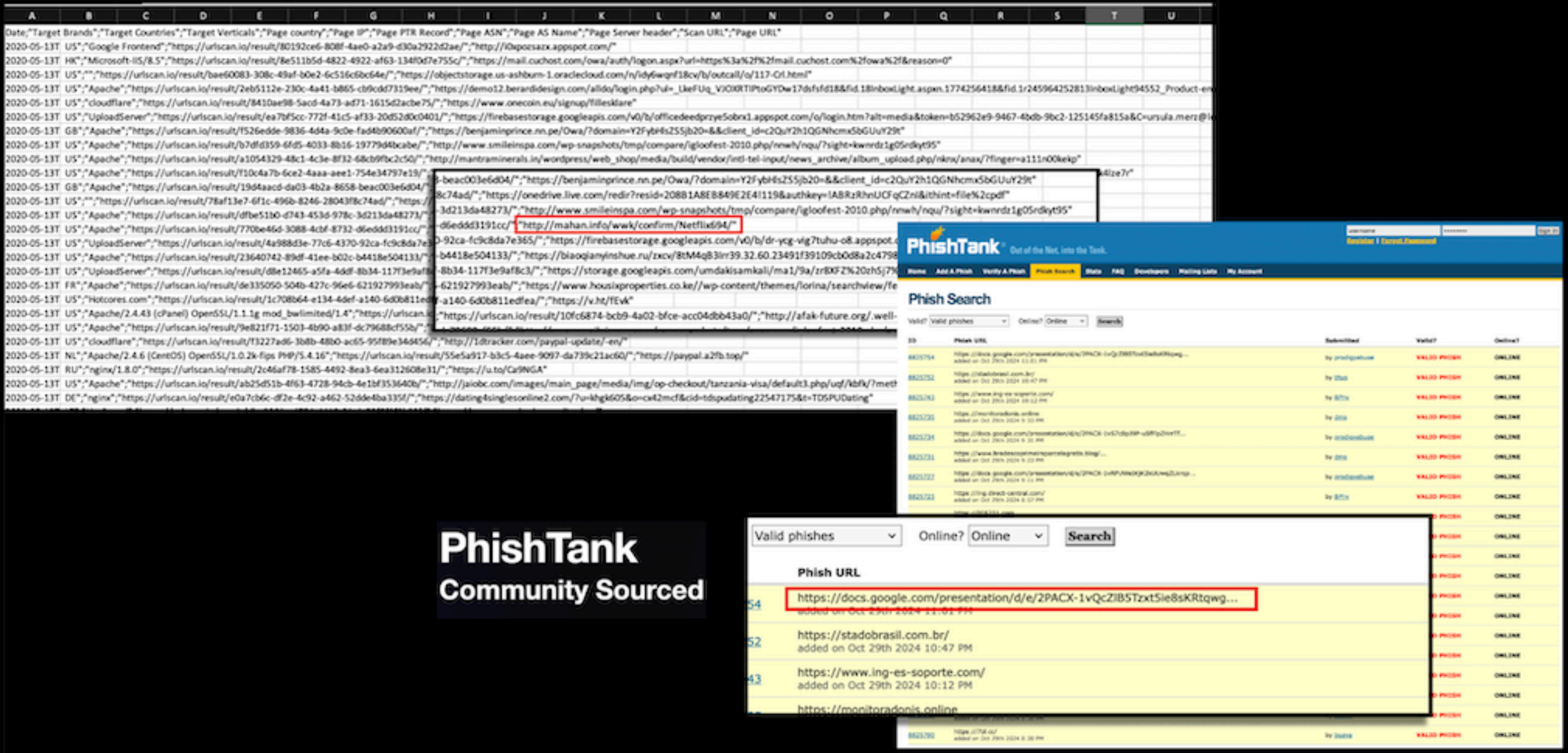
Summary of HTTP Status Codes

- HTTP 404: 9405 occurrences
- HTTP 200: 4 occurrences

This is a malware feed

Other feeds are better, such as urlscan.io and phishintank phishing feeds

URLScan.io



```
Date;"Target Brands";"Target Countries";"Target Verticals";"Page country";"Page IP";"Page PTR Record";"Page ASN";"Page AS Name";"Page Server header";"Scan URL";"Page URL"
2020-05-13T  US";"Google Frontend";"https://urlscan.io/result/80192ce6-808f-4ae0-a2a9-d30a2922d2ae/";"http://i0xpozsazx.appspot.com/"
2020-05-13T  HK";"Microsoft-IIS/8.5";"https://urlscan.io/result/8e511b5d-4822-4922-af63-134f0d7e755c/";"https://mail.cuchost.com/owa/auth/logon.aspx?url=https%3a%2f%2fmail.cuchost.com%2fowa%2f&reason=0"
2020-05-13T  US";"";"https://urlscan.io/result/bae60083-308c-49af-b0e2-6c516c6bc64e/";"https://objectstorage.us-ashburn-1.oraclecloud.com/n/idy6wqnf18cv/b/outcall/o/117-Crl.html"
2020-05-13T  US";"Apache";"https://urlscan.io/result/2eb5112e-230c-4a41-b865-cb9cdd7319ee/";"https://demo12.berardidesign.com/alldo/login.php?ul=_LkeFUq_VJOXRTIPtoGYDw17dsfsfd18&fid.18InboxLight.aspxn.1774256418&fid.1r245964252813InboxLight94552_Product-en
2020-05-13T  US";"cloudflare";"https://urlscan.io/result/8410ae98-5acd-4a73-ad71-1615d2acbe75/";"https://www.onecoin.eu/signup/filleskiare"
2020-05-13T  US";"UploadServer";"https://urlscan.io/result/ea7bf5cc-772f-41c5-af33-20d52d0c0401/";"https://firebasestorage.googleapis.com/v0/b/officedeedprzyeSobrx1.appspot.com/o/login.htm?alt=media&token=b52962e9-9467-4bdb-9bc2-125145fa815a&C=ursula.merz@l
2020-05-13T  GB";"Apache";"https://urlscan.io/result/f526edde-9836-4d4a-9c0e-fad4b90600af/";"https://benjaminprince.nn.pe/Owa/?domain=Y2FybHls255jb20=&&client_id=c2QuY2h1QGNhcmx5bGUuY29t"
2020-05-13T  US";"Apache";"https://urlscan.io/result/b7dfd359-6fd5-4033-8b16-19779d4bcabe/";"http://www.smileinspa.com/wp-snapshots/tmp/compare/igloofest-2010.php/nmwh/nqu/?sight=kwnrdz1g05rdkyt95"
2020-05-13T  US";"Apache";"https://urlscan.io/result/a1054329-48c1-4c3e-8f32-68cb9fbc2c50/";"http://mantraminerals.in/wordpress/web_shop/media/build/vendor/intl-tel-input/news_archive/album_upload.php/nkra/anax/?finger=a111n00kekp"
2020-05-13T  US";"Apache";"https://urlscan.io/result/f10c4a7b-6ce2-4aaa-aee1-754e34797e19/";
2020-05-13T  GB";"Apache";"https://urlscan.io/result/19d4aacd-da03-4b2a-8658-beac003e6d04/";
2020-05-13T  US";"";"https://urlscan.io/result/78af13e7-6f1c-496b-8246-28043f8c74ad/";"https://
2020-05-13T  US";"Apache";"https://urlscan.io/result/dfbe51b0-d743-453d-978c-3d213da48273/";
2020-05-13T  US";"Apache";"https://urlscan.io/result/770be46d-3088-4cbf-8732-d6eddd3191cc/";
2020-05-13T  US";"UploadServer";"https://urlscan.io/result/4a988d3e-77c6-4370-92ca-fc9c8da7e
2020-05-13T  US";"Apache";"https://urlscan.io/result/23640742-89df-41ee-b02c-b4418e504133/";
2020-05-13T  US";"UploadServer";"https://urlscan.io/result/d8e12465-a5fa-4ddf-8b34-117f3e9af8
2020-05-13T  FR";"Apache";"https://urlscan.io/result/de335050-504b-427c-96e6-621927993eab/";
2020-05-13T  US";"Hotcores.com";"https://urlscan.io/result/1c708b64-e134-4def-a140-6d0b811ed
2020-05-13T  US";"Apache/2.4.43 (cPanel) OpenSSL/1.1.1g mod_bwlimited/1.4";"https://urlscan.io
2020-05-13T  US";"Apache";"https://urlscan.io/result/9e821f71-1503-4b90-a83f-dc79688cf55b/";
2020-05-13T  US";"cloudflare";"https://urlscan.io/result/f3227ad6-3b8b-48b0-ac65-95f89e34d456/";"http://1dtracker.com/paypal-update/-en/"
2020-05-13T  NL";"Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips PHP/5.4.16";"https://urlscan.io/result/55e5a917-b3c5-4aee-9097-da739c21ac60/";"https://paypal.a2fb.top/"
2020-05-13T  RU";"nginx/1.8.0";"https://urlscan.io/result/2c46af78-1585-4492-8ea3-6ea312608e31/";"https://u.to/Ca9NGA"
2020-05-13T  US";"Apache";"https://urlscan.io/result/ab25d51b-4f63-4728-94cb-4e1bf353640b/";"http://jaiobc.com/images/main_page/media/img/op-checkout/tanzania-visa/default3.php/uqf/kbfk/?meth
2020-05-13T  DE";"nginx";"https://urlscan.io/result/e0a7cb6c-df2e-4c92-a462-52dde4ba335f/";"https://dating4singlesonline2.com/?u=khgk605&o=cx42mcf&cid=tdspudating22547175&t=TDSPUDating"
```

```
3-beac003e6d04/";"https://benjaminprince.nn.pe/Owa/?domain=Y2FybHls255jb20=&&client_id=c2QuY2h1QGNhcmx5bGUuY29t"
8c74ad/";"https://onedrive.live.com/redir?resid=208B1A8EB849E2E4!119&authkey=!ABRzRhnUCFqCZnI&ithint=file%2cpdf"
-3d213da48273/";"http://www.smileinspa.com/wp-snapshots/tmp/compare/igloofest-2010.php/nmwh/nqu/?sight=kwnrdz1g05rdkyt95"
-d6eddd3191cc/";"http://mahan.info/wwk/confirm/Netflix694/"
0-92ca-fc9c8da7e365/";"https://firebasestorage.googleapis.com/v0/b/dr-ycg-vig7tuhu-o8.appspot.c
-b4418e504133/";"https://biaoqianyinshue.ru/zxcv/8tM4qB3lrr39.32.60.23491f39109cb0d8a2c4798
-8b34-117f3e9af8c3/";"https://storage.googleapis.com/umdakisamkali/ma1/9a/zrBXFZ%20zh5j7%
-621927993eab/";"https://www.housixproperties.co.ke//wp-content/themes/lorina/searchview/fe
-a140-6d0b811edfea/";"https://v.ht/fEvk"
;"https://urlscan.io/result/10fc6874-bcb9-4a02-bfce-acc04dbb43a0/";"http://afak-future.org/.well-
```

PhishTank
Community Sourced

Valid phishes ▾ Online? Online ▾ Search

Phish URL

54 https://docs.google.com/presentation/d/e/2PACX-1vQcZlB5TzxtSie8sKRtqwg...
added on Oct 29th 2024 11:01 PM

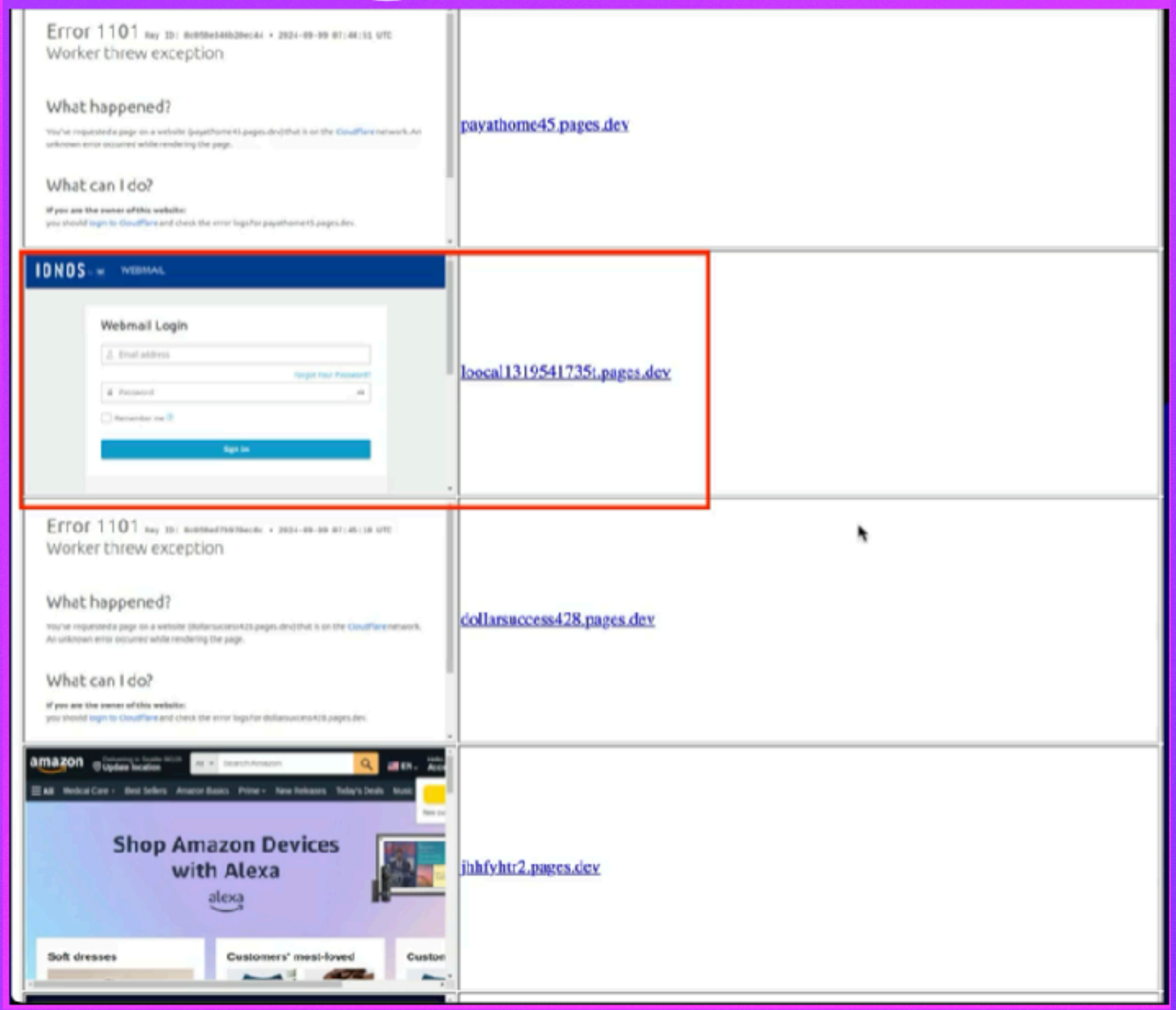
52 https://stadobrasil.com.br/
added on Oct 29th 2024 10:47 PM

43 https://www.ing-es-soporte.com/
added on Oct 29th 2024 10:12 PM

https://monitoradonis.online

Once you have your list of URLs and you’ve crawled them to get screenshots, you may notice that a lot of screenshots are not worthy of being used as malicious screenshot to compare other screenshots to. At this point, you have to go through the screenshots, removing anything that doesn’t apply.

Removing non-relevant



payathome45.pages.dev

loocal1319541735t.pages.dev

dollarsuccess428.pages.dev

jhhfyhtr2.pages.dev

In this example of screenshots, some are good for a malicious dataset, but you can see that others aren’t so great.

To quickly remove screenshots that won’t work, I will introduce some of the main ideas behind locating similar screenshots. The information in this portion of the presentation will be used in various ways throughout the many different similarity analysis Techniques.

Distance & Similarity

The primary algorithm and technique behind all similarity detection is Levenshtein. I’ll be demonstrating using it as well as various hashing comparison mechanisms that use it under the hood of their own processes.

Levenshtein Distance

Levenshtein is used best on short strings of text.
In this example, how many changes to you need to make
to turn the word ‘bats’ into the word ‘cats’. It’s just one letter, so
this would be a Levenshtein distance of 1

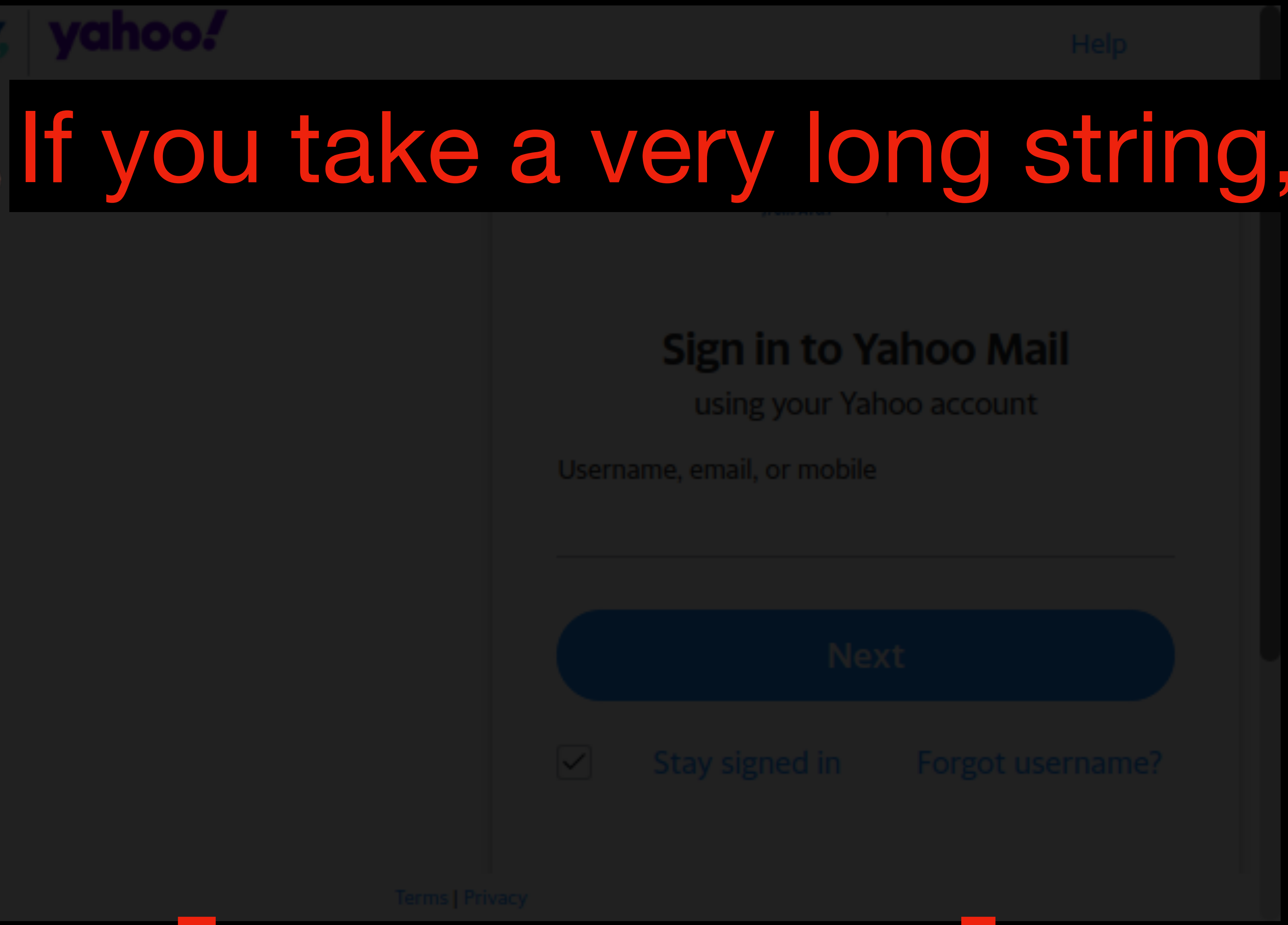
BATS

↓

CATS

Levenshtein Distance of 1

If you take a very long string, you will get a much larger Levenshtein distance.



Currently. yahoo!

Help

Sign in to Yahoo Mail

using your Yahoo account

Username, email, or mobile

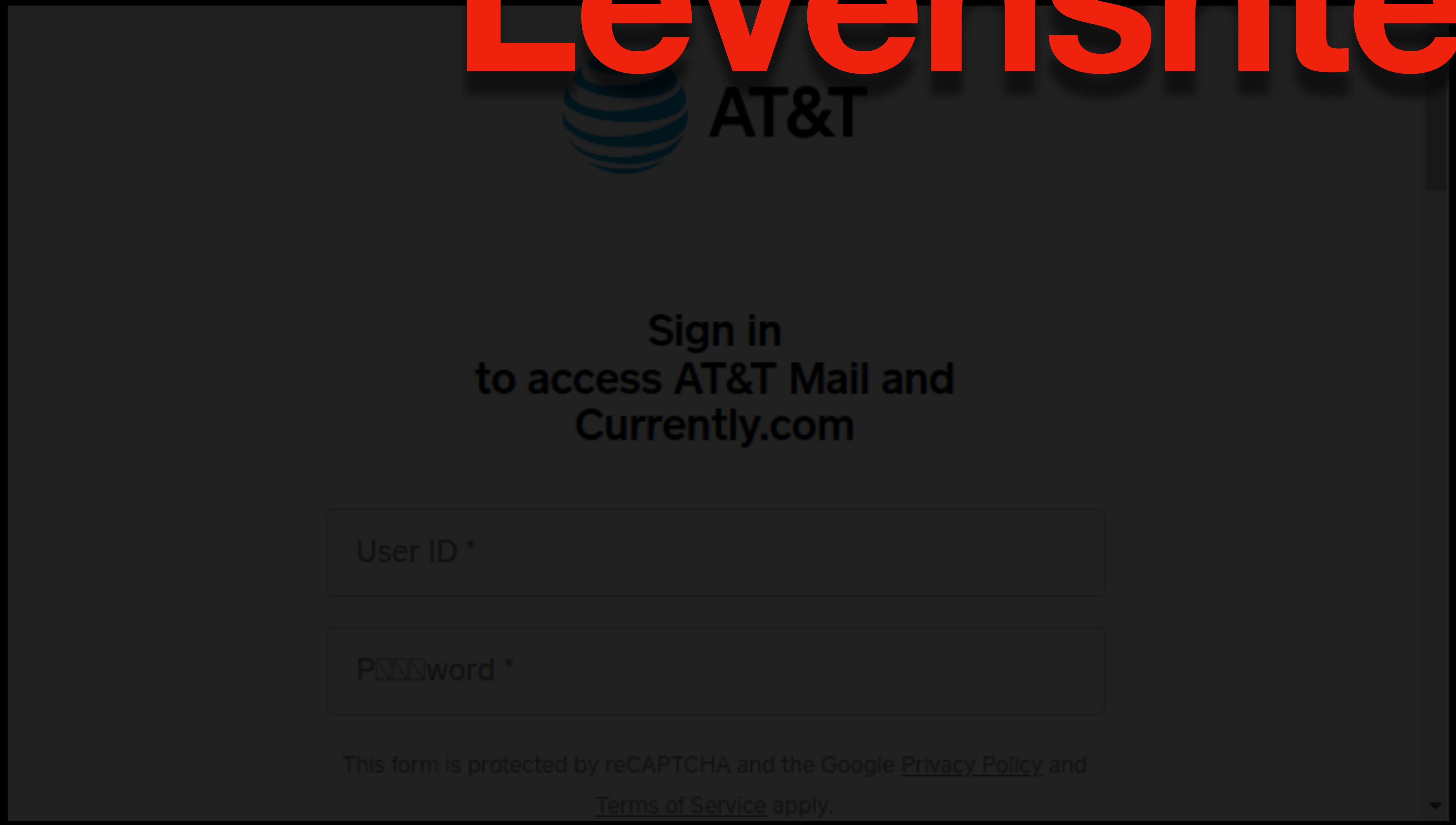
Next

☑ Stay signed in Forgot username?

Terms | Privacy

yahoo.com gmail.com outlook.com aol.com
Stay signed in
Forgot username?
Create an account
Yahoo makes it easy to enjoy what matters most in your world.
Best in class Yahoo Mail, breaking local, national and global news, finance, sports, music, movies and more
Enter password to finish sign in
Password Next
TermsPrivacy
x
Yahoo works best with the latest versions of the browsers. You're using an outdated or unsupported brows
browser version now.
More Info

Levenshtein Distance of 717



AT&T

Sign in
to access AT&T Mail and
Currently.com

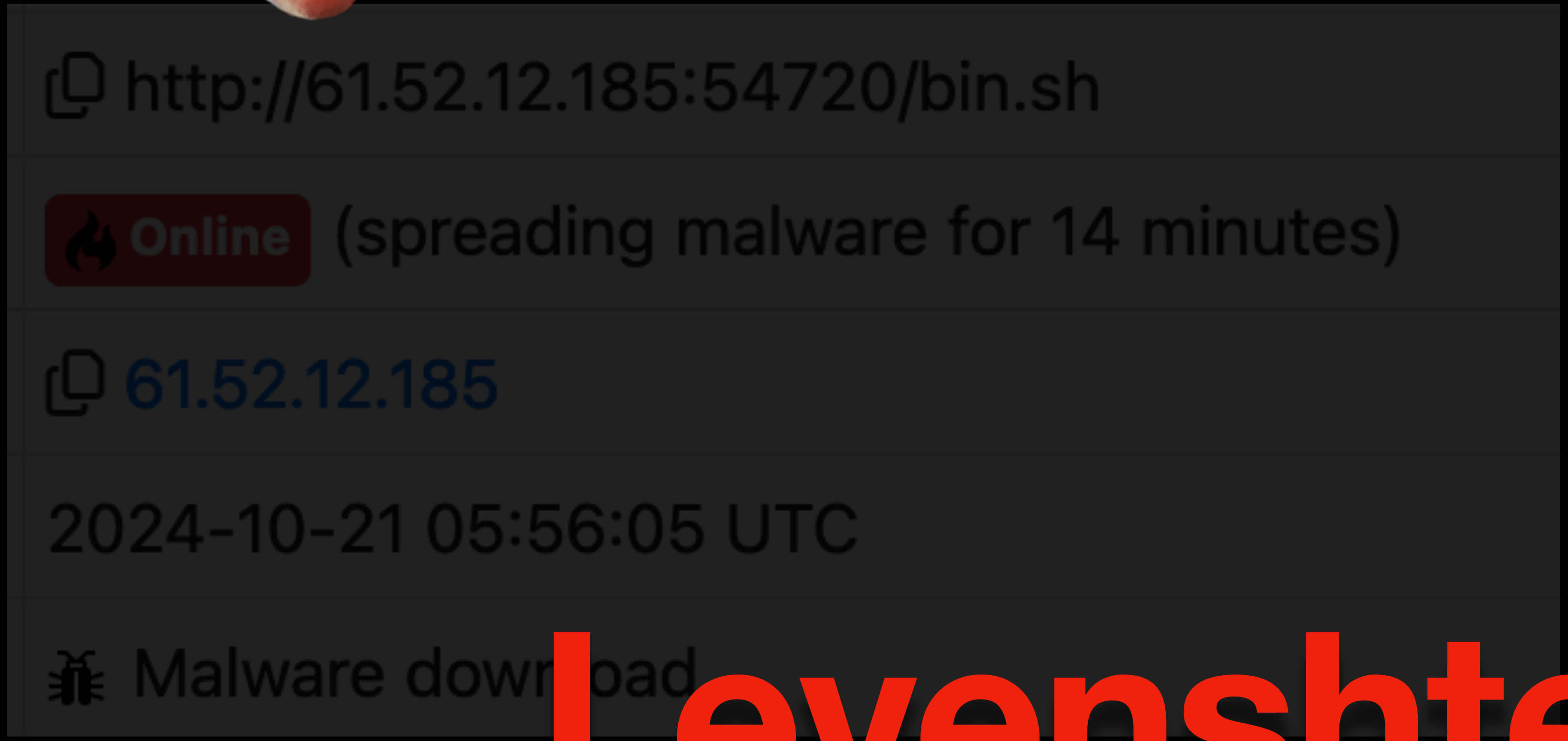
User ID *

P⊠⊠⊠word *

This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

Processing file: ./009-288-49.weeblysite.com.txt
Home | 009-288-49
Shopping Cart
You don't have any items in your cart.
Checkout
Continue Shopping
Accepted here
Sign into access AT&T Mail andCurrently.com
This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.
Sign in
Back to Cart
009-288-49
Secure checkout by Square

If you take a shorter string, like these two URLs, you get a smaller one



http://61.52.12.185:54720/bin.sh

Online (spreading malware for 14 minutes)

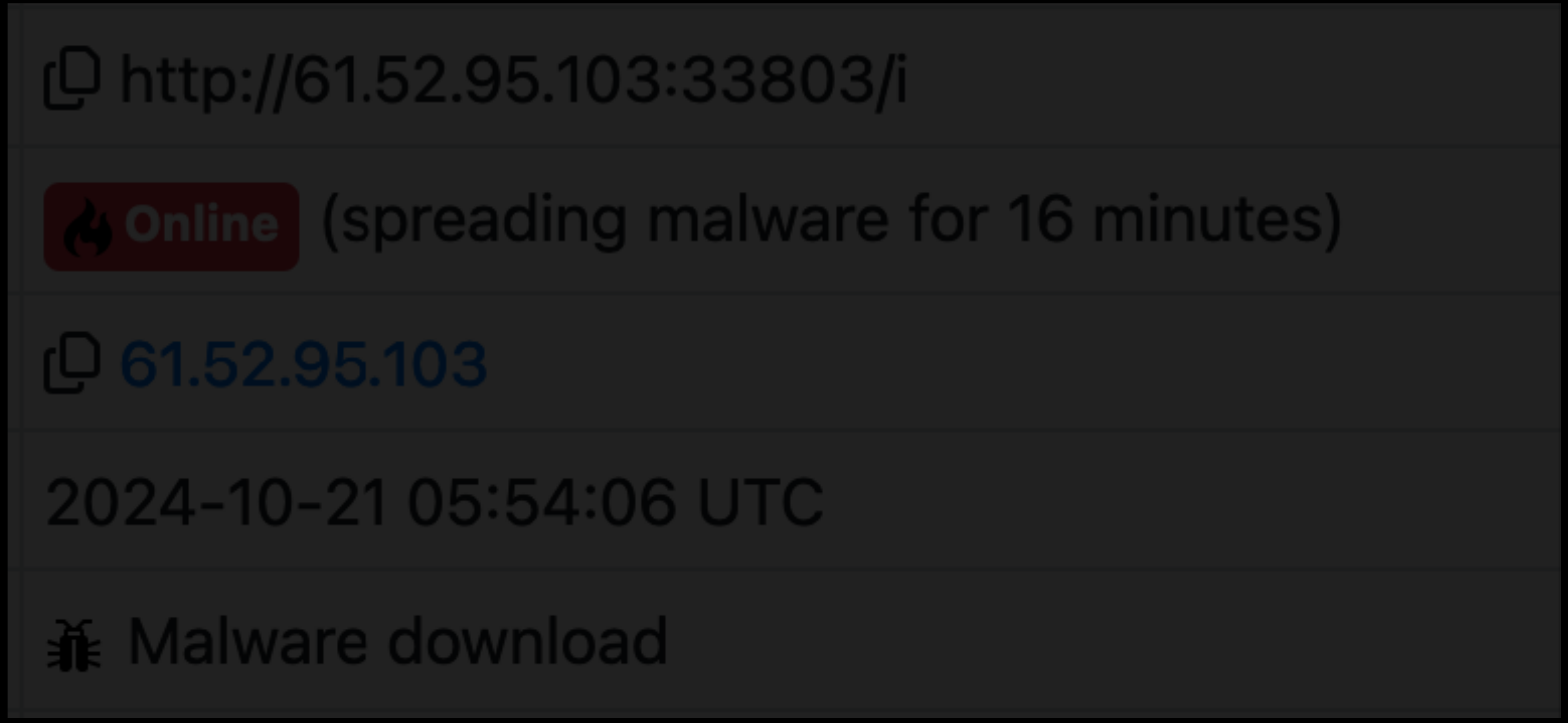
61.52.12.185

2024-10-21 05:56:05 UTC

Malware download

http://61.52.12.185:54720/bin.sh

Levenshtein Distance of 14



http://61.52.95.103:33803/i

Online (spreading malware for 16 minutes)

61.52.95.103

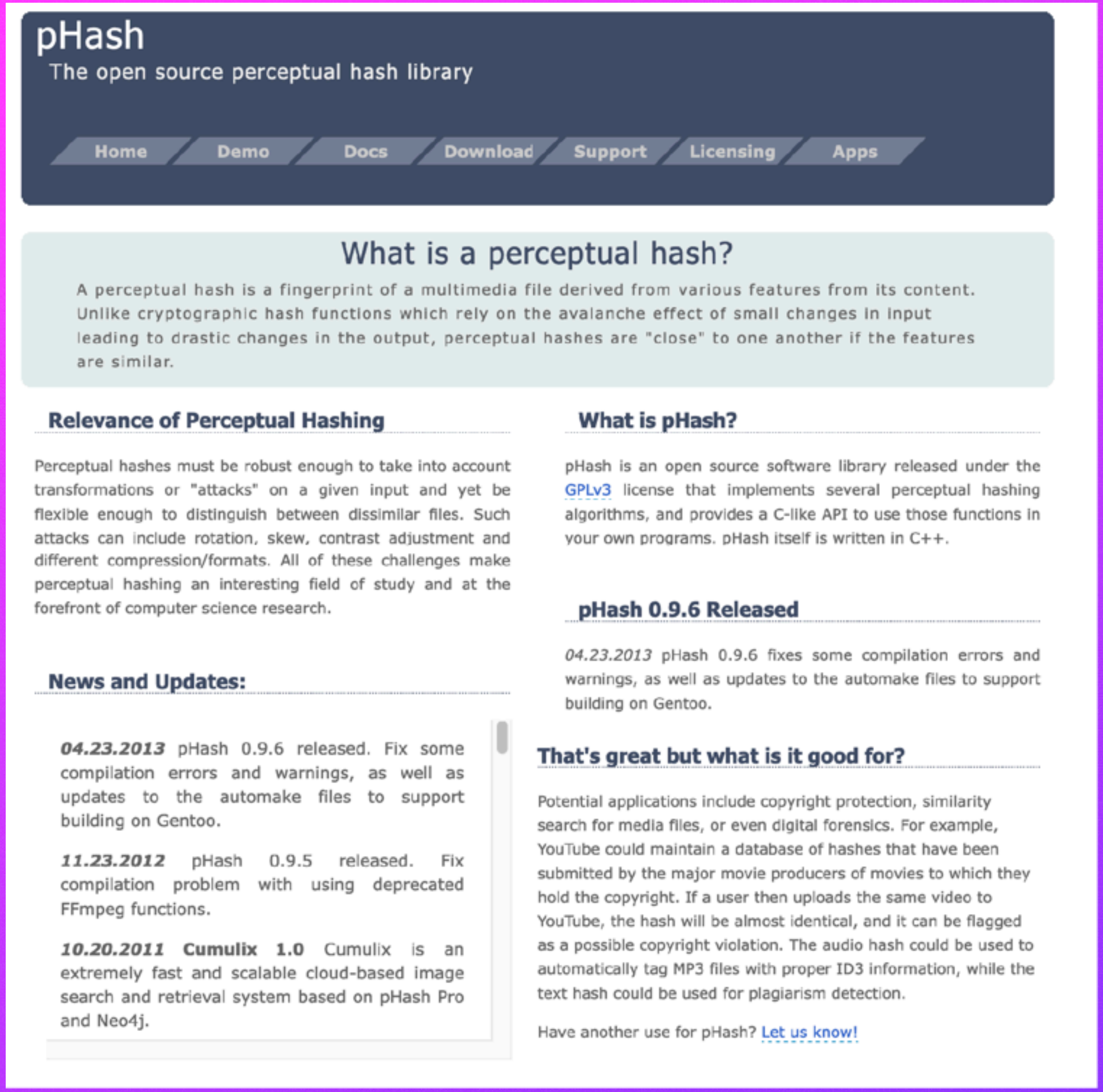
2024-10-21 05:54:06 UTC

Malware download

http://61.52.95.103:33803/i

Using PHASH To Classify Images

Before continuing with Levenshtein, I want to talk about a hashing algorithm designed for images called PHASH, then we will apply Levenshtein to it.



pHash

The open source perceptual hash library

What is a perceptual hash?

A perceptual hash is a fingerprint of a multimedia file derived from various features from its content. Unlike cryptographic hash functions which rely on the avalanche effect of small changes in input leading to drastic changes in the output, perceptual hashes are "close" to one another if the features are similar.

Relevance of Perceptual Hashing

Perceptual hashes must be robust enough to take into account transformations or "attacks" on a given input and yet be flexible enough to distinguish between dissimilar files. Such attacks can include rotation, skew, contrast adjustment and different compression/formats. All of these challenges make perceptual hashing an interesting field of study and at the forefront of computer science research.

News and Updates:

04.23.2013 pHash 0.9.6 released. Fix some compilation errors and warnings, as well as updates to the automake files to support building on Gentoo.

11.23.2012 pHash 0.9.5 released. Fix compilation problem with using deprecated FFmpeg functions.

10.20.2011 **Cumulix 1.0** Cumulix is an extremely fast and scalable cloud-based image search and retrieval system based on pHash Pro and Neo4j.

What is pHash?

pHash is an open source software library released under the GPLv3 license that implements several perceptual hashing algorithms, and provides a C-like API to use those functions in your own programs. pHash itself is written in C++.

pHash 0.9.6 Released

04.23.2013 pHash 0.9.6 fixes some compilation errors and warnings, as well as updates to the automake files to support building on Gentoo.

That's great but what is it good for?

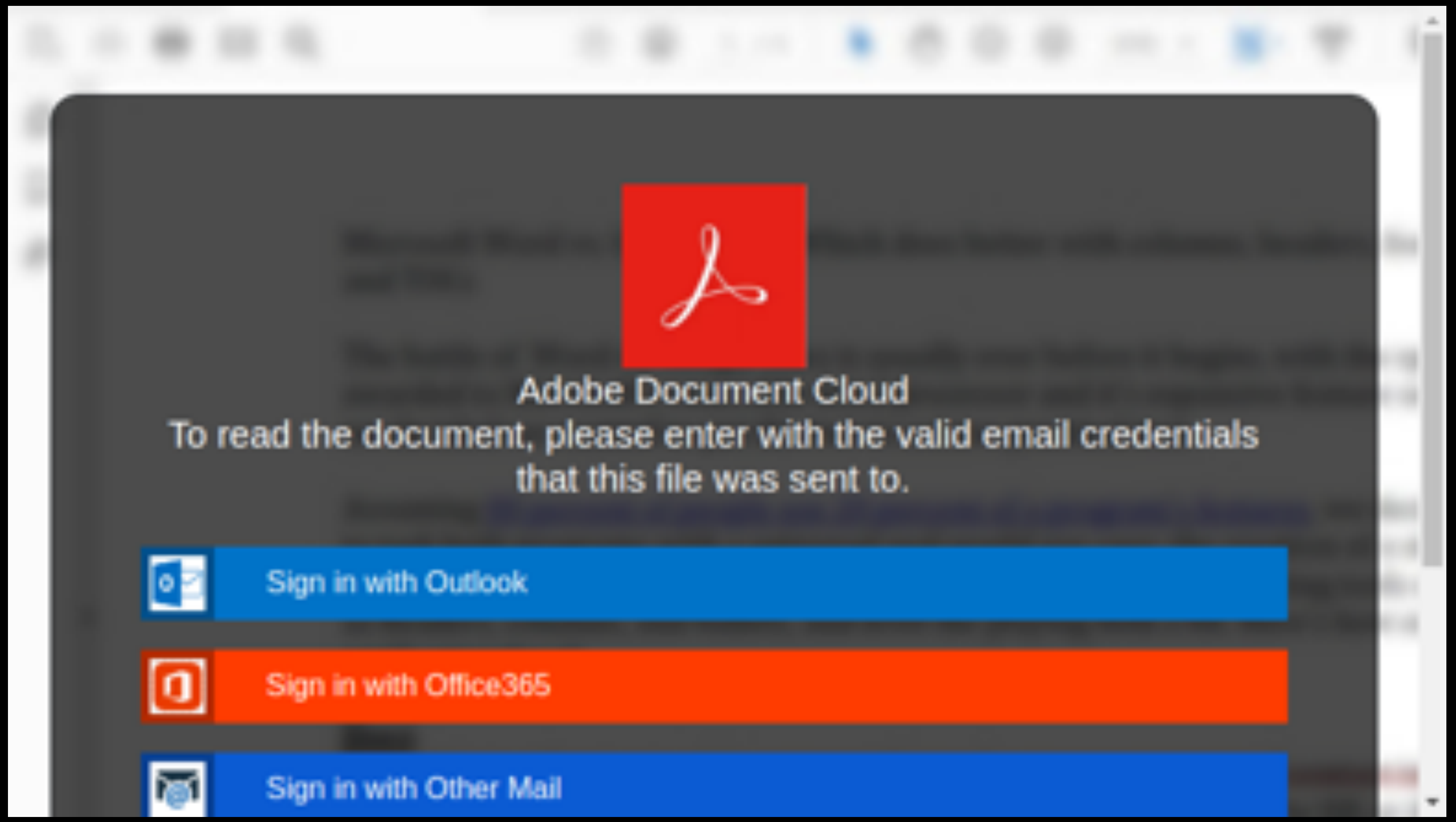
Potential applications include copyright protection, similarity search for media files, or even digital forensics. For example, YouTube could maintain a database of hashes that have been submitted by the major movie producers of movies to which they hold the copyright. If a user then uploads the same video to YouTube, the hash will be almost identical, and it can be flagged as a possible copyright violation. The audio hash could be used to automatically tag MP3 files with proper ID3 information, while the text hash could be used for plagiarism detection.

Have another use for pHash? Let us know!

Using **PHASH** To Classify Images

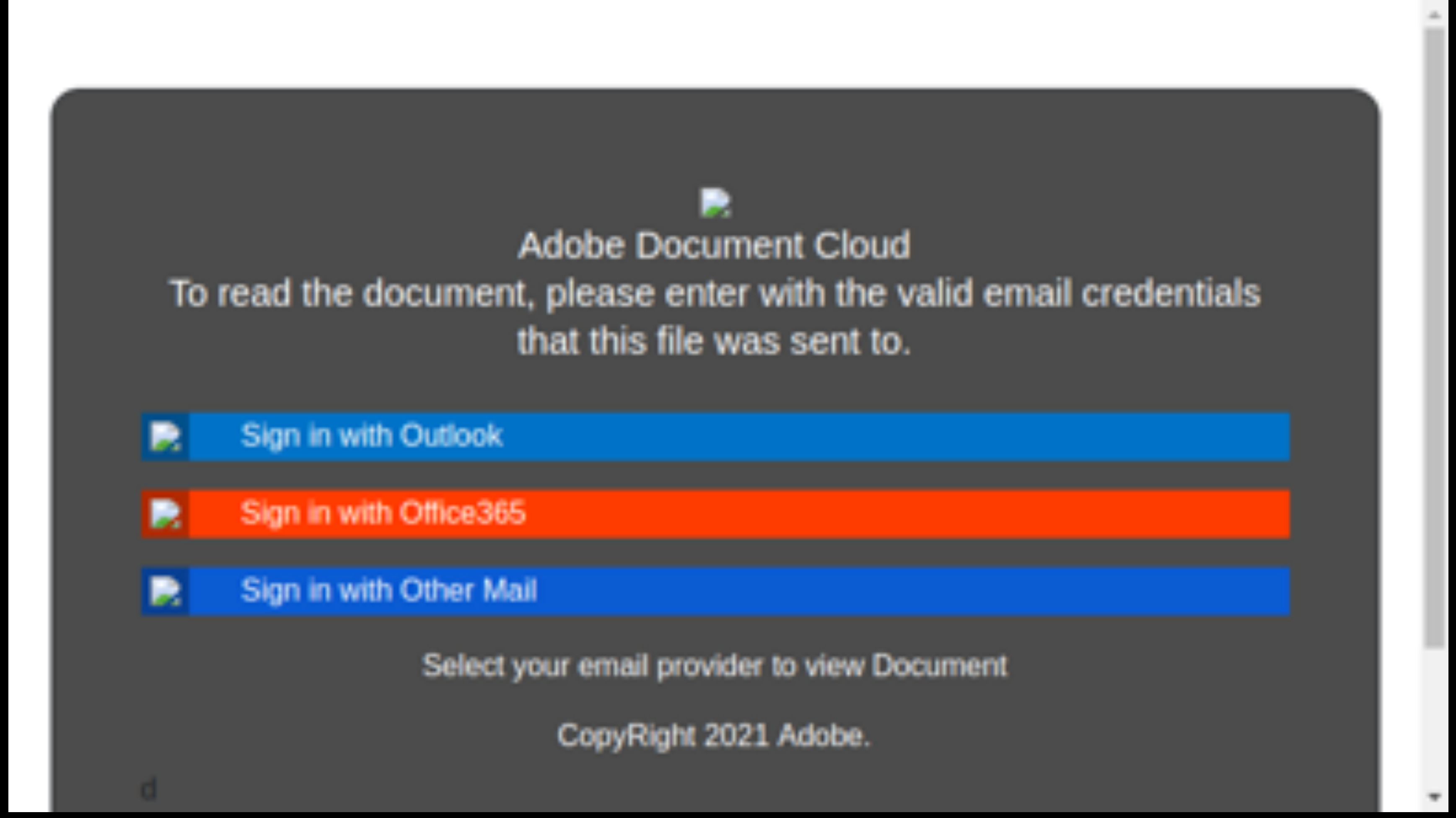
These are NOT the same but they are similar

PHASH is different than SHA256.
SHA256 makes a longer hash, but a very small change in the hash means a drastically different file.



Adobe Document Cloud
To read the document, please enter with the valid email credentials that this file was sent to.

SHA256:
84926feadda7dfeb59777ab0e4b4cc60a977b6e1



Adobe Document Cloud
To read the document, please enter with the valid email credentials that this file was sent to.

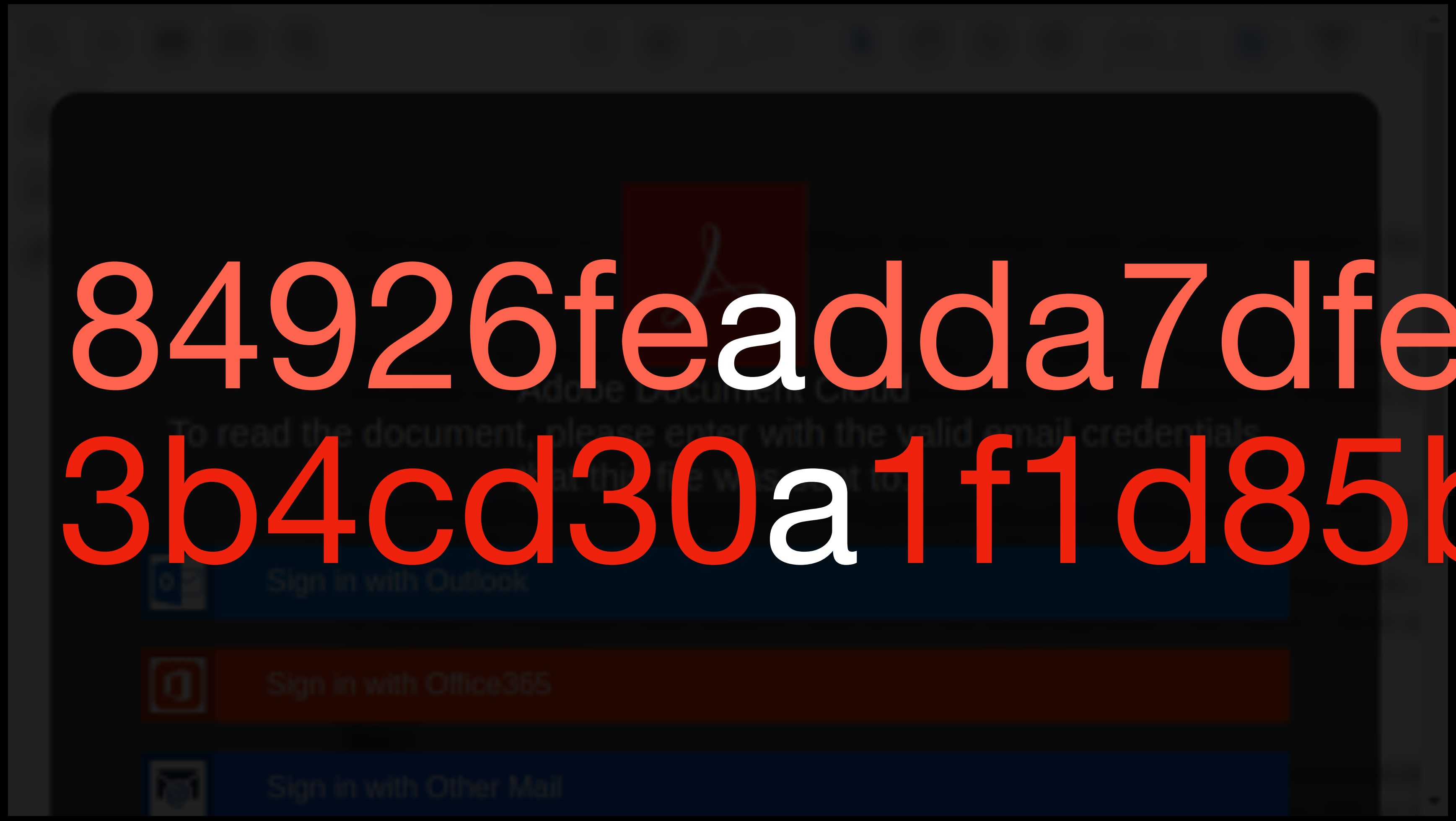
Select your email provider to view Document

CopyRight 2021 Adobe.

SHA256:
3b4cd30a1f1d85b28601641afde04375169854f8

PHASH is different than SHA256.
SHA256 makes a longer hash, but a very small change in the hash means a drastically different file.

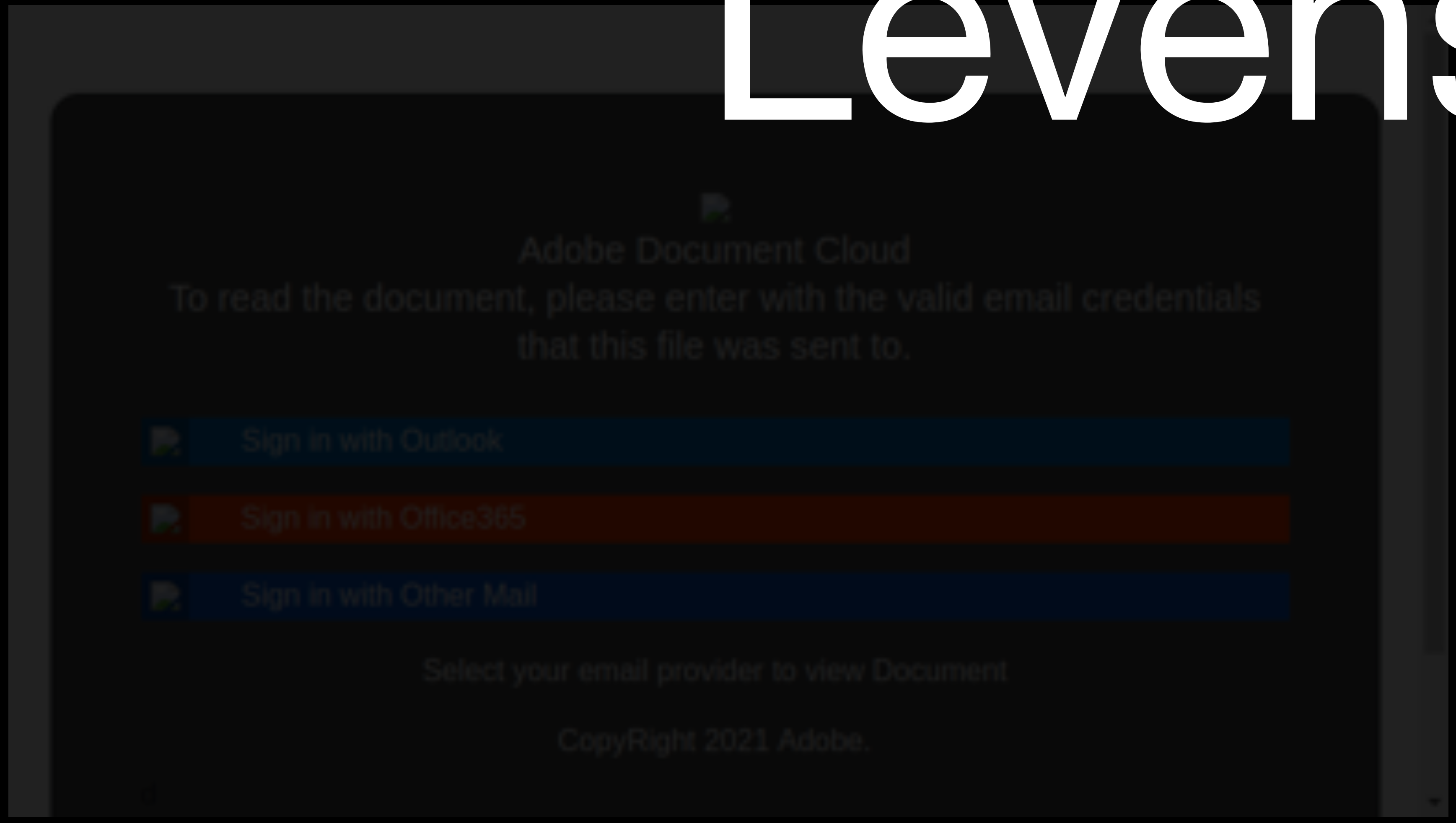
84926feadda7dfeb59777ab0e4b4cc60a977b6e1
3b4cd30a1f1d85b28601641afde04375169854f8



SHA256:
84926feadda7dfeb59777ab0e4b4cc60a977b6e1

Levenshtein Distance

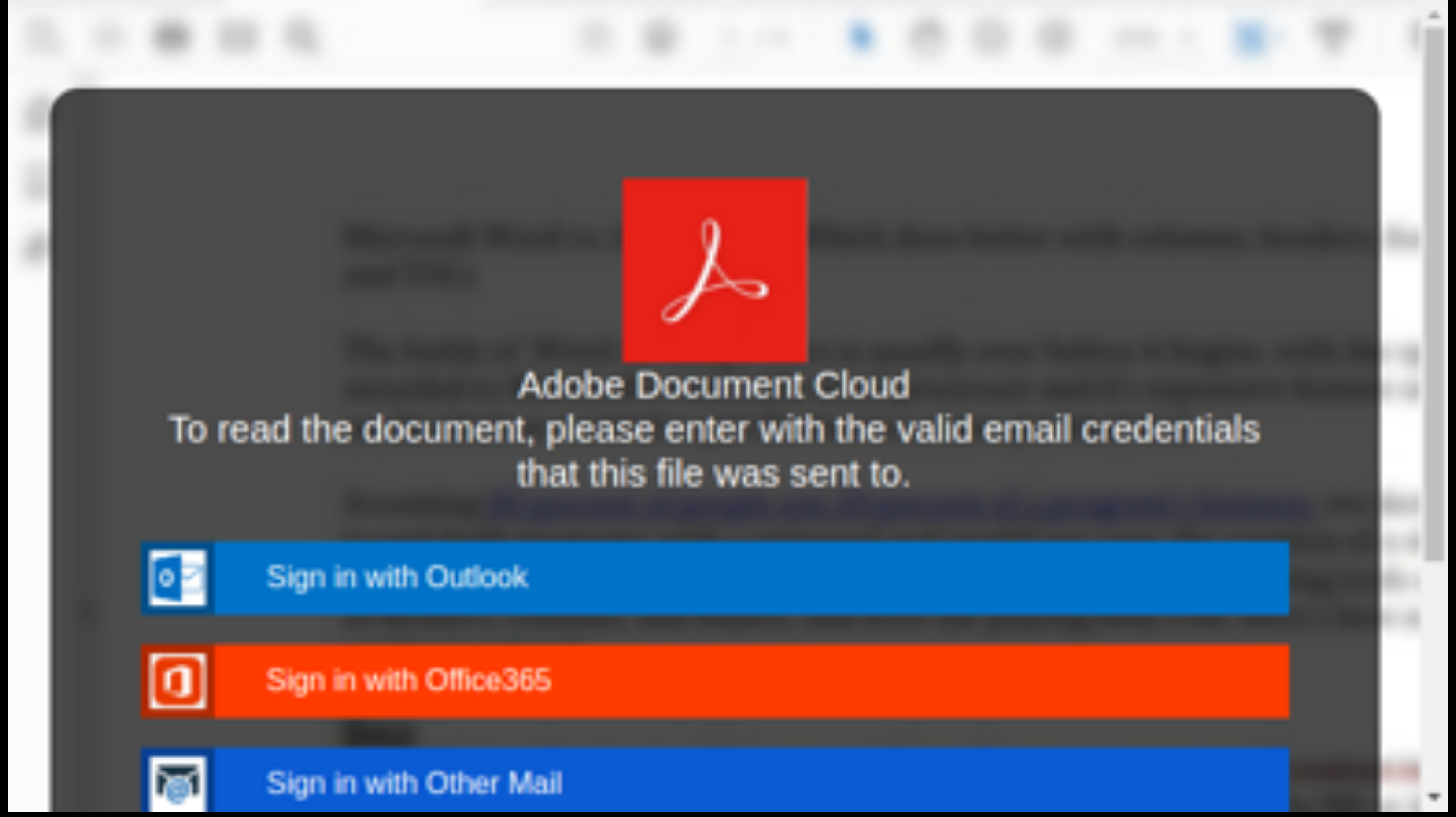
36


Adobe Document Cloud
To read the document, please enter with the valid email credentials that this file was sent to.

Select your email provider to view Document
CopyRight 2021 Adobe.

SHA256:
3b4cd30a1f1d85b28601641afde04375169854f8

PHASH is shorter, but a difference in the hashes don't suggest a drastic change in the images.



Adobe Document Cloud

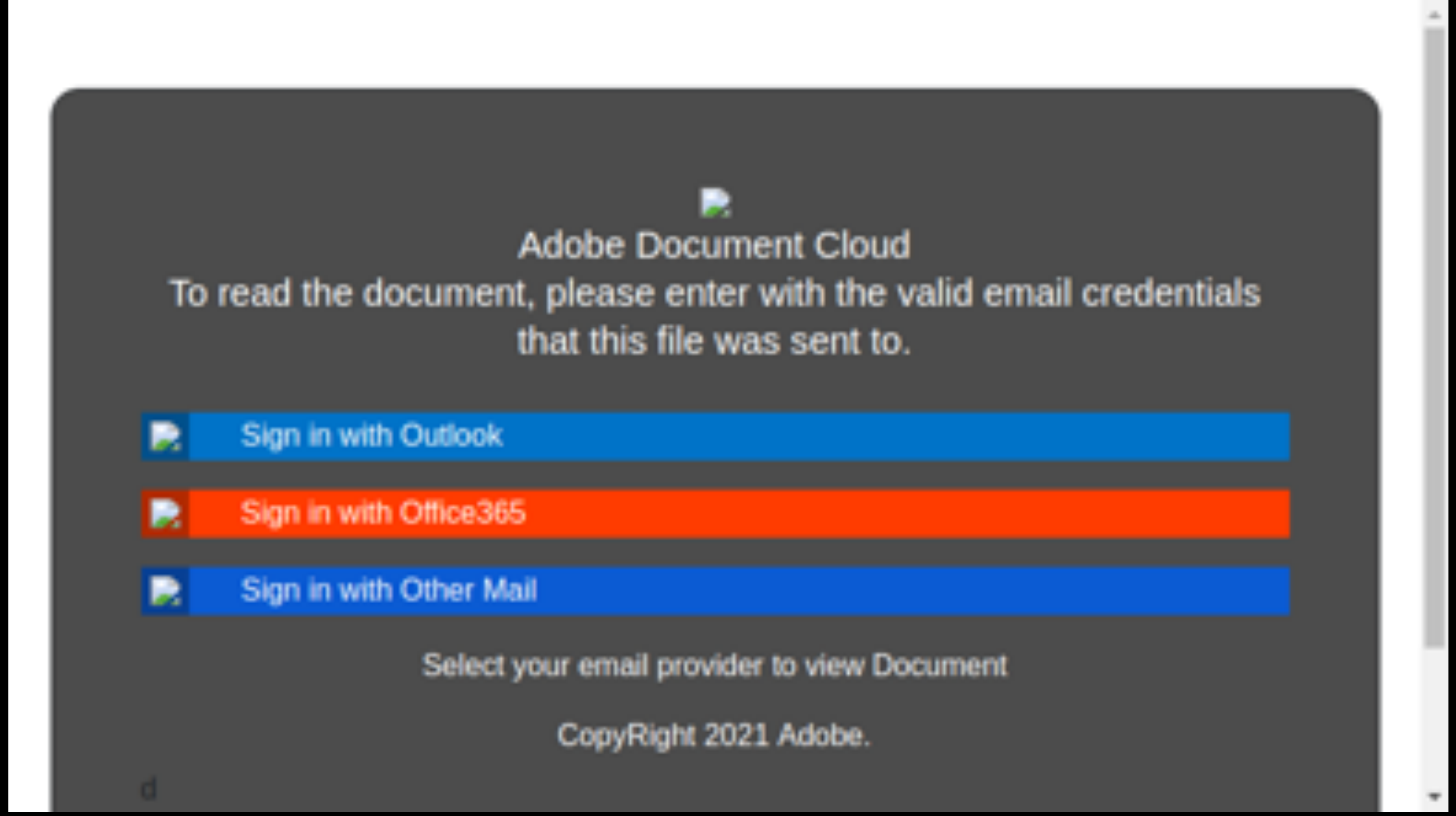
To read the document, please enter with the valid email credentials that this file was sent to.

Sign in with Outlook

Sign in with Office365

Sign in with Other Mail

PHASH:
aa95c1d5d595d485



Adobe Document Cloud

To read the document, please enter with the valid email credentials that this file was sent to.

Sign in with Outlook

Sign in with Office365

Sign in with Other Mail

Select your email provider to view Document

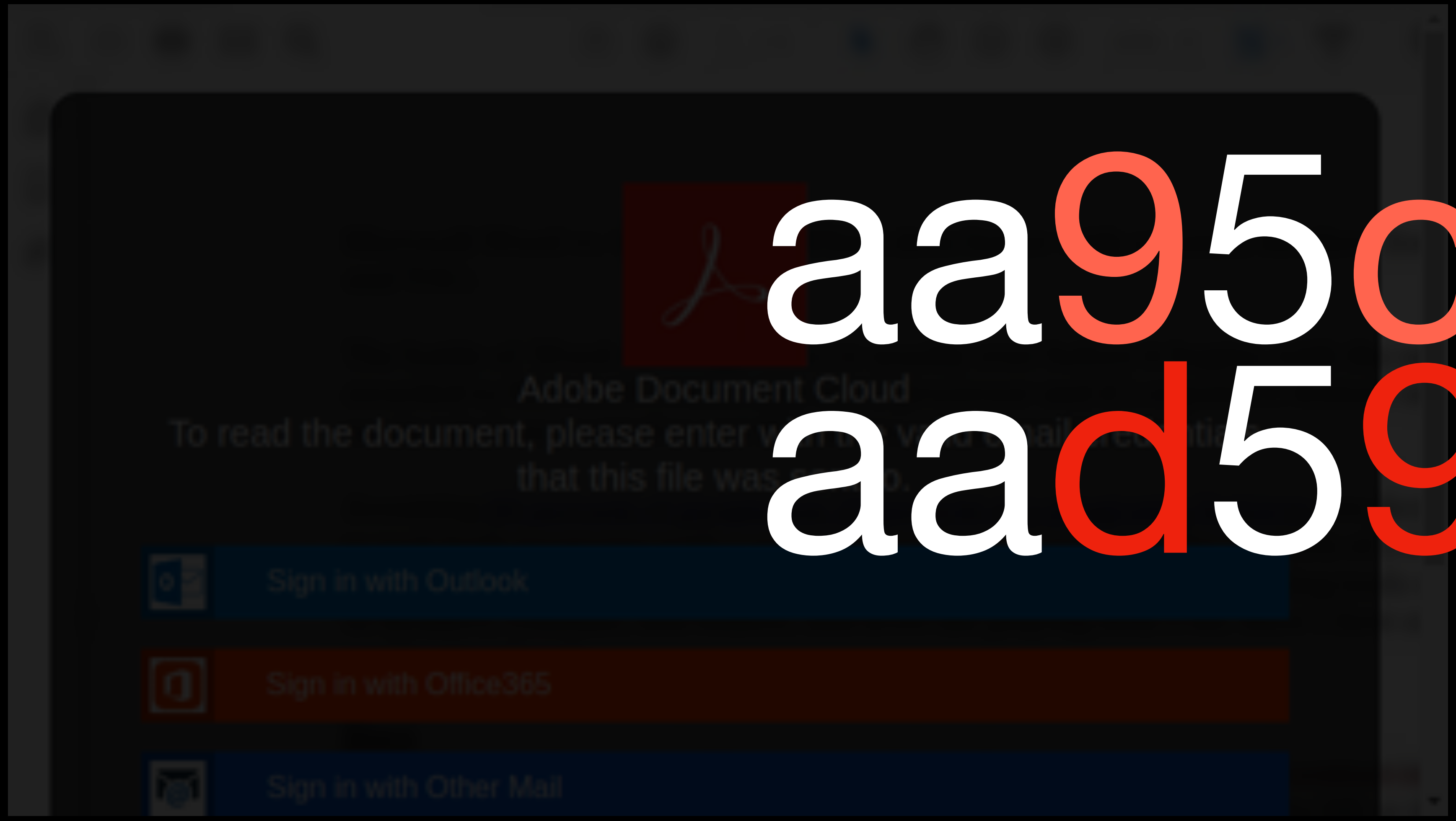
CopyRight 2021 Adobe.

PHASH:
aad595c1d595d590

PHASH is shorter, but a difference in the hashes don’t suggest a drastic change in the images. Levenshtein distance can be applied to these shorter hashes with promising results.

aa95c1d5d595d485

aad595c1d595d590



Adobe Document Cloud

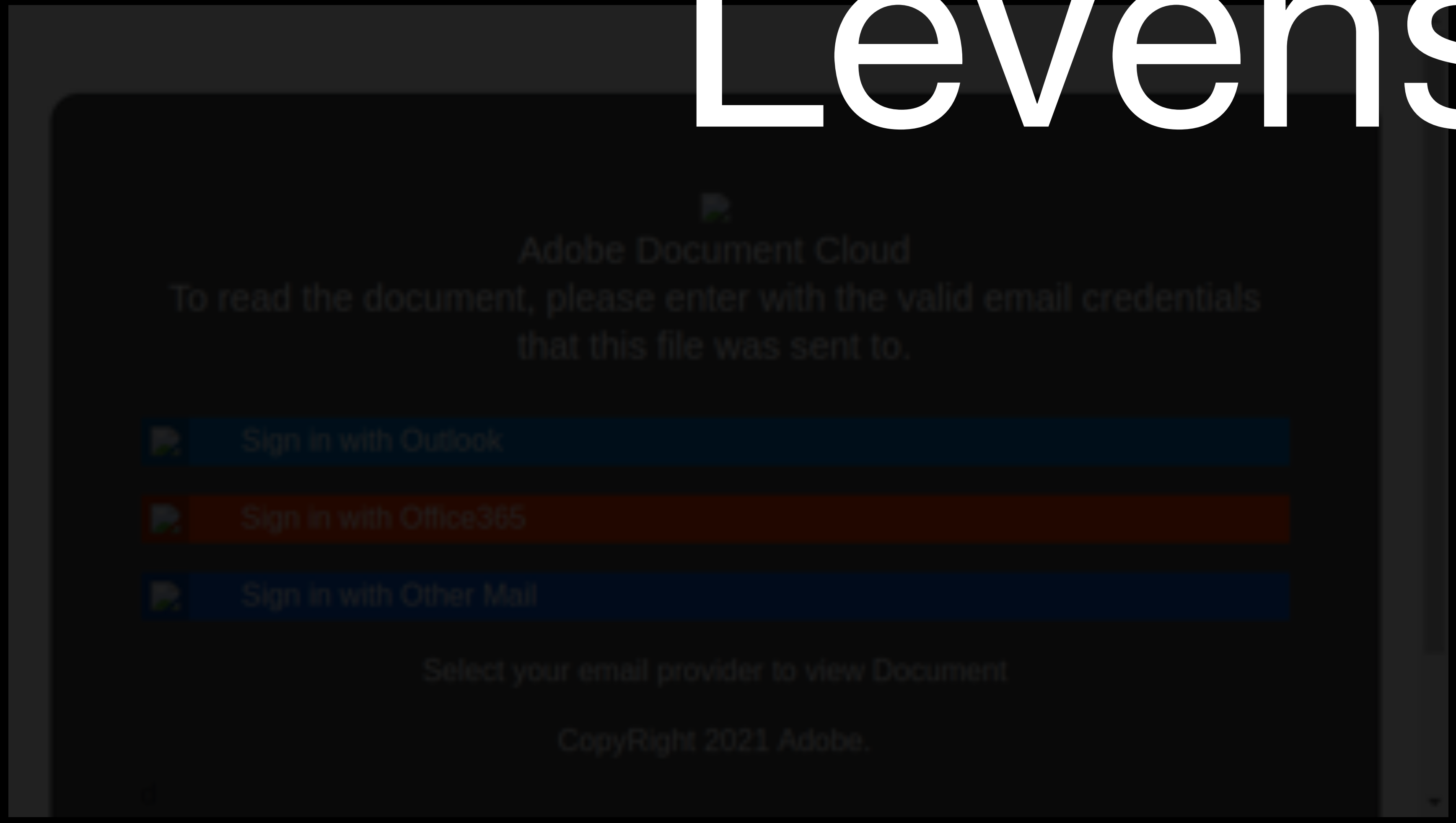
To read the document, please enter with the valid email credentials that this file was sent to.

- Sign in with Outlook
- Sign in with Office365
- Sign in with Other Mail

PHASH:
aa95c1d5d595d485

Levenshtein Distance

7



Adobe Document Cloud

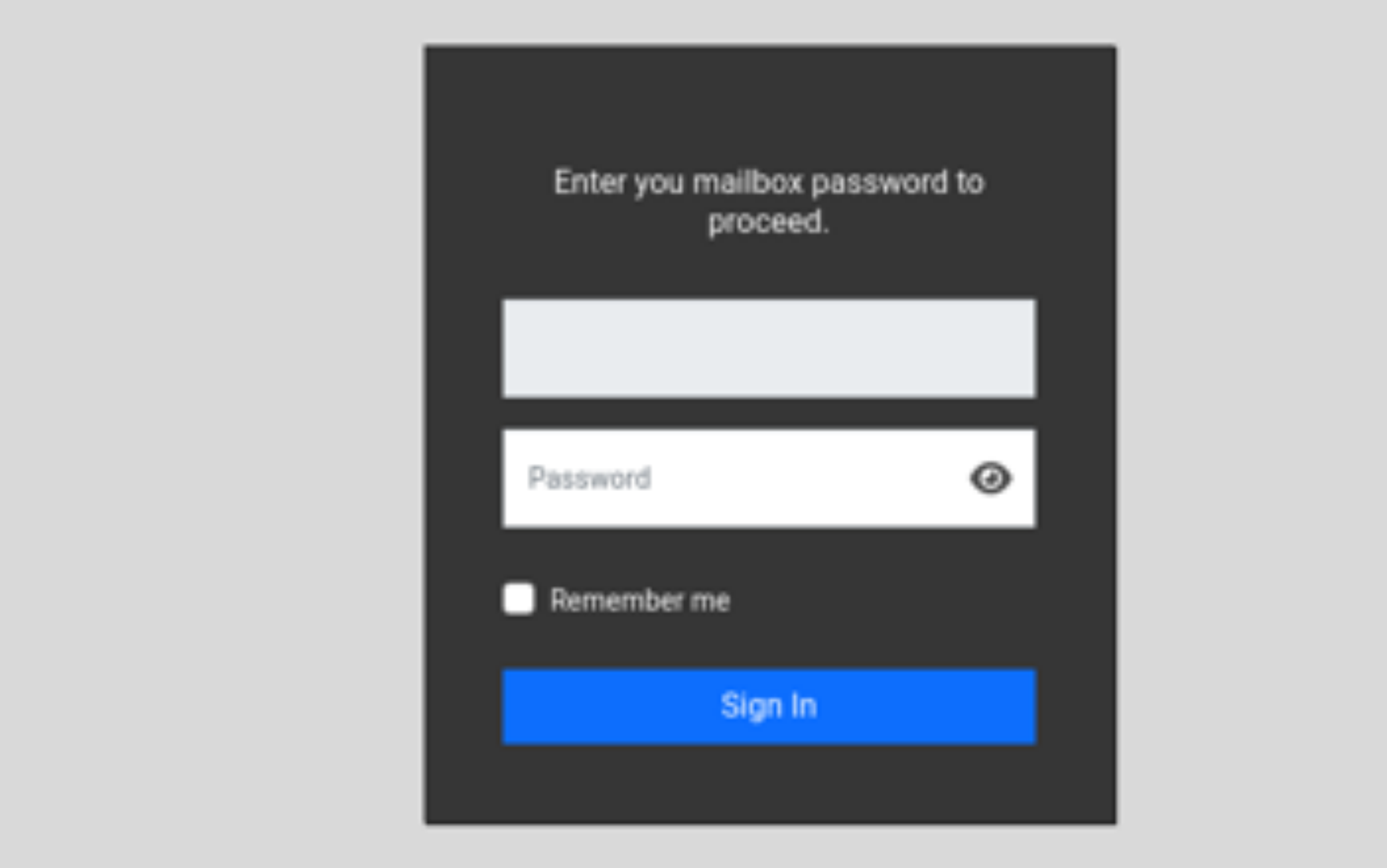
To read the document, please enter with the valid email credentials that this file was sent to.

- Sign in with Outlook
- Sign in with Office365
- Sign in with Other Mail

Select your email provider to view Document

CopyRight 2021 Adobe.

PHASH:
aad595c1d595d590


Enter you mailbox password to proceed.

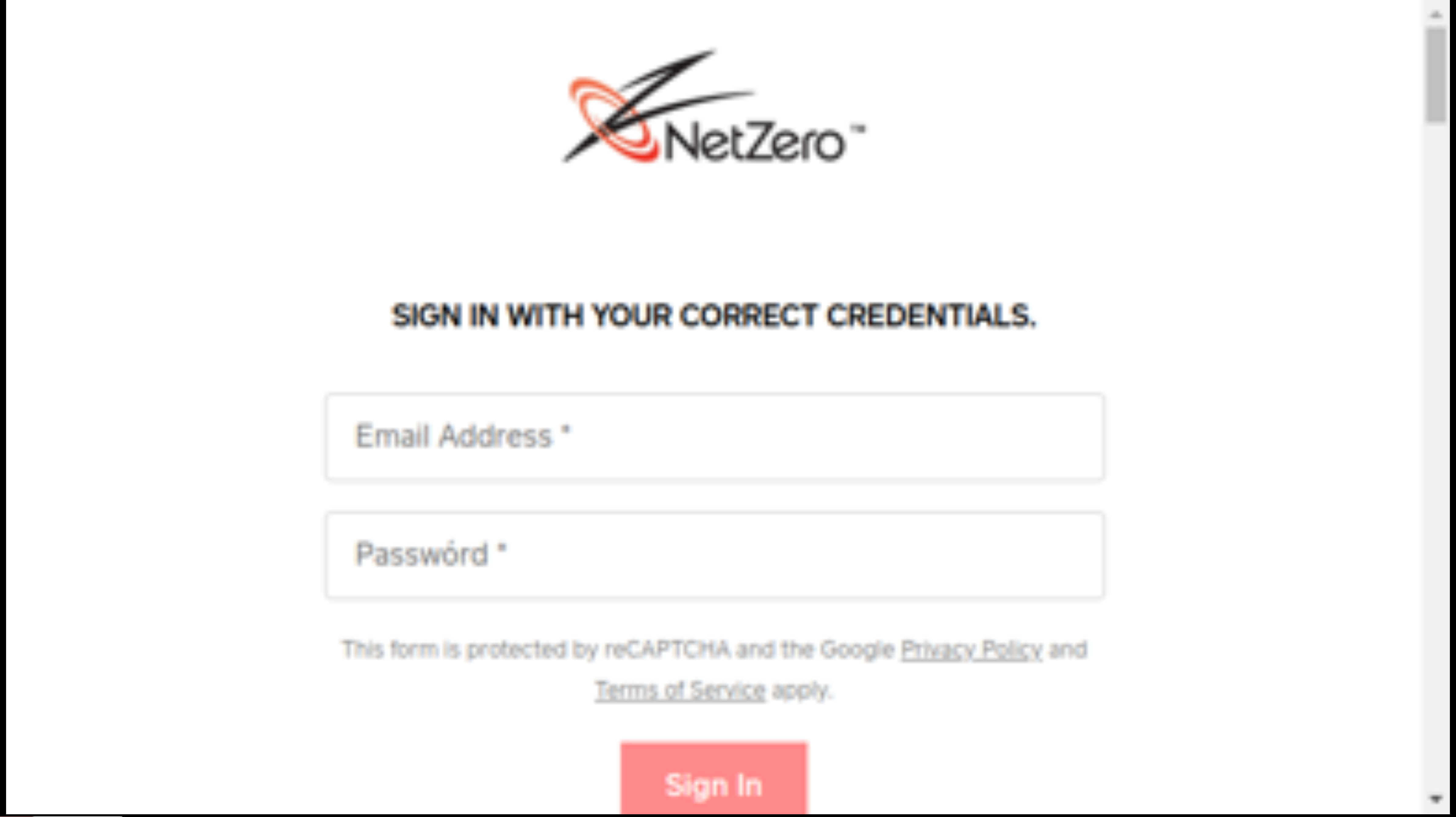
Password

☐ Remember me

These two images are obviously different, but they have some similarities. The PHASH's are different with a lev distance of 14

e59832669b9ac699

Levenshtein: 14



SIGN IN WITH YOUR CORRECT CREDENTIALS.

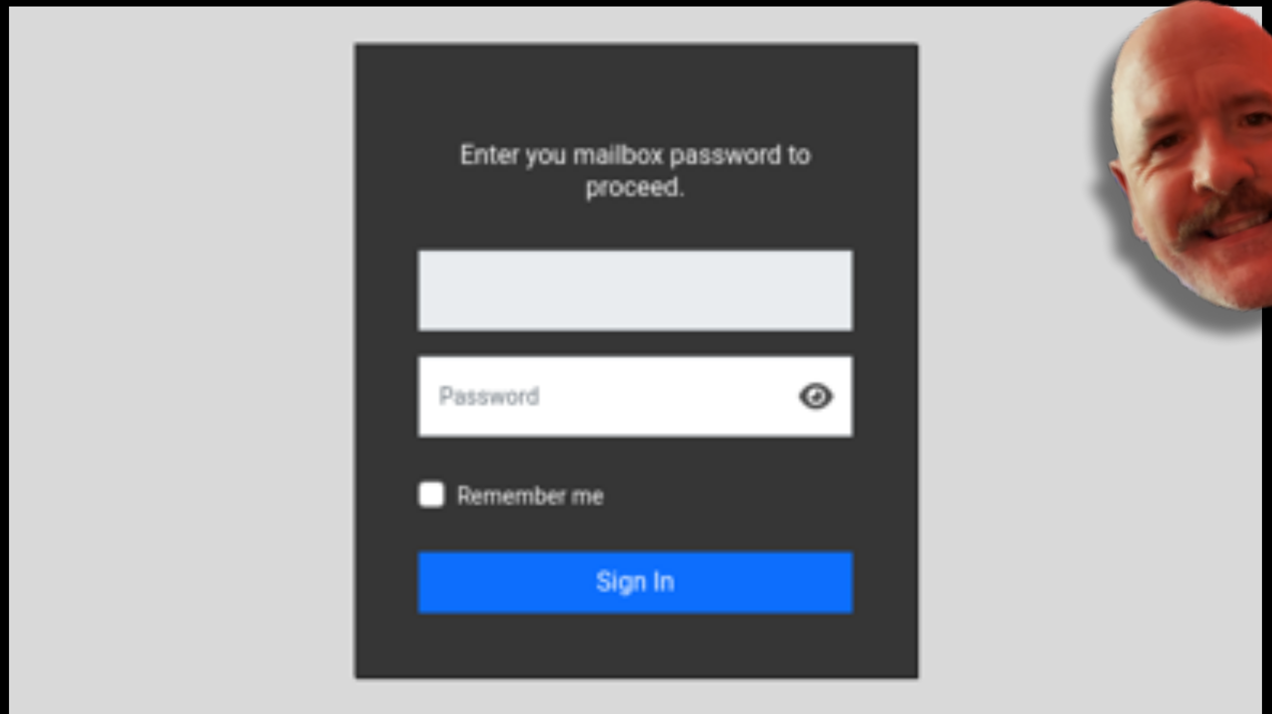
Email Address *

Password *

This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

b3517399334ccccc

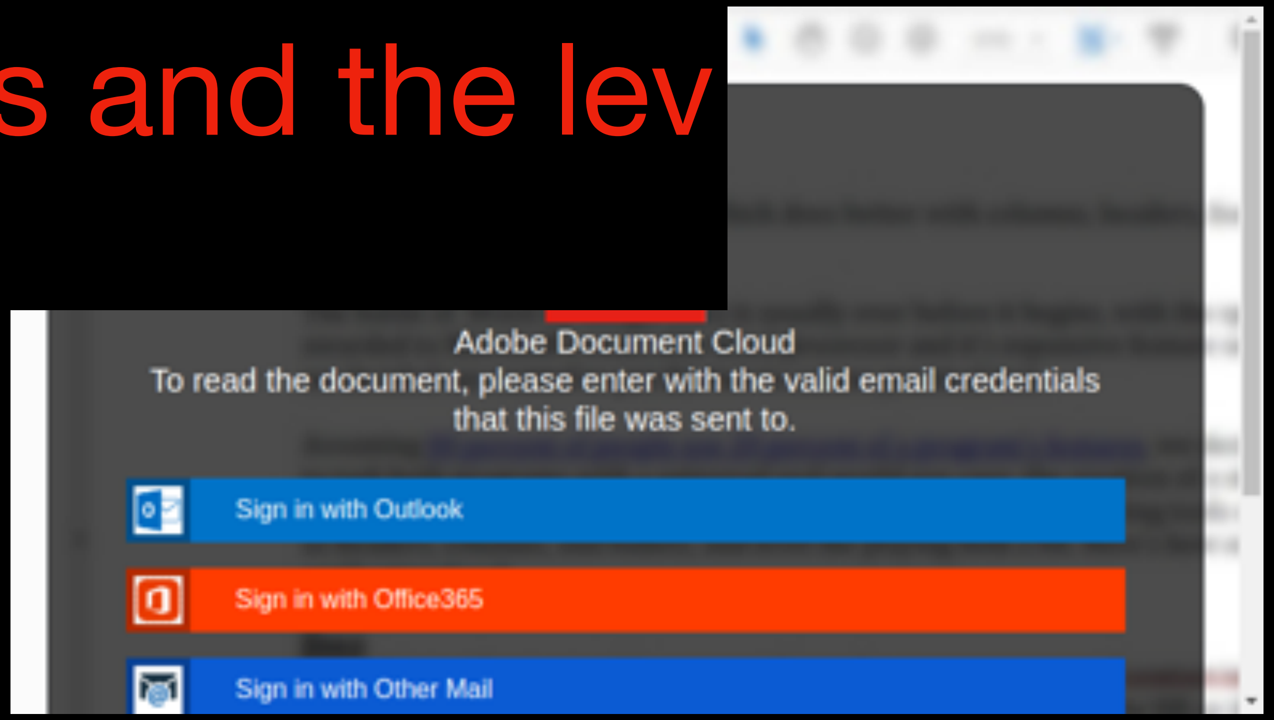
Showing multiple images and the lev
distances



Enter you mailbox password to proceed.

e59832669b9ac699

Levenshtein: 14



Adobe Document Cloud

To read the document, please enter with the valid email credentials that this file was sent to.

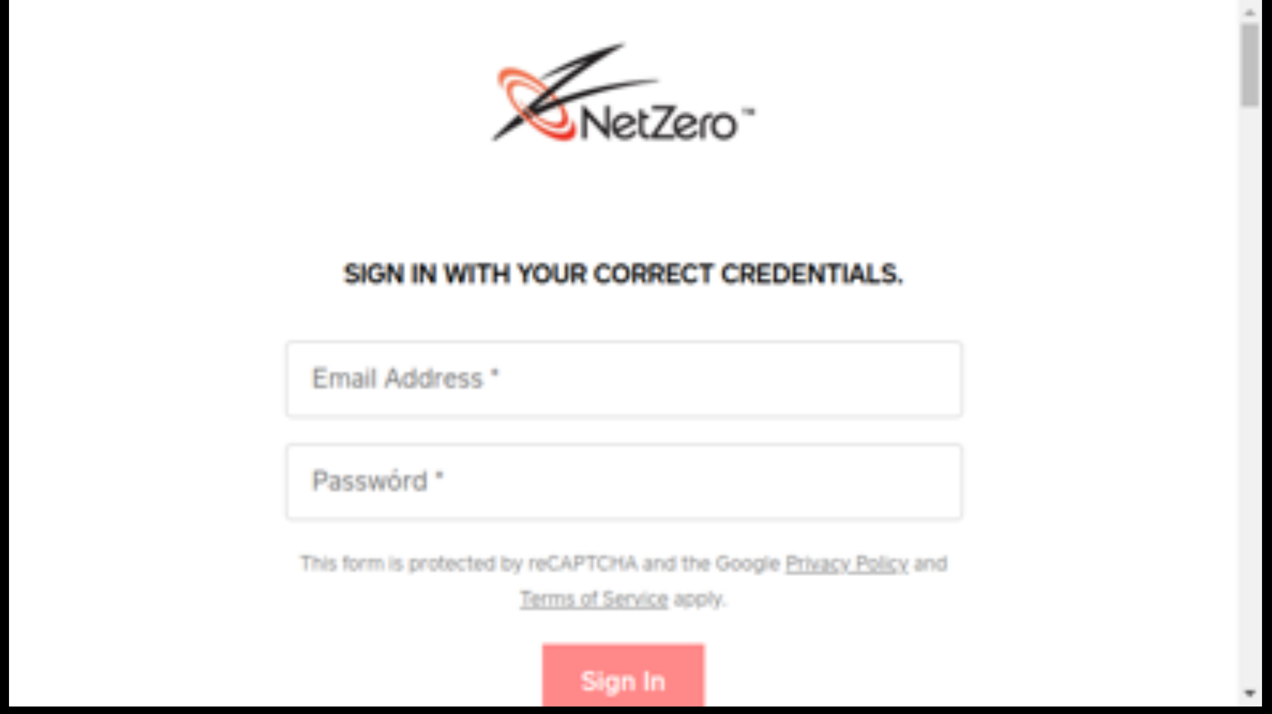
aa95c1d5d595d485

Levenshtein: 14

Levenshtein: 15

Levenshtein: 14

Levenshtein: 7

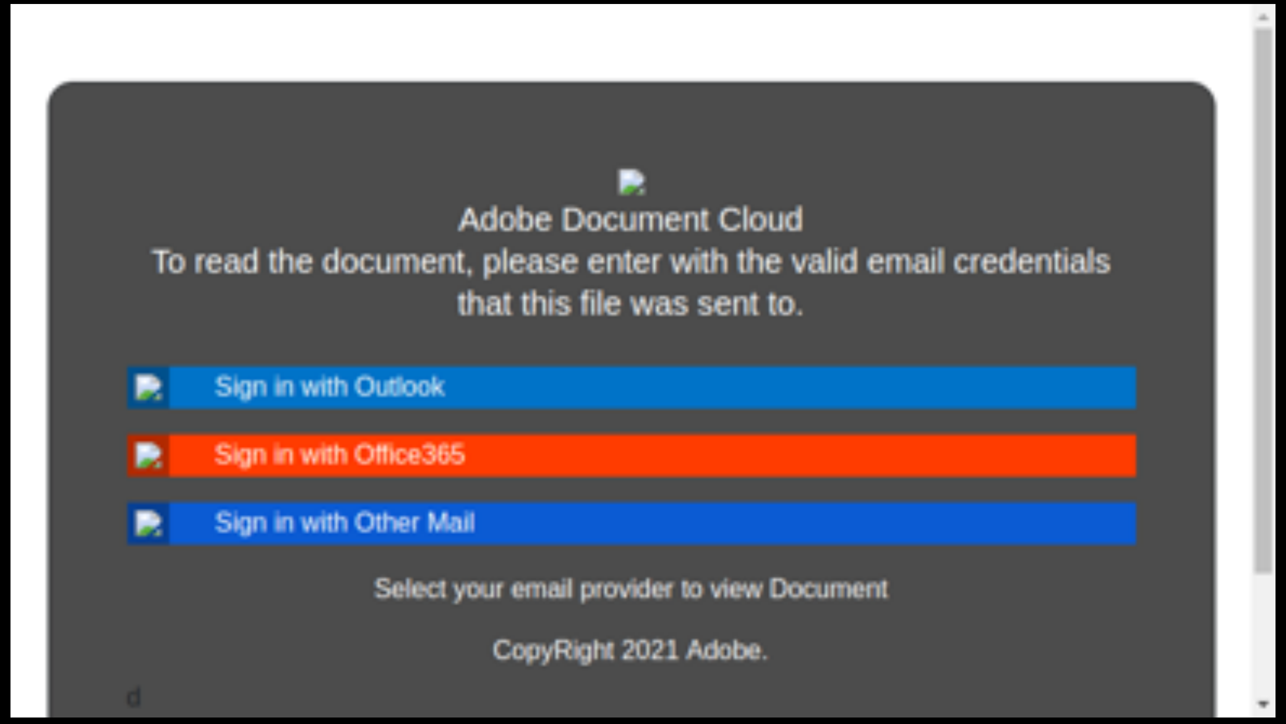


SIGN IN WITH YOUR CORRECT CREDENTIALS.

This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

b3517399334ccccc

Levenshtein: 16



Adobe Document Cloud

To read the document, please enter with the valid email credentials that this file was sent to.

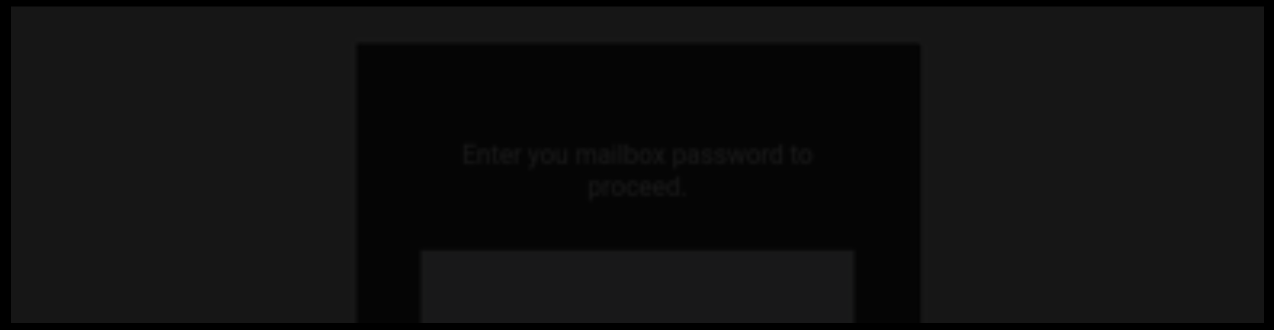
Select your email provider to view Document

CopyRight 2021 Adobe.

aad595c1d595d590

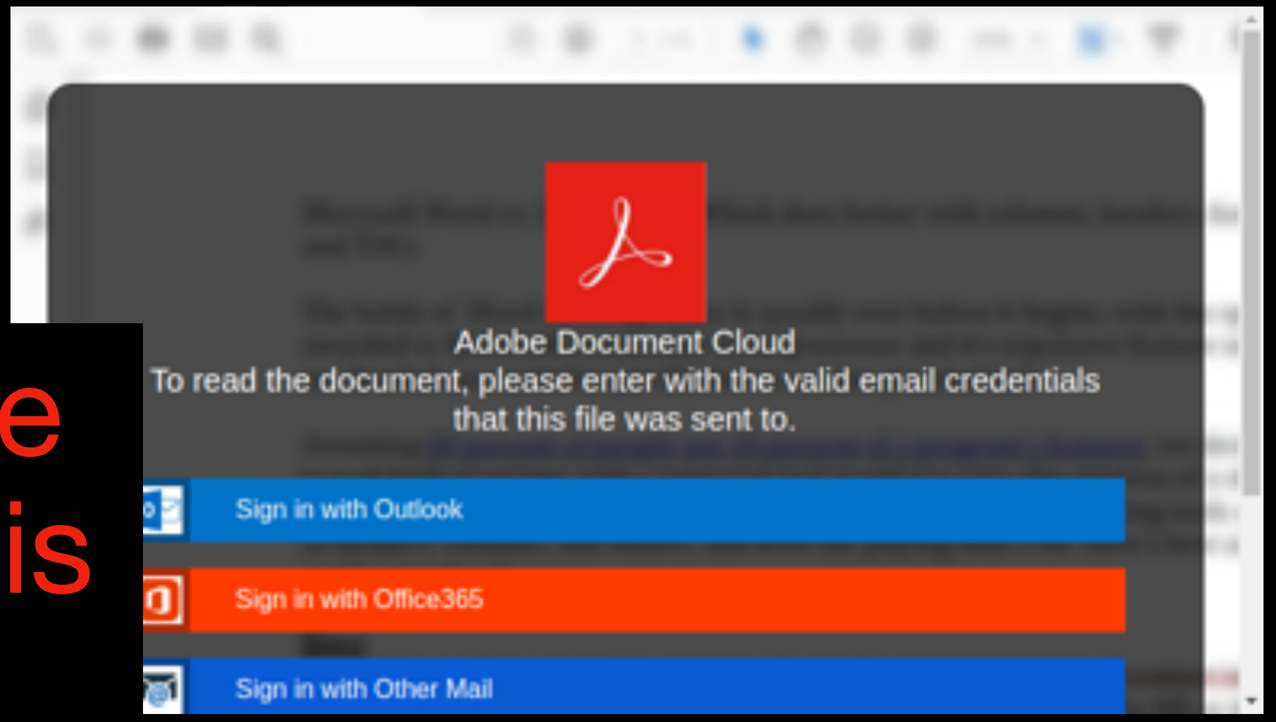
Setting a Threshold

The images that look the most alike, but which are still different have a lev distance of 7. I can use this analysis to start setting a threshold.



Enter you mailbox password to proceed.

e59832669b9ac699



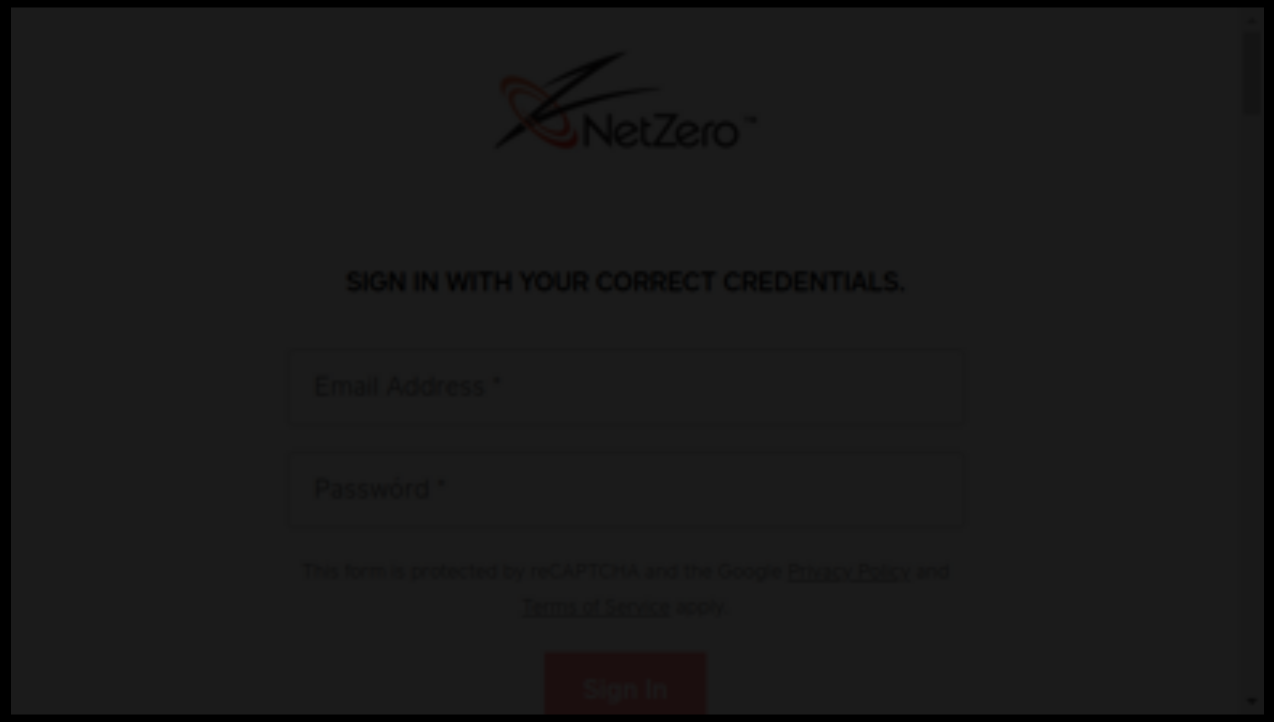
Adobe Document Cloud
To read the document, please enter with the valid email credentials that this file was sent to.

aa95c1d5d595d485

Levenshtein: 15

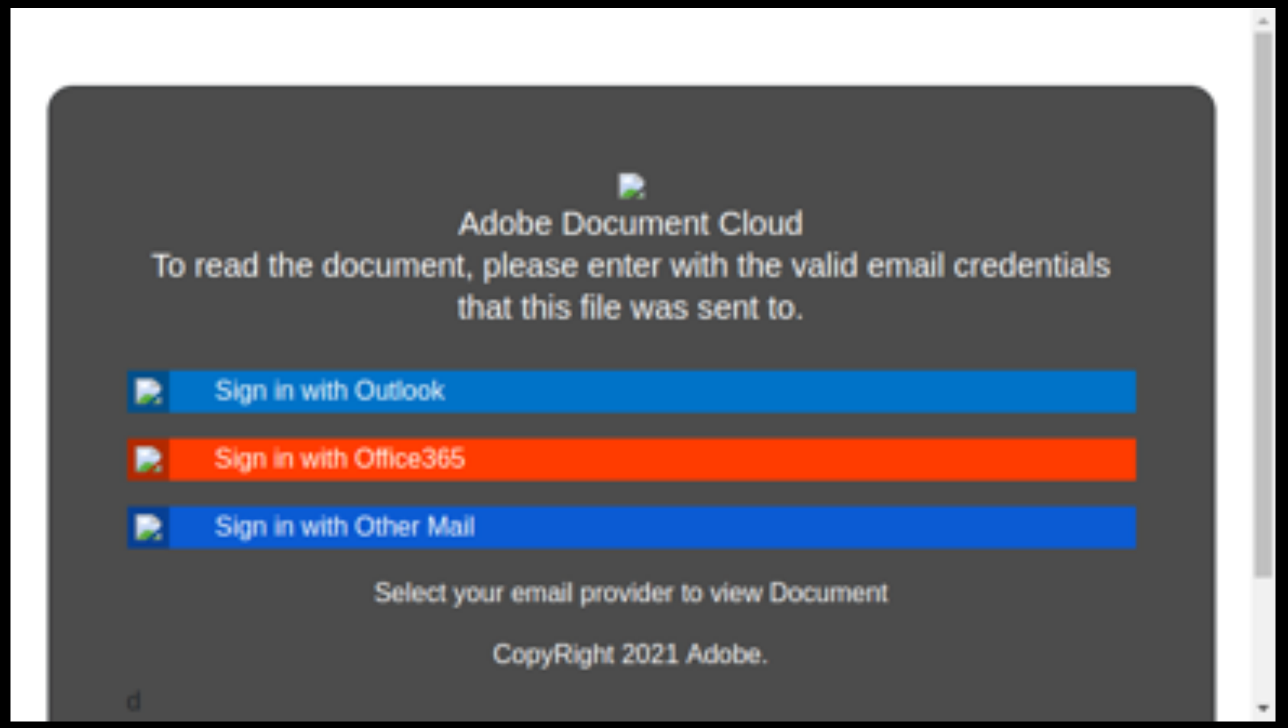
Levenshtein: 14

Levenshtein: 7



SIGN IN WITH YOUR CORRECT CREDENTIALS.

Levenshtein: 16



Adobe Document Cloud
To read the document, please enter with the valid email credentials that this file was sent to.

Select your email provider to view Document

CopyRight 2021 Adobe.

b3517399334ccccc

aad595c1d595d590

Now I can get back to removing non-relevant images in the malicious dataset. I’m going to group them using a threshold and delete non-relevant images as groups.

Grouping Images

Using the threshold

To demonstrate what grouping all the crawled screenshots looks like, I’ll be taking a random bunch of screenshots like this:

Webmail

USERNAME

PASSWORD

Enter email password

advertisedszp.shop needs to review the security of your connection before proceeding.

Error 1101

What can I do?

If you are the owner of this website:

404 Not Found

Sign In

to myAT&T

User ID

Password

Forgot user ID?

Forgot password?

Save user ID

Enter you mailbox password to proceed.

Password

Remember me

Sign In

amazon

Enter the characters you see below

Type the characters you see in this image:

PXMTBL

Try different image

Type characters

signup-swellnetwork.net

Verify you are human by completing the action below.

I am human

signup-swellnetwork.net needs to review the security of your

Webmail

USERNAME

PASSWORD

Enter email password

quialitsuzoxm.shop

Verify you are human by completing the action below.

Verify you are human

quialitsuzoxm.shop needs to review the security of your connection before proceeding.

Enter you mailbox password to proceed.

Password

Remember me

Sign In

Microsoft

Sign in

Email, phone, or Skype

Next

DocuSign

Signature Required

View Document With Your Work Email

Outlook

Microsoft

Sign in

Email, phone, or Skype

Next

amazon

Enter the characters you see below

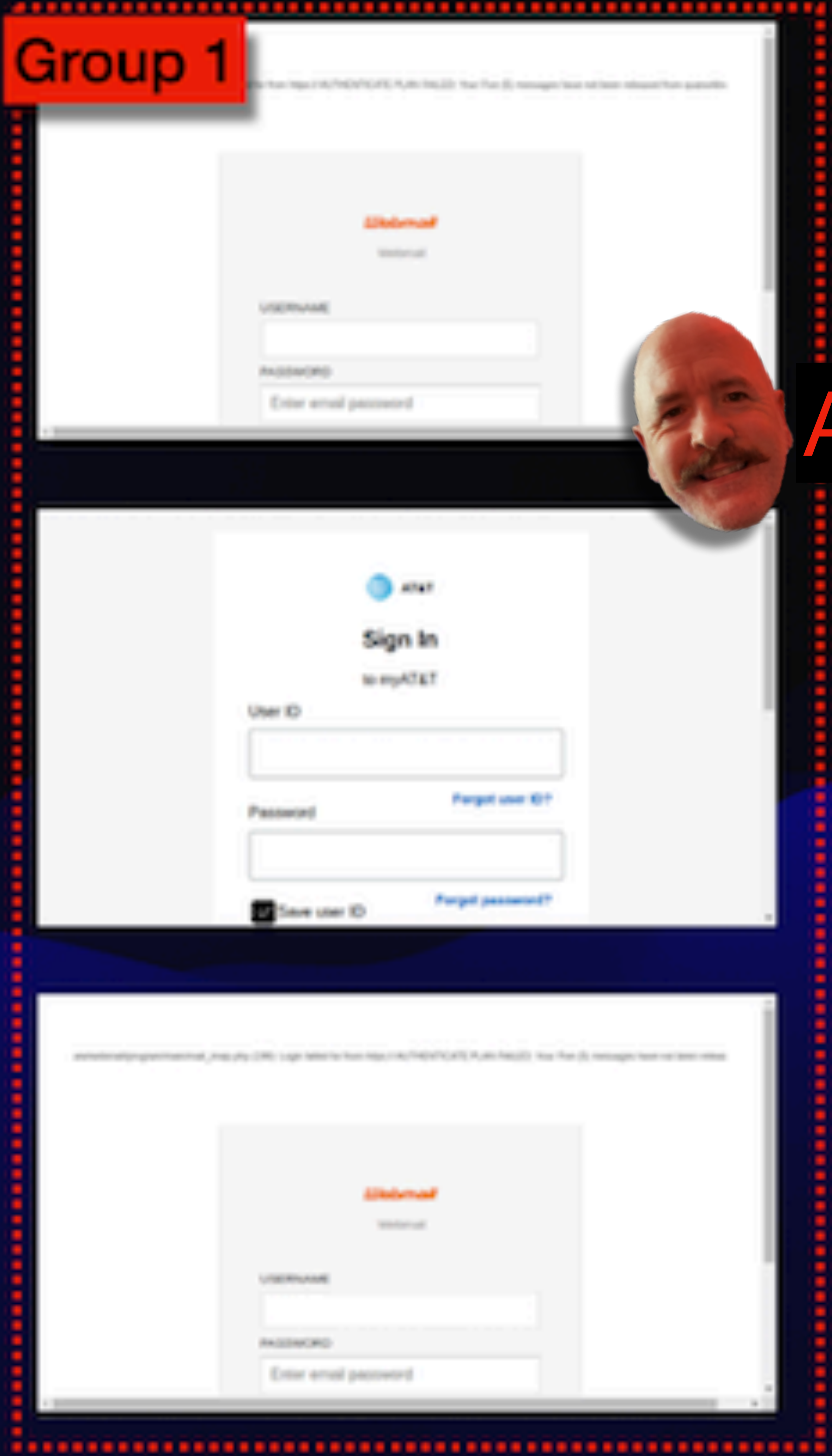
Type the characters you see in this image:

EHBUFN

Try different image

Type characters

And group similar together.


Group 1

Webmail

USERNAME

PASSWORD

Enter email password

Sign In

to myAT&T

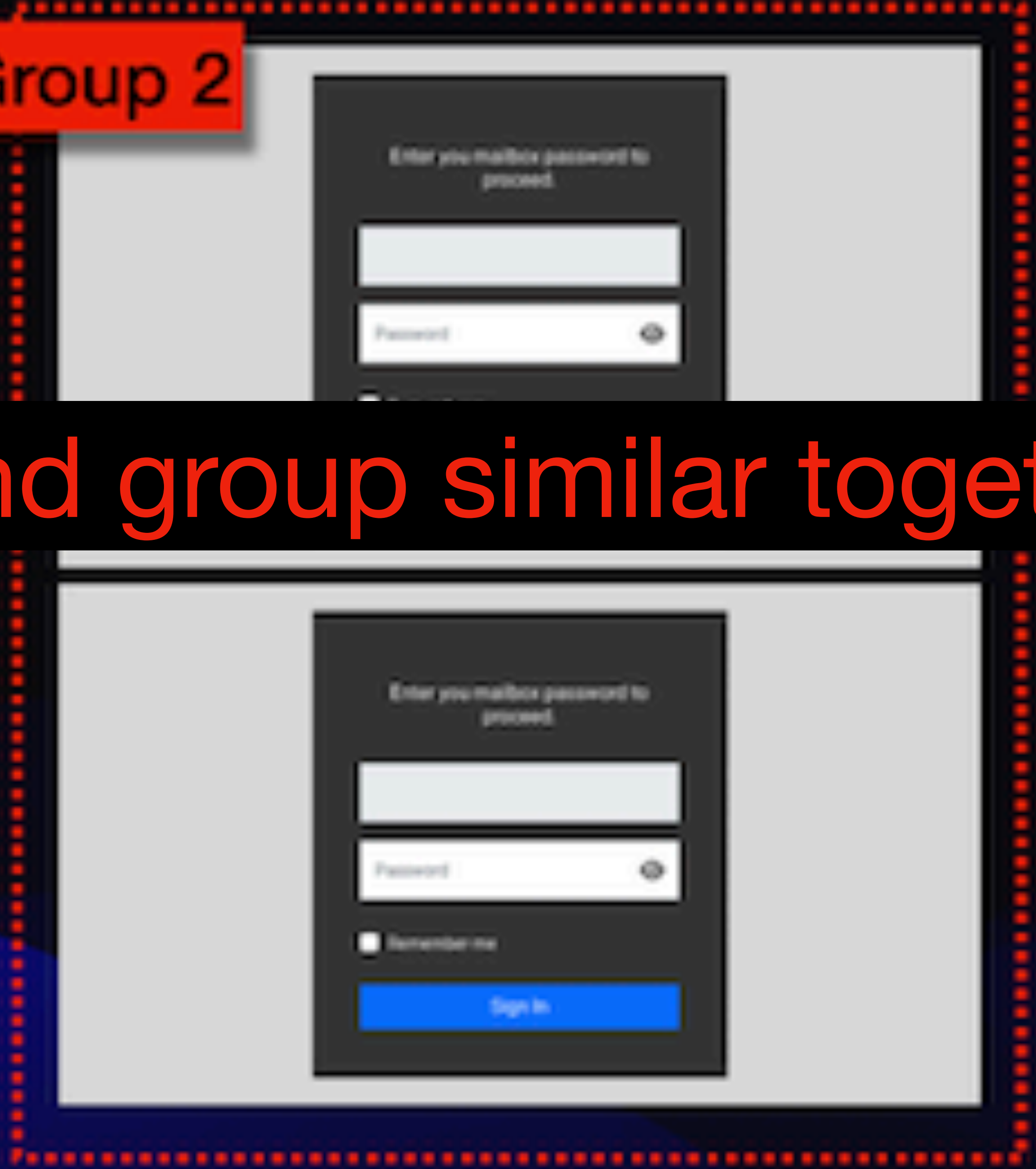
User ID

Password

Forgot user ID?

Forgot password?

Save user ID

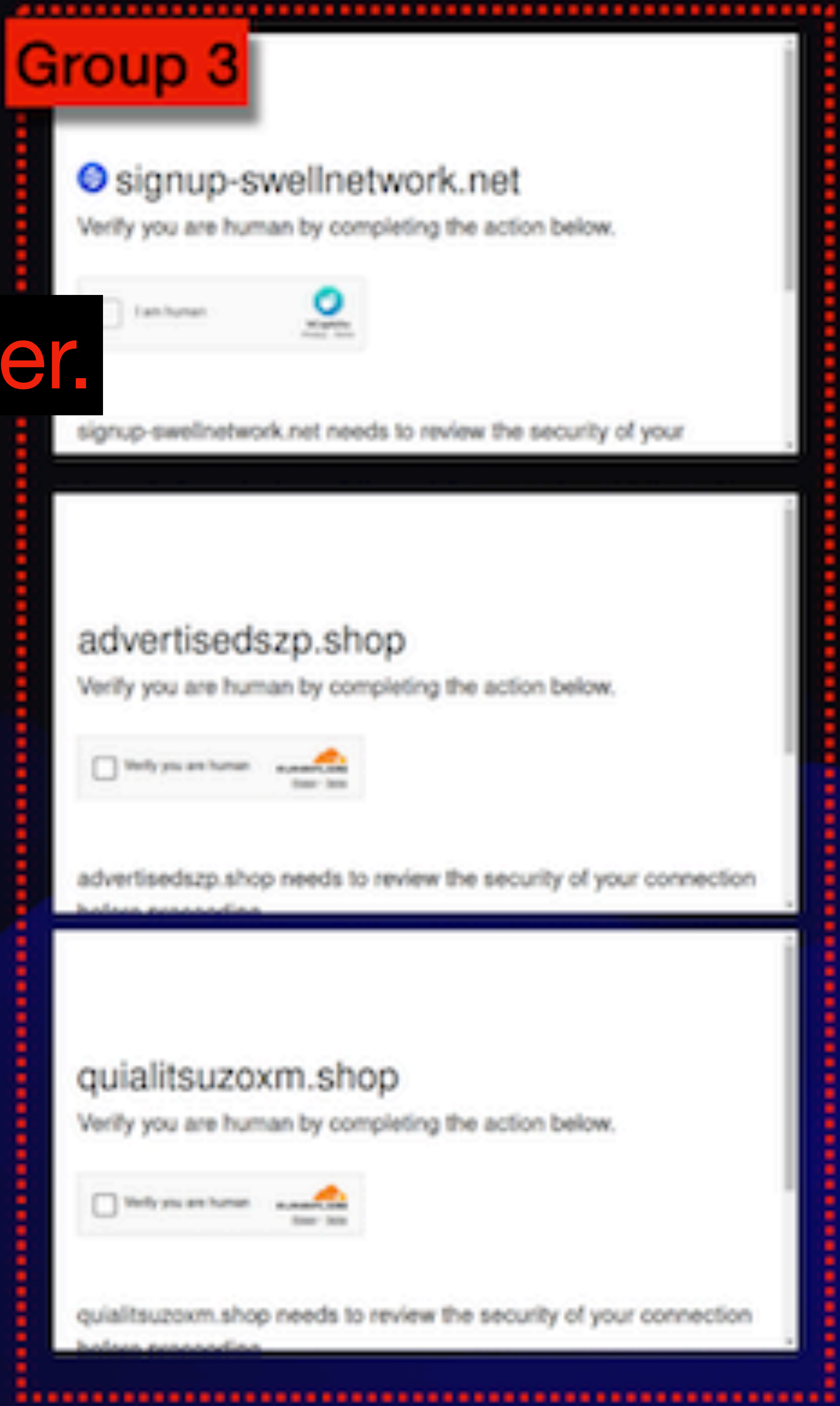

Group 2

Enter you mailbox password to proceed.

Password

Remember me

Sign In


Group 3

signup-swellnetwork.net

Verify you are human by completing the action below.

☐ I am human

signup-swellnetwork.net needs to review the security of your

advertisedszp.shop

Verify you are human by completing the action below.

☐ Verify you are human

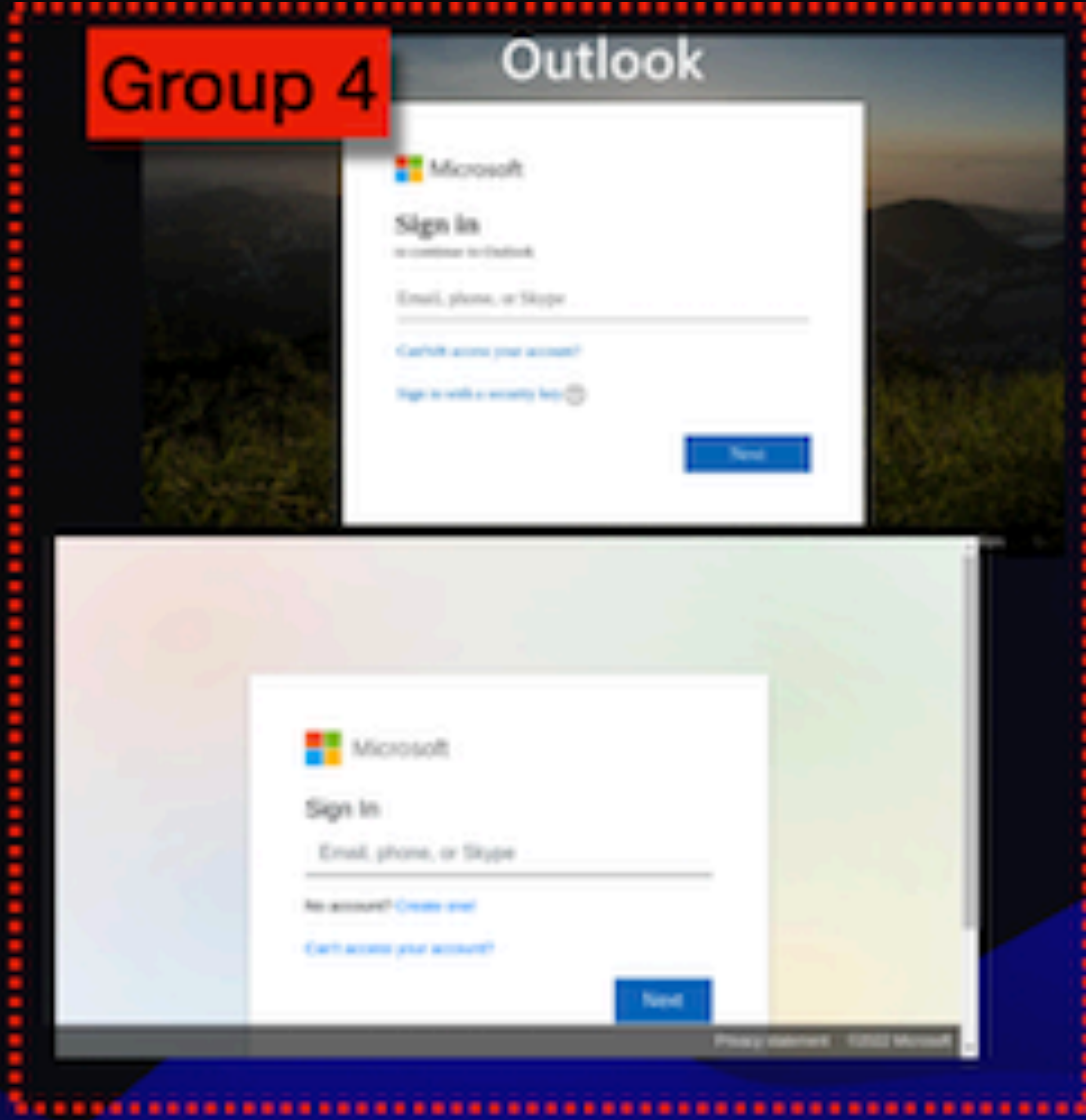
advertisedszp.shop needs to review the security of your connection

qualitsuzoxm.shop

Verify you are human by completing the action below.

☐ Verify you are human

qualitsuzoxm.shop needs to review the security of your connection


Group 4

Outlook

Microsoft

Sign in

to continue to Outlook

Email, phone, or Skype

Can't access your account?

Next

Microsoft

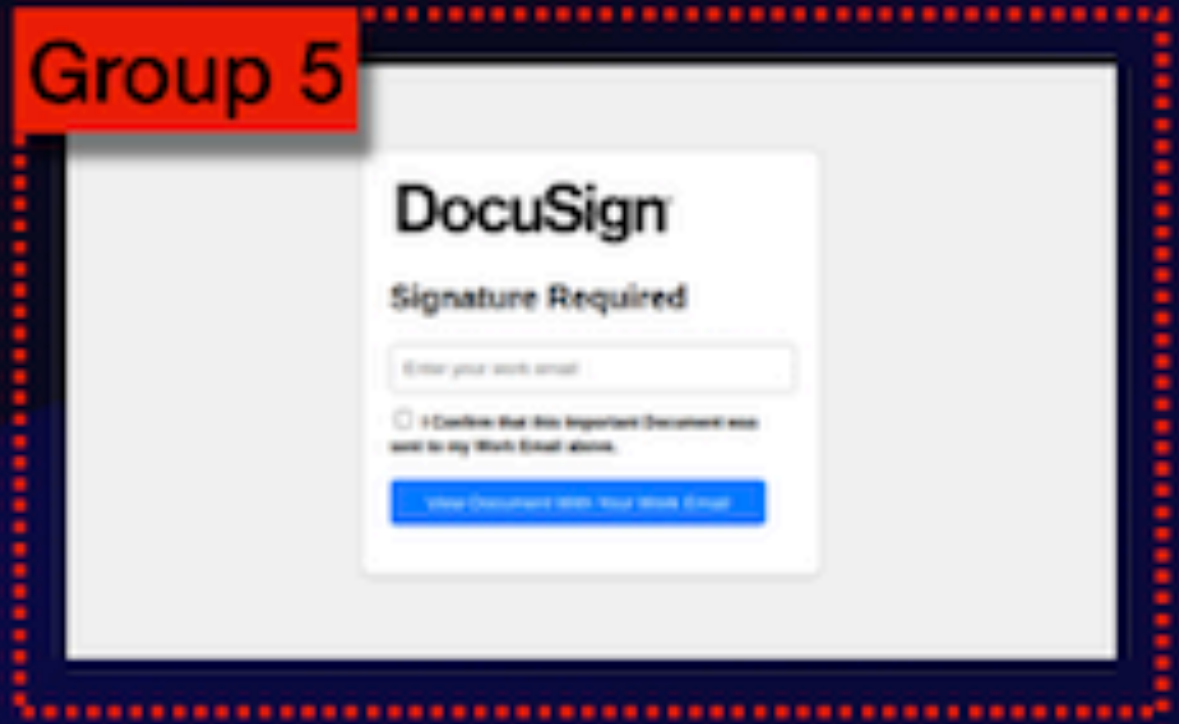
Sign in

Email, phone, or Skype

No account? Create one!

Can't access your account?

Next


Group 5

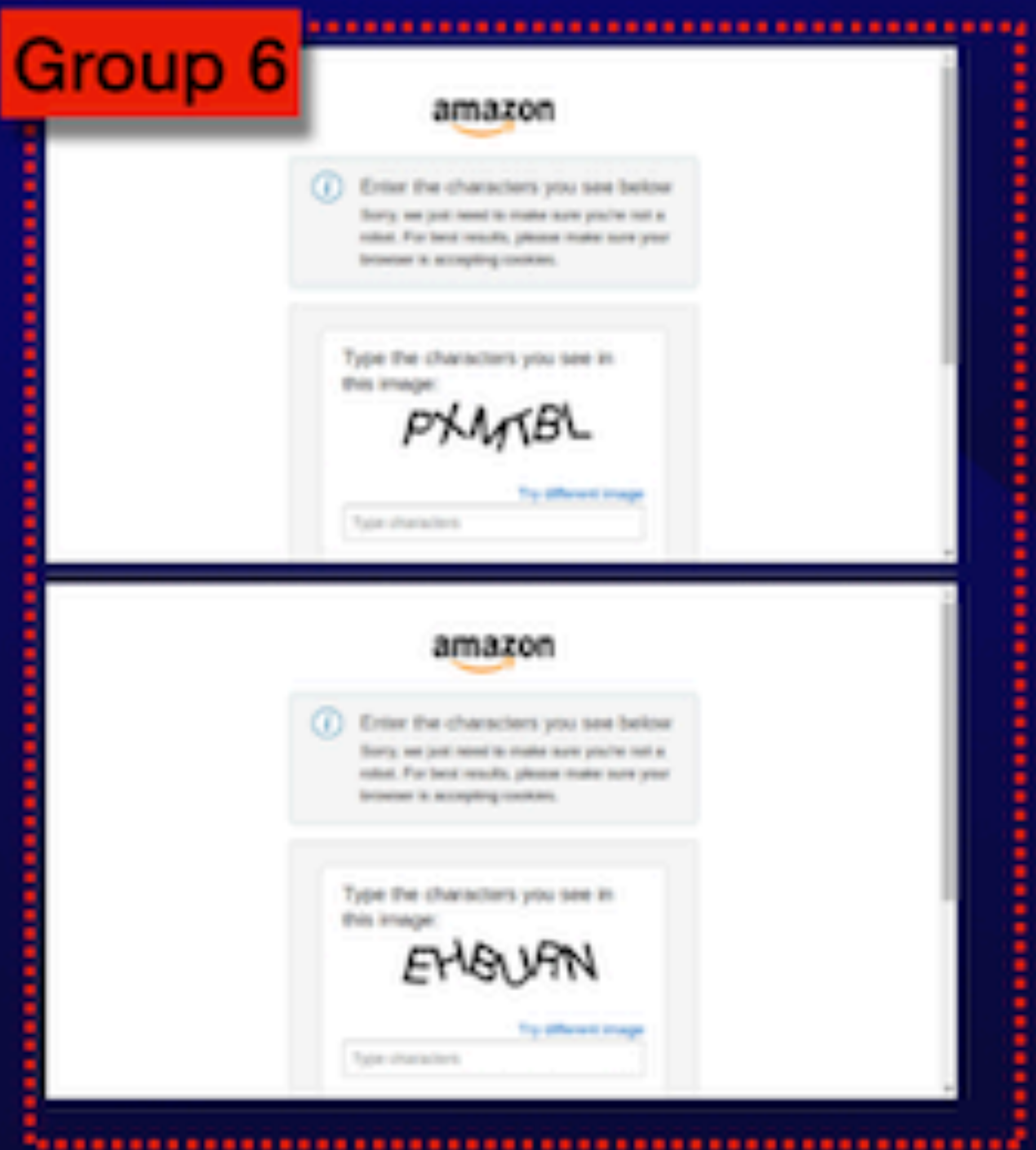
DocuSign

Signature Required

Enter your work email

☐ I Confirm that this important Document was sent to my Work Email above.

View Document With Your Work Email


Group 6

amazon

Enter the characters you see below

Type the characters you see in this image:

PXMTBL

Try different image

Type characters

amazon

Enter the characters you see below

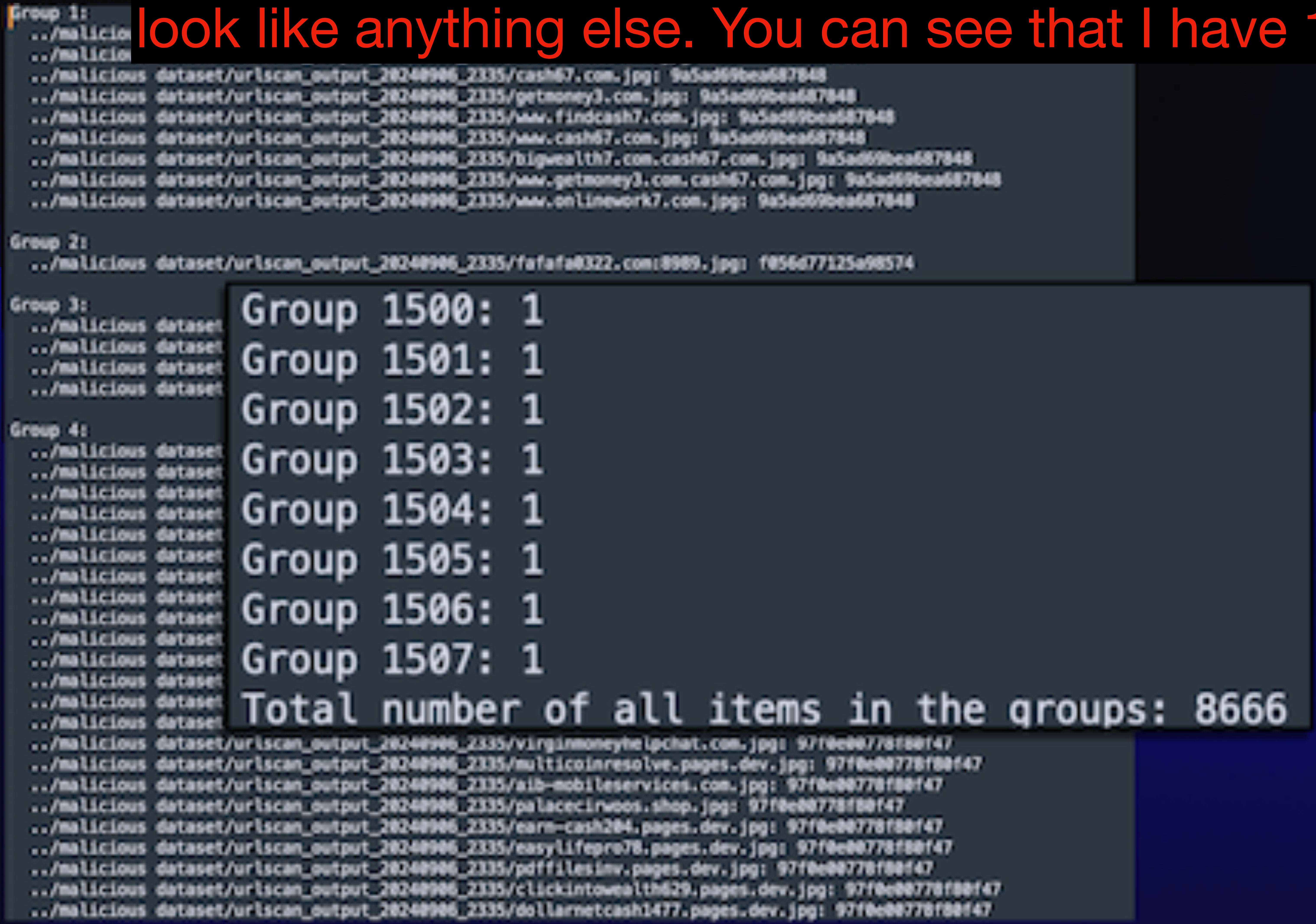
Type the characters you see in this image:

EHBUFN

Try different image

Type characters

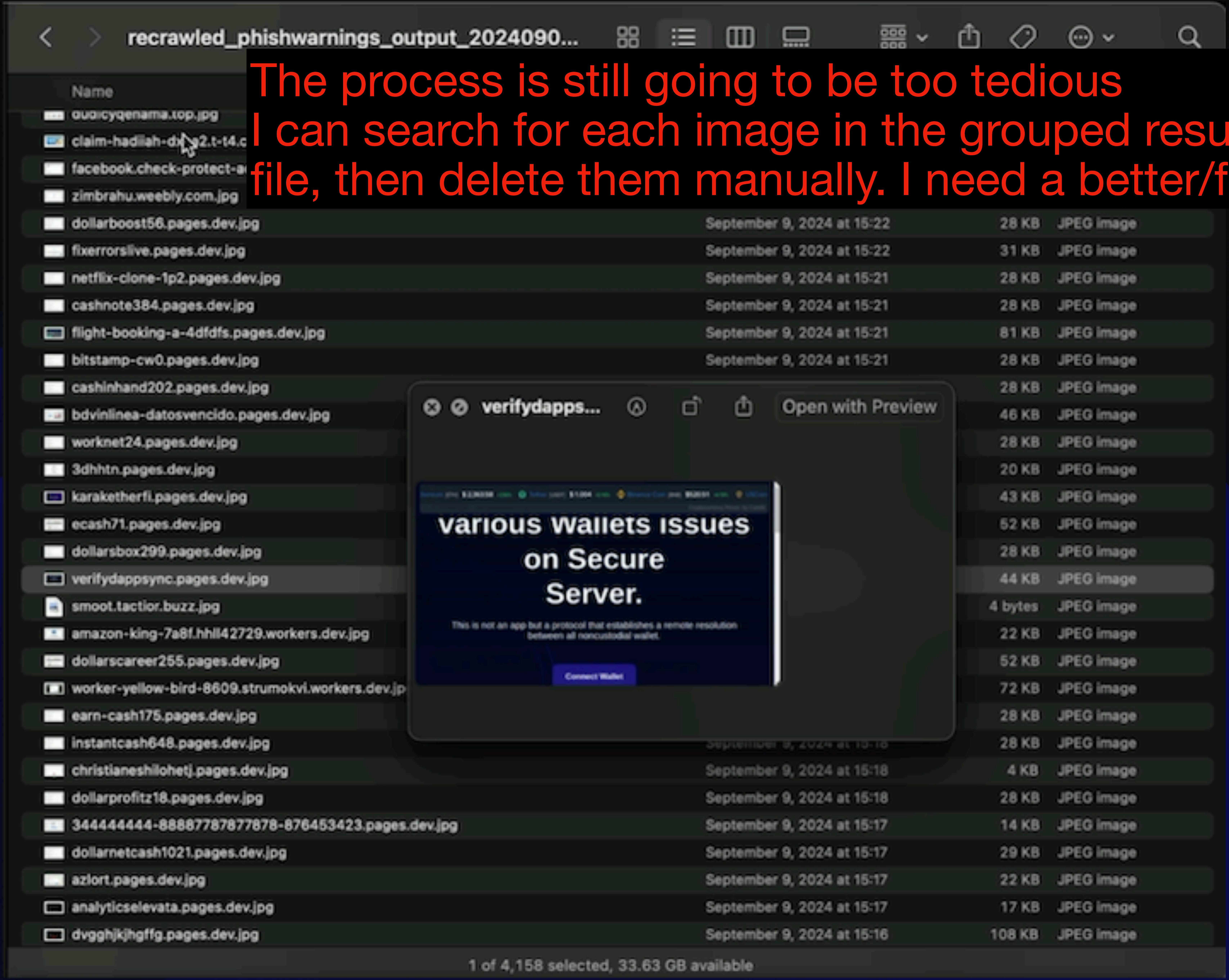
Here's the initial process of grouping that I did using Python.
It just output a text file with the file locations that matched a group all together.
However, all images are in groups, including images that don't look like anything else. You can see that I have 1,507 groups.



```
Group 1:
  ../malicio
  ../malicio
  ../malicious dataset/urlscan_output_20240906_2335/cash67.com.jpg: 9a5ad69bea687848
  ../malicious dataset/urlscan_output_20240906_2335/getmoney3.com.jpg: 9a5ad69bea687848
  ../malicious dataset/urlscan_output_20240906_2335/www.findcash7.com.jpg: 9a5ad69bea687848
  ../malicious dataset/urlscan_output_20240906_2335/www.cash67.com.jpg: 9a5ad69bea687848
  ../malicious dataset/urlscan_output_20240906_2335/bigwealth7.com.cash67.com.jpg: 9a5ad69bea687848
  ../malicious dataset/urlscan_output_20240906_2335/www.getmoney3.com.cash67.com.jpg: 9a5ad69bea687848
  ../malicious dataset/urlscan_output_20240906_2335/www.onlinework7.com.jpg: 9a5ad69bea687848

Group 2:
  ../malicious dataset/urlscan_output_20240906_2335/fafafa0322.com:8989.jpg: f056d77125a98574

Group 3:
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset

Group 4:
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset
  ../malicious dataset/urlscan_output_20240906_2335/virginmoneyhelpchat.com.jpg: 97f0e00778f80f47
  ../malicious dataset/urlscan_output_20240906_2335/multicoinresolve.pages.dev.jpg: 97f0e00778f80f47
  ../malicious dataset/urlscan_output_20240906_2335/aib-mobileservices.com.jpg: 97f0e00778f80f47
  ../malicious dataset/urlscan_output_20240906_2335/palacecirwoos.shop.jpg: 97f0e00778f80f47
  ../malicious dataset/urlscan_output_20240906_2335/earn-cash204.pages.dev.jpg: 97f0e00778f80f47
  ../malicious dataset/urlscan_output_20240906_2335/easylifepro78.pages.dev.jpg: 97f0e00778f80f47
  ../malicious dataset/urlscan_output_20240906_2335/pdffilesinv.pages.dev.jpg: 97f0e00778f80f47
  ../malicious dataset/urlscan_output_20240906_2335/clickintowealth629.pages.dev.jpg: 97f0e00778f80f47
  ../malicious dataset/urlscan_output_20240906_2335/dollarnetcash1477.pages.dev.jpg: 97f0e00778f80f47
```

```
Group 1500: 1
Group 1501: 1
Group 1502: 1
Group 1503: 1
Group 1504: 1
Group 1505: 1
Group 1506: 1
Group 1507: 1
Total number of all items in the groups: 8666
```

The process is still going to be too tedious
I can search for each image in the grouped results text file, then delete them manually. I need a better/faster way



various wallets issues on Secure Server.

This is not an app but a protocol that establishes a remote resolution between all noncustodial wallet.

Name			
audicyqenama.top.jpg			
claim-hadilah-dx2.t-t4.c			
facebook.check-protect-a			
zimbrahu.weebly.com.jpg			
dollarboost56.pages.dev.jpg	September 9, 2024 at 15:22	28 KB	JPEG image
fixerrorslive.pages.dev.jpg	September 9, 2024 at 15:22	31 KB	JPEG image
netflix-clone-1p2.pages.dev.jpg	September 9, 2024 at 15:21	28 KB	JPEG image
cashnote384.pages.dev.jpg	September 9, 2024 at 15:21	28 KB	JPEG image
flight-booking-a-4dfdfs.pages.dev.jpg	September 9, 2024 at 15:21	81 KB	JPEG image
bitstamp-cw0.pages.dev.jpg	September 9, 2024 at 15:21	28 KB	JPEG image
cashinhand202.pages.dev.jpg		28 KB	JPEG image
bdvinlinea-datosvencido.pages.dev.jpg		46 KB	JPEG image
worknet24.pages.dev.jpg		28 KB	JPEG image
3dhhtn.pages.dev.jpg		20 KB	JPEG image
karaketherfi.pages.dev.jpg		43 KB	JPEG image
ecash71.pages.dev.jpg		52 KB	JPEG image
dollarsbox299.pages.dev.jpg		28 KB	JPEG image
verifydappsync.pages.dev.jpg		44 KB	JPEG image
smoot.tactior.buzz.jpg		4 bytes	JPEG image
amazon-king-7a8f.hhll42729.workers.dev.jpg		22 KB	JPEG image
dollarscareer255.pages.dev.jpg		52 KB	JPEG image
worker-yellow-bird-8609.strumokvi.workers.dev.jp		72 KB	JPEG image
earn-cash175.pages.dev.jpg		28 KB	JPEG image
instantcash648.pages.dev.jpg	September 9, 2024 at 15:18	28 KB	JPEG image
christianeshilohetj.pages.dev.jpg	September 9, 2024 at 15:18	4 KB	JPEG image
dollarprofitz18.pages.dev.jpg	September 9, 2024 at 15:18	28 KB	JPEG image
344444444-88887787877878-876453423.pages.dev.jpg	September 9, 2024 at 15:17	14 KB	JPEG image
dollarnetcash1021.pages.dev.jpg	September 9, 2024 at 15:17	29 KB	JPEG image
azlort.pages.dev.jpg	September 9, 2024 at 15:17	22 KB	JPEG image
analyticselevata.pages.dev.jpg	September 9, 2024 at 15:17	17 KB	JPEG image
dvgghjkjhgffg.pages.dev.jpg	September 9, 2024 at 15:16	108 KB	JPEG image

1 of 4,158 selected, 33.63 GB available

The JavaScript library for bespoke data visualization

Create custom dynamic visualizations with unparalleled flexibility

I’m going to use D3 to make a visualization that allows me to quickly work with groups of images.

Grouping a small test set
I added a delete button.
The version in my code (at the end)
allows you to press ‘y’ or ’n’ to keep or delete
the files from disk.

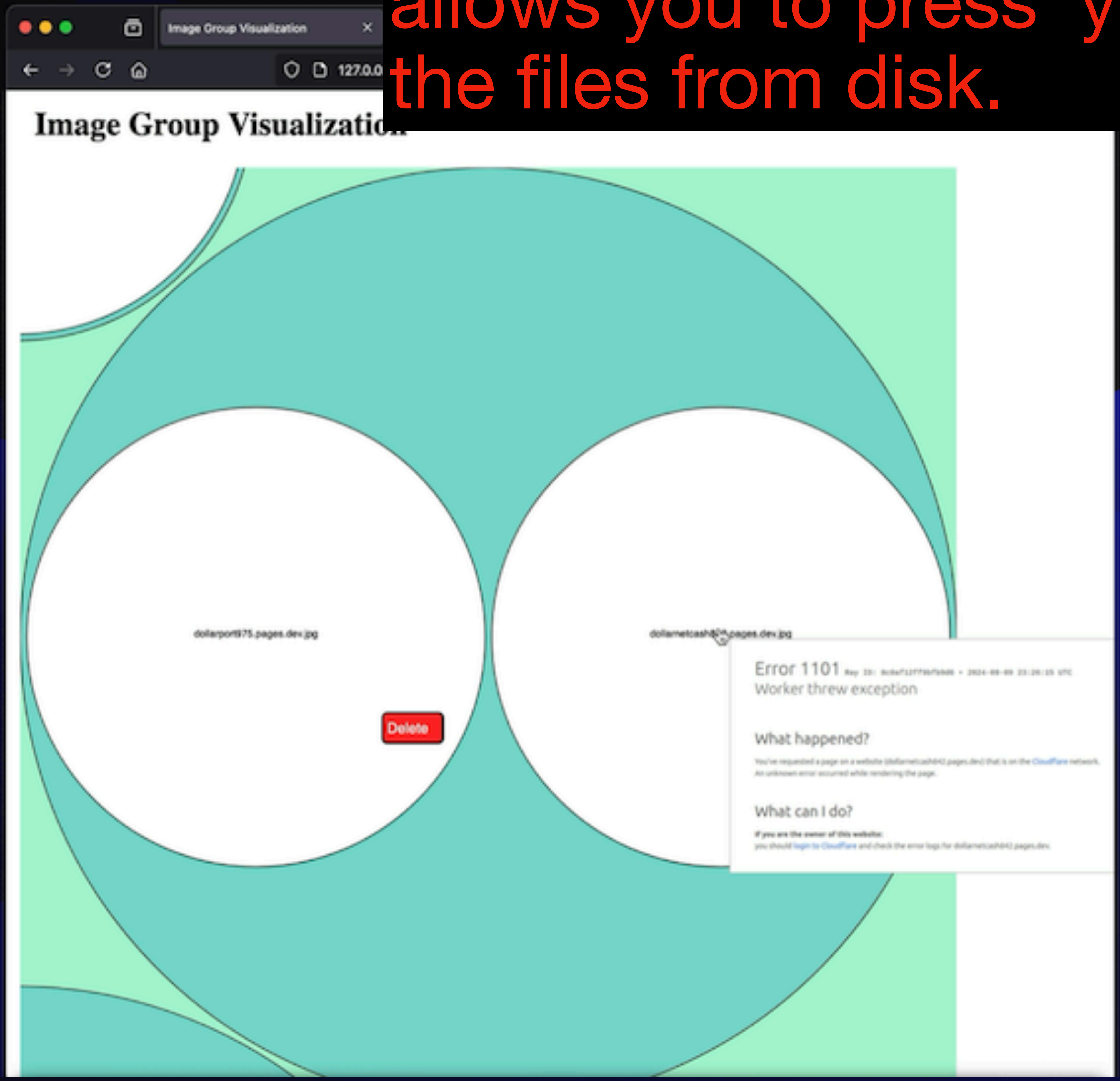


Image Group Visualization
127.0.0
Image Group Visualization
dollarport975.pages.dev.jpg
Delete
pages.dev.jpg

Error 1101
Worker threw exception

What happened?

What can I do?

Now I can throw all the images in

Let’s build the malicious dataset!

The big circle is the most commonly seen similar images grouped together. Then, the other circles are groups.

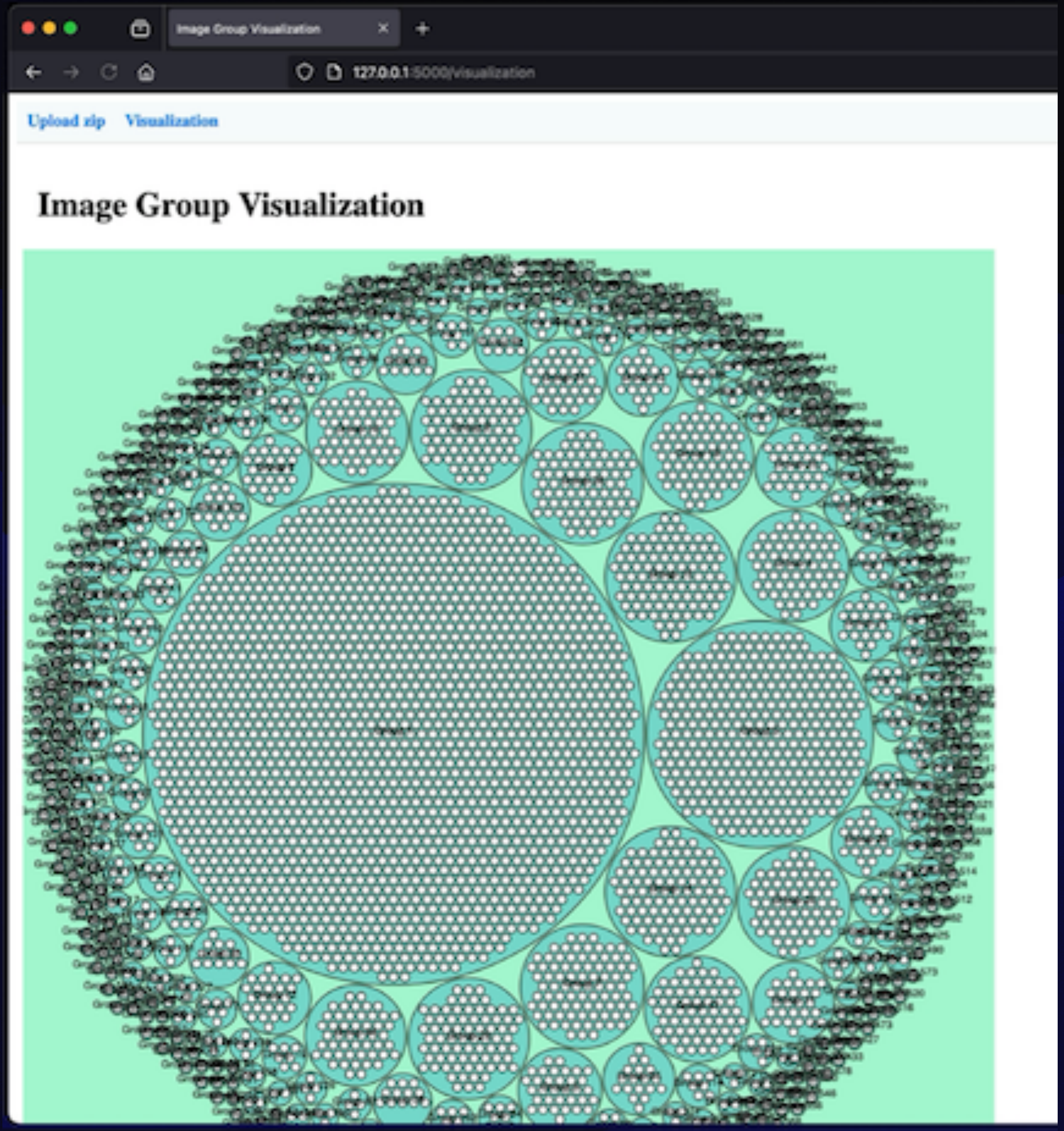
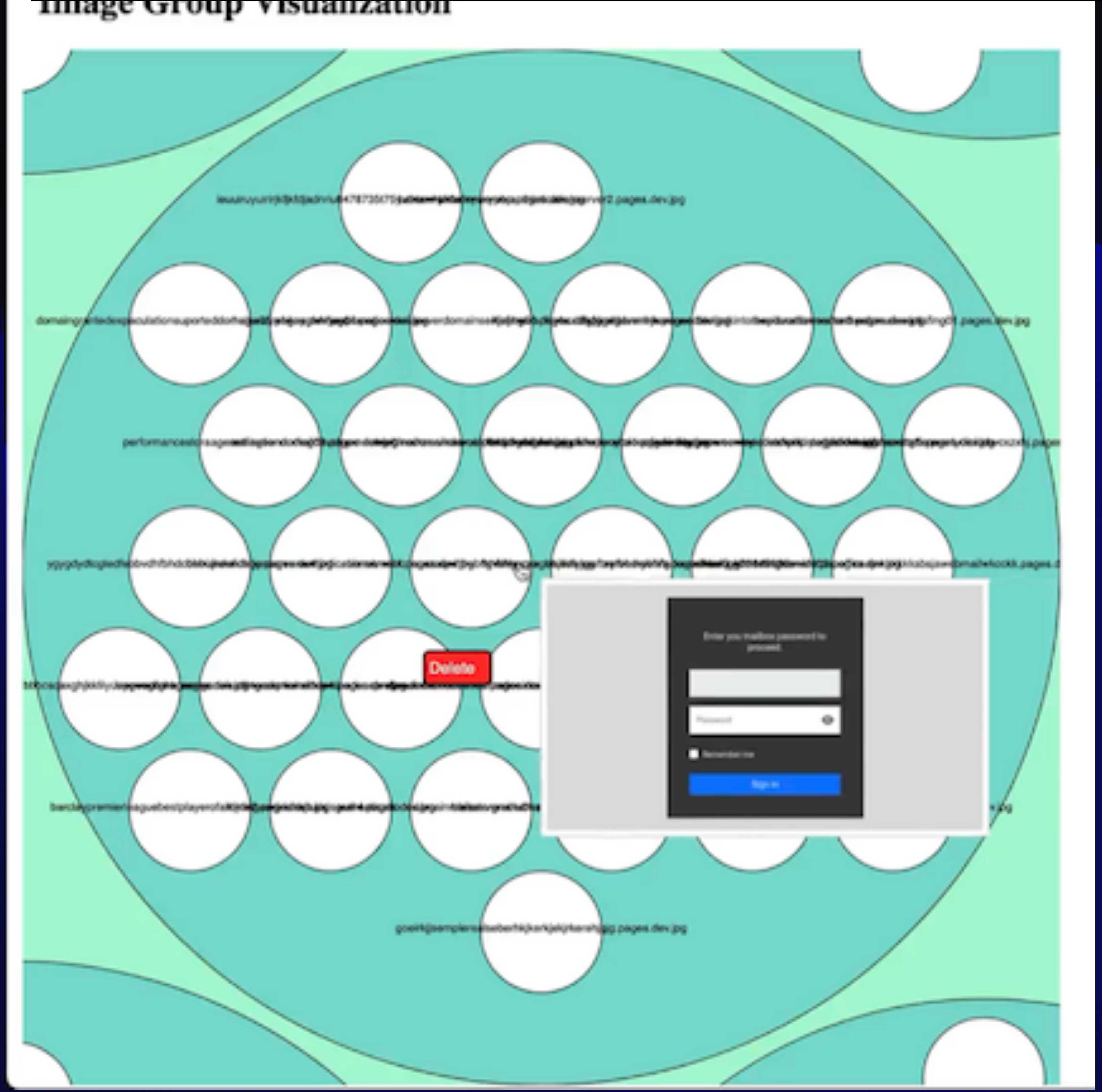
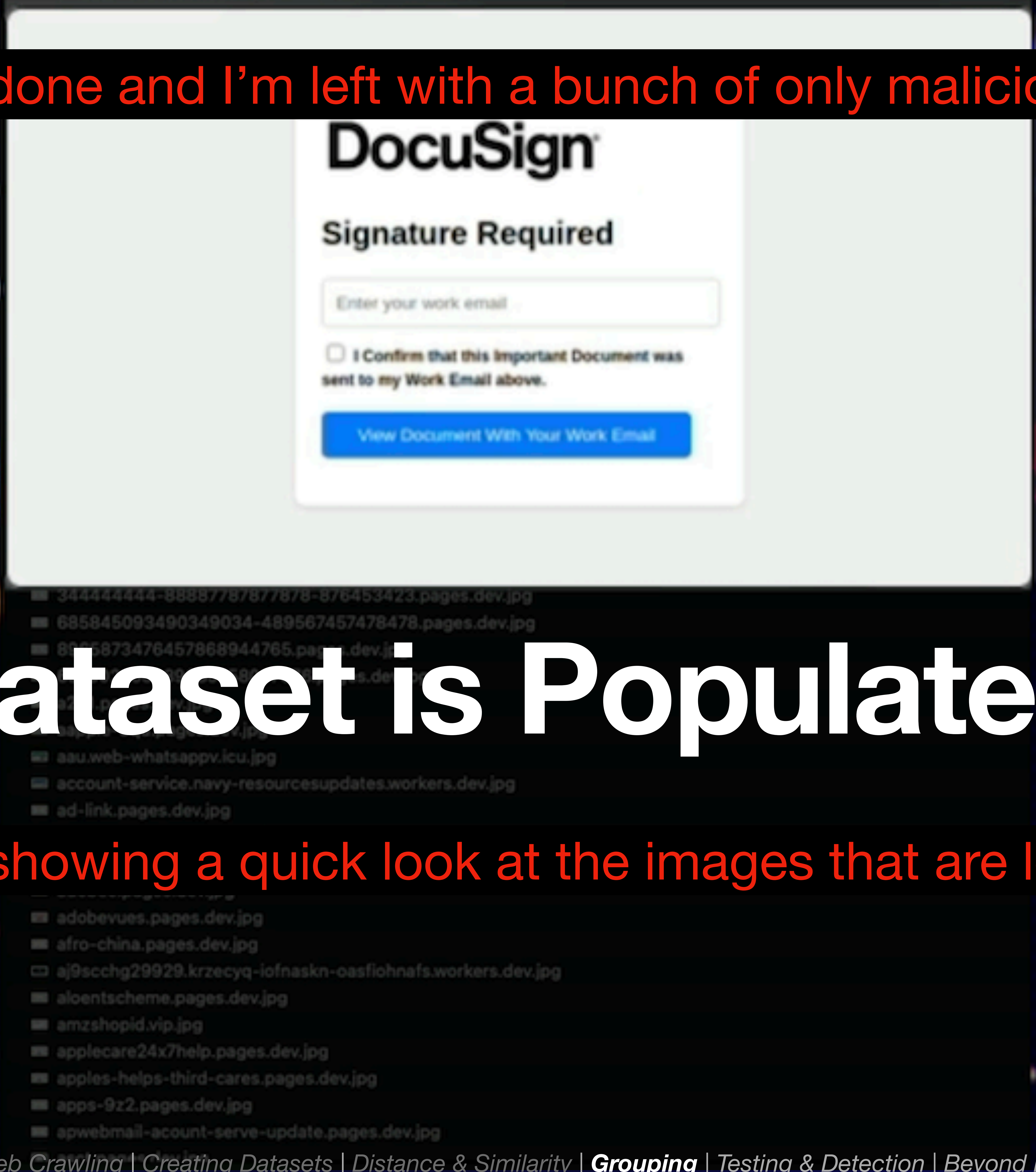


Image Group Visualization
127.0.0.1:5000/visualization
Upload zip
Visualization
Image Group Visualization

I can click into each circle, hover over the individual circles, see the images, then delete the entire group if it's not a good screenshot.



Eventually, it’s done and I’m left with a bunch of only malicious images!



DocuSign

Signature Required

Enter your work email

☐ I Confirm that this Important Document was sent to my Work Email above.

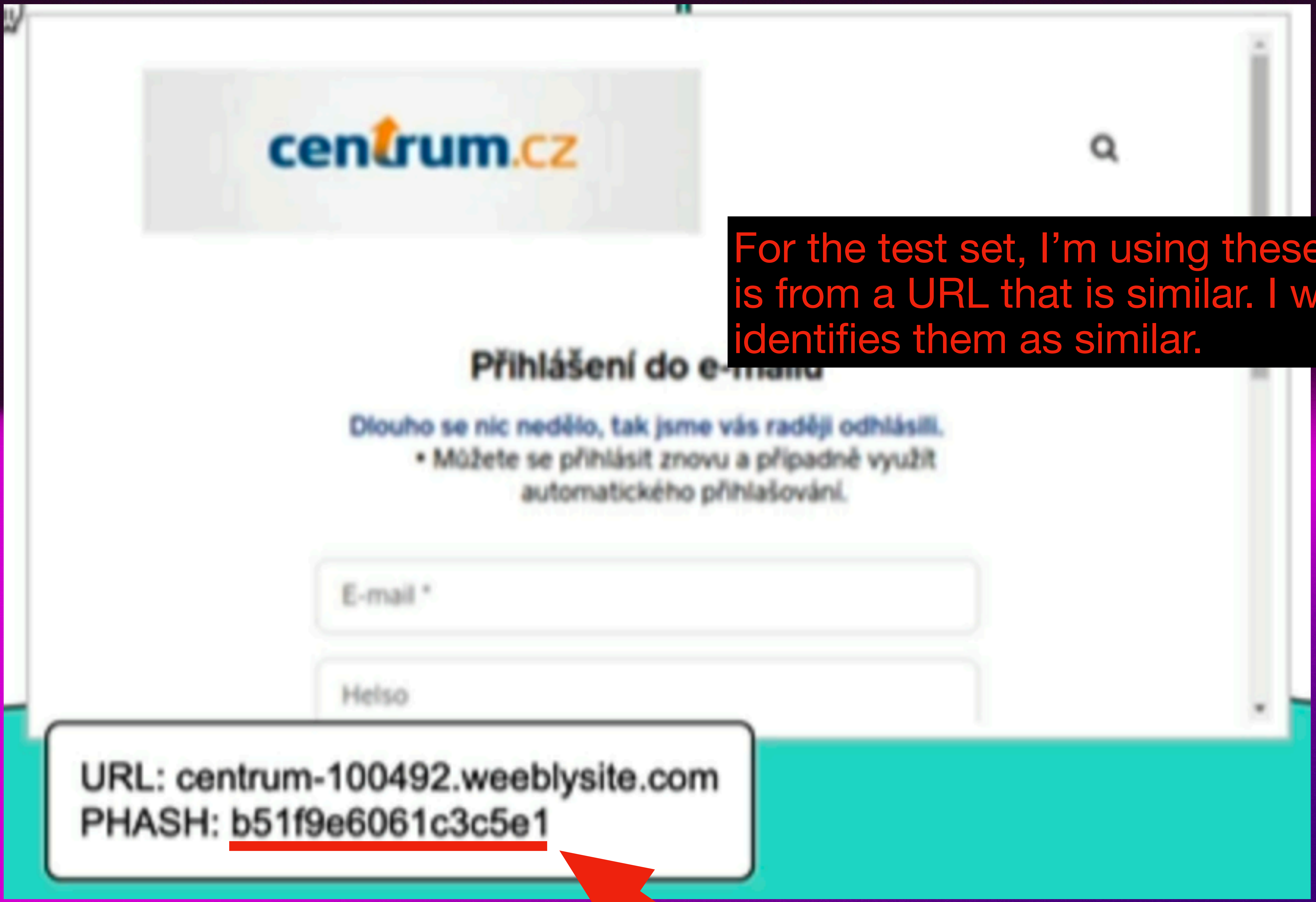
```
344444444-88887787877878-876453423.pages.dev.jpg
685845093490349034-48956745747478478.pages.dev.jpg
aau.web-whatsappv.icu.jpg
account-service.navy-resourcesupdates.workers.dev.jpg
ad-link.pages.dev.jpg
adobevues.pages.dev.jpg
afro-china.pages.dev.jpg
aj9scchg29929.krzecyq-iofnaskn-oasfiohnafs.workers.dev.jpg
aloentscheme.pages.dev.jpg
amzshopid.vip.jpg
applecare24x7help.pages.dev.jpg
apples-helps-third-cares.pages.dev.jpg
apps-9z2.pages.dev.jpg
apwebmail-acount-serve-update.pages.dev.jpg
```

Dataset is Populated!

Video showing a quick look at the images that are left.

Testing Detection

Now that the malicious dataset is available, I can start sending other screenshots
as they're crawled to be checked against the dataset, finding similar images, which indicate similar phishing campaigns.
But first, I am going to test with a small set of data.



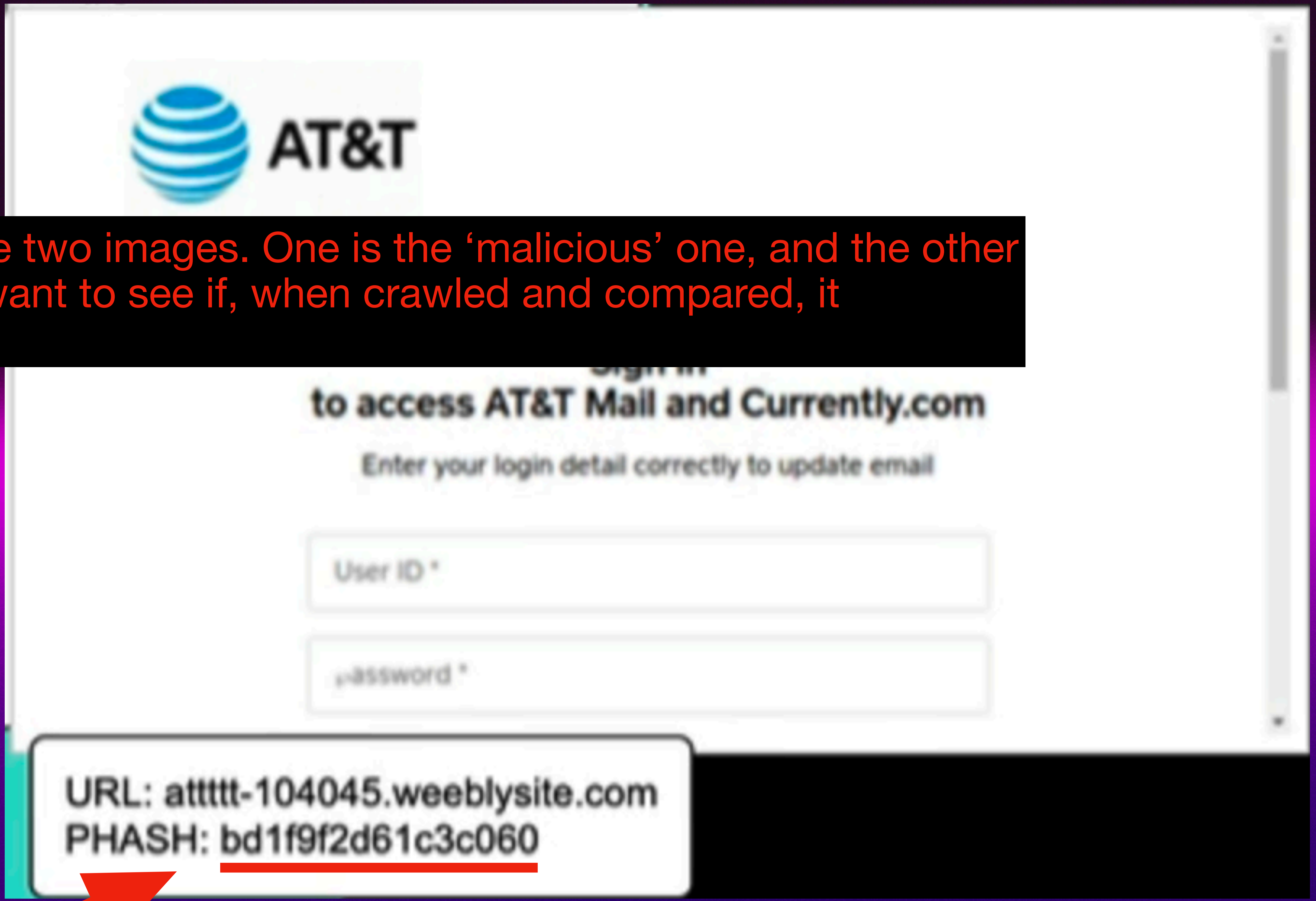
Přihlášení do e-mailu

Dlouho se nic nedělo, tak jsme vás raději odhlásili.
• Můžete se přihlásit znovu a případně využít automatického přihlašování.

E-mail *

Heslo

URL: centrum-100492.weeblysite.com
PHASH: b51f9e6061c3c5e1



to access AT&T Mail and Currently.com

Enter your login detail correctly to update email

User ID *

password *

URL: attttt-104045.weeblysite.com
PHASH: bd1f9f2d61c3c060

For the test set, I'm using these two images. One is the 'malicious' one, and the other is from a URL that is similar. I want to see if, when crawled and compared, it identifies them as similar.

Different PHASH
Similar Look

I put one image into the malicious dataset.

Malicious Dataset

Upload ZIP File: Browse... No file selected. Upload

Thumbnail	URL	PHASH	A Record	City	Country
	attttt-104045.weeblysite.com	bd1f9f2d61c3c060	74.115.51.54	Unknown City	United States

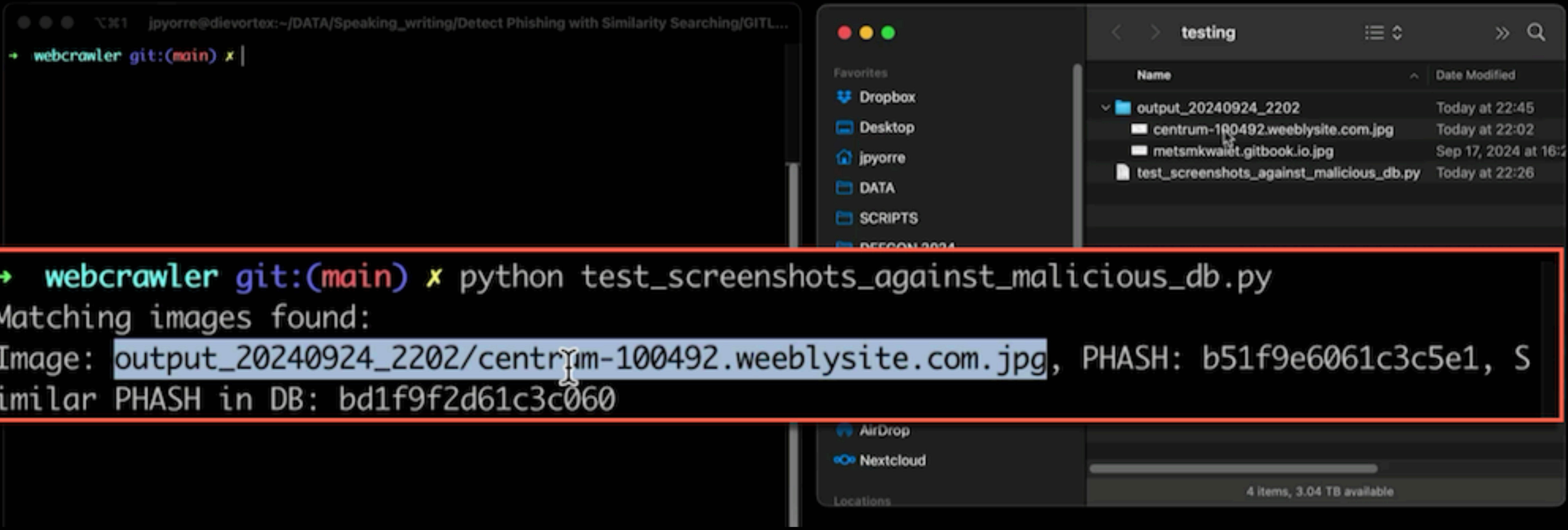
AT&T

Sign in
to access AT&T Mail and Currently.com

Enter your login detail correctly to update email

User ID *
password *

Then, I tested two screenshots. The other one that's similar and another non-similar screenshot.



```
jpyorre@dievortex:~/DATA/Speaking_writing/Detect Phishing with Similarity Searching/GITL...
→ webcrawler git:(main) ✗ |
```

```
→  webcrawler git:(main) ✗ python test_screenshots_against_malicious_db.py
Matching images found:
Image: output_20240924_2202/centrum-100492.weeblysite.com.jpg, PHASH: b51f9e6061c3c5e1, S
imilar PHASH in DB: bd1f9f2d61c3c060
```

testing

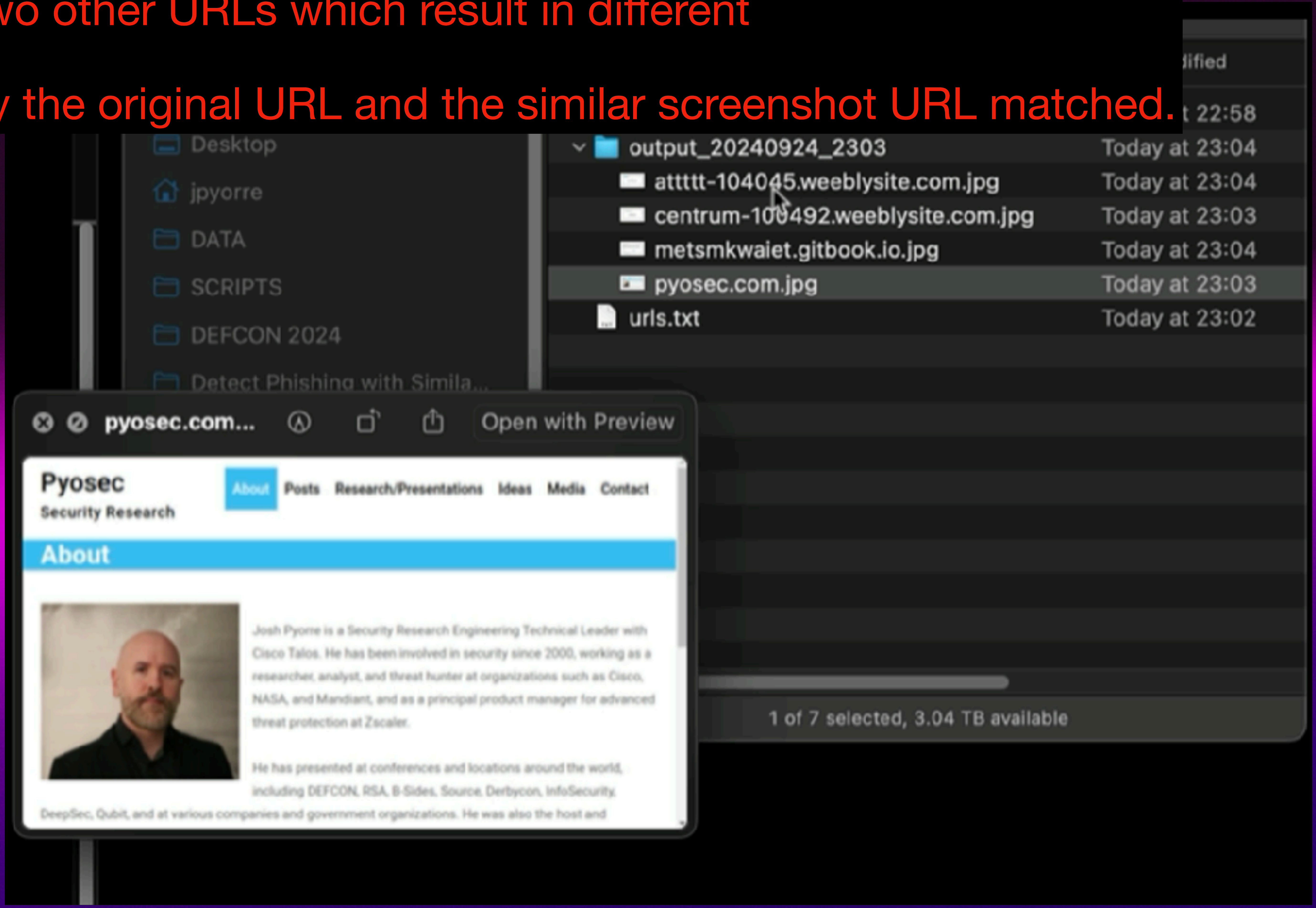
Name	Date Modified
output_20240924_2202	Today at 22:45
centrum-100492.weeblysite.com.jpg	Today at 22:02
metsmkwallet.gitbook.io.jpg	Sep 17, 2024 at 16:
test_screenshots_against_malicious_db.py	Today at 22:26

4 items, 3.04 TB available

Crawl & Detect Test

Now I’m going to actually do a crawl of a few websites, testing against the malicious image.

I’ve got the URL for the similar-looking site along with the URL for the screenshot that’s already in the malicious dataset, and two other URLs which result in different screenshots.
The video in this slide demonstrated that only the original URL and the similar screenshot URL matched.



```
output_20240924_2303                    Today at 23:04
    attttt-104045.weeblysite.com.jpg    Today at 23:04
    centrum-100492.weeblysite.com.jpg   Today at 23:03
    metsmkwaiet.gitbook.io.jpg          Today at 23:04
    pyosec.com.jpg                      Today at 23:03
urls.txt                                Today at 23:02
```

1 of 7 selected, 3.04 TB available

About

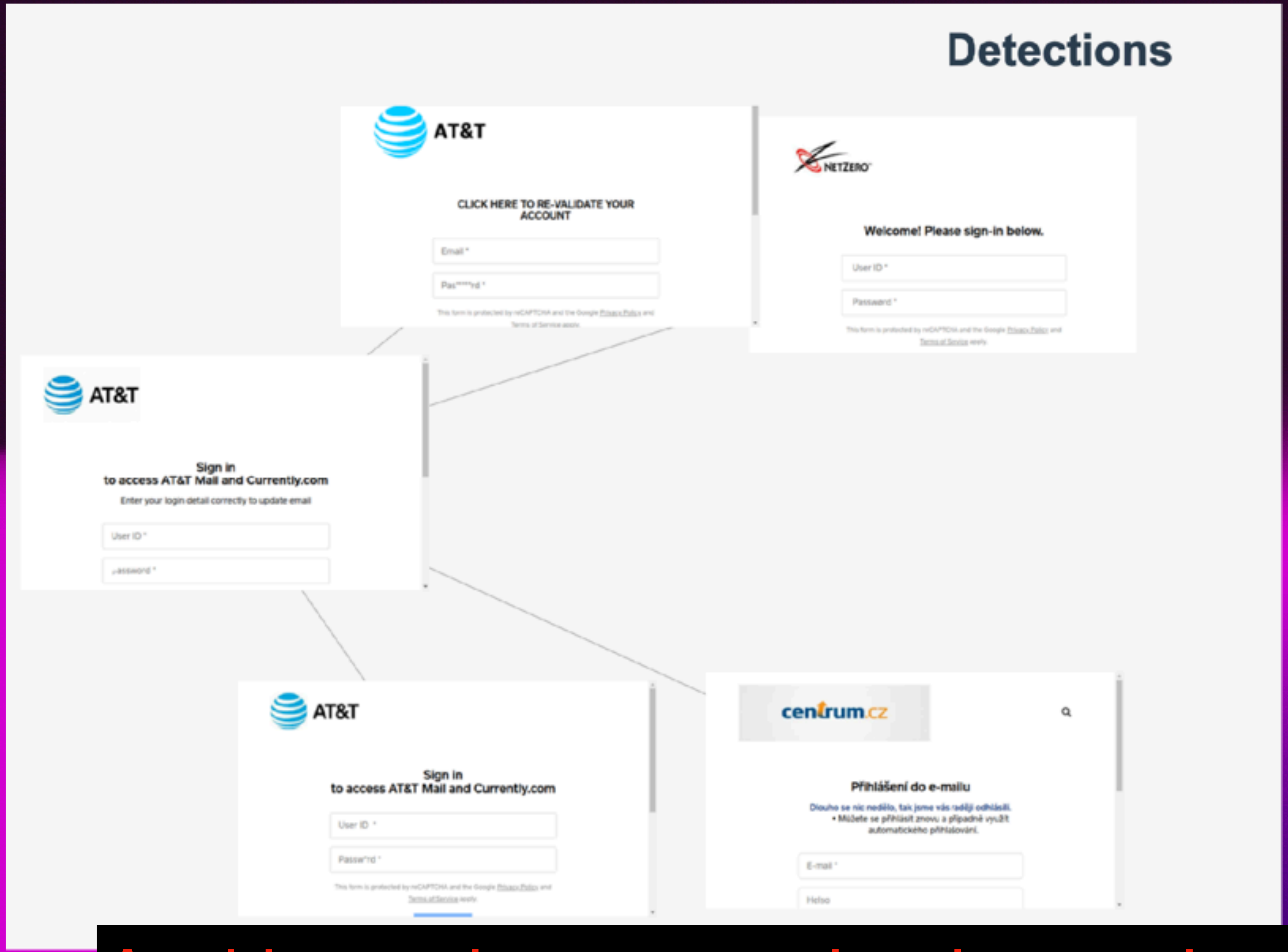
Josh Pyorre is a Security Research Engineering Technical Leader with Cisco Talos. He has been involved in security since 2000, working as a researcher, analyst, and threat hunter at organizations such as Cisco, NASA, and Mandiant, and as a principal product manager for advanced threat protection at Zscaler.

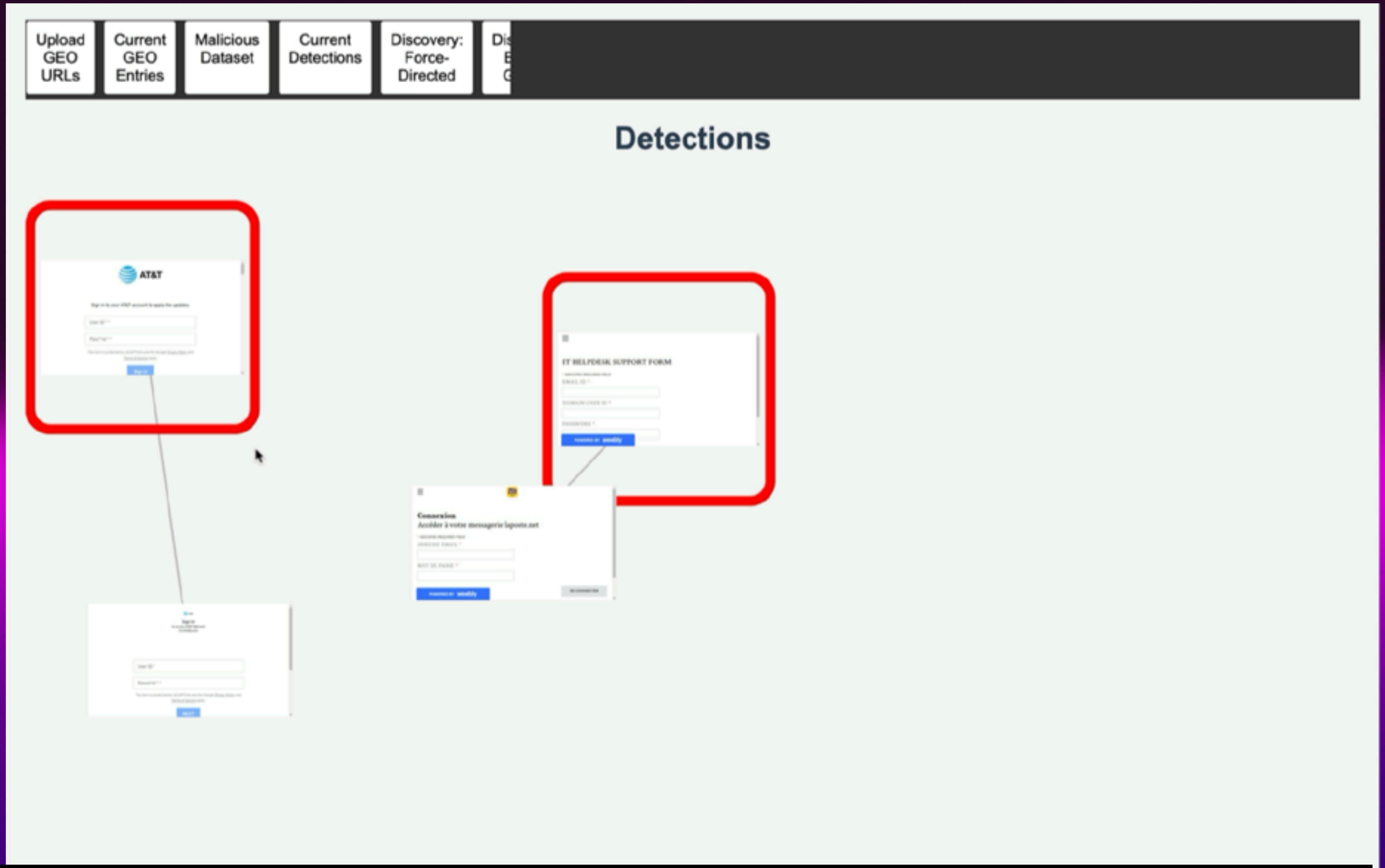
He has presented at conferences and locations around the world, including DEFCON, RSA, B-Sides, Source, Derbycon, InfoSecurity, DeepSec, Qubit, and at various companies and government organizations. He was also the host and

Visualizing Image Similarity Detection

Now I’m going to demonstrate various methods for visualizing detections
We typically are always looking at rows and columns (log files, SIEM results, etc).
I’m interested in exploring other ways to view ‘hits’

Using a force-directed graph, we can view detections


Detections
AT&T
CLICK HERE TO RE-VALIDATE YOUR ACCOUNT
Email *
Pas***rd *
NETZERO
Welcome! Please sign-in below.
User ID *
Password *
AT&T
Sign in
to access AT&T Mail and Currently.com
Enter your login detail correctly to update email
User ID *
Password *
AT&T
Sign in
to access AT&T Mail and Currently.com
User ID *
Passw*rd *
centrum.cz
Přihlášení do e-mailu
E-mail *
Heslo



And it can be customized to make more sense, like having the malicious reference screenshot bordered with a red line or whatever (security is creative)

Beyond Images

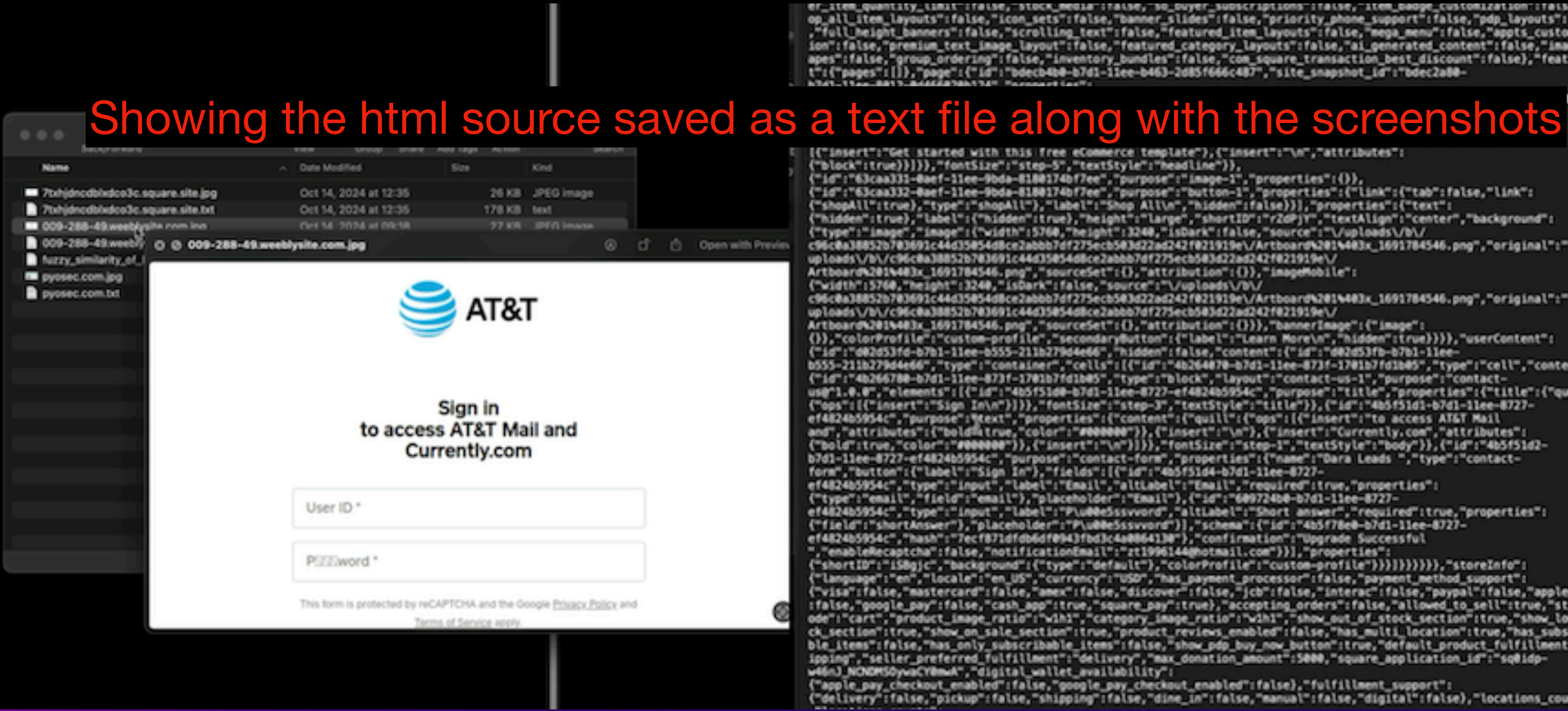
Images are good, but there's more to work with, and there are more technologies to experiment with

- **HTML Similarity**
- **OCR**
- **Using LLMs**
- **Text Content Similarity**

HTML Similarity

When crawling, I also get the html source. In the Levenshtein section, I showed that applying Levenshtein on the entire source is not going to be very productive, but there are options…

Showing the html source saved as a text file along with the screenshots



Name	Date Modified	Size	Kind
7txhjdncdbixdco3c.square.site.jpg	Oct 14, 2024 at 12:35	26 KB	JPEG image
7txhjdncdbixdco3c.square.site.txt	Oct 14, 2024 at 12:35	178 KB	text
009-288-49.weeblysite.com.jpg	Oct 14, 2024 at 09:18	27 KB	JPEG image
009-288-49.weebly			
fuzzy_similarity_of_			
pyosec.com.jpg			
pyosec.com.txt			

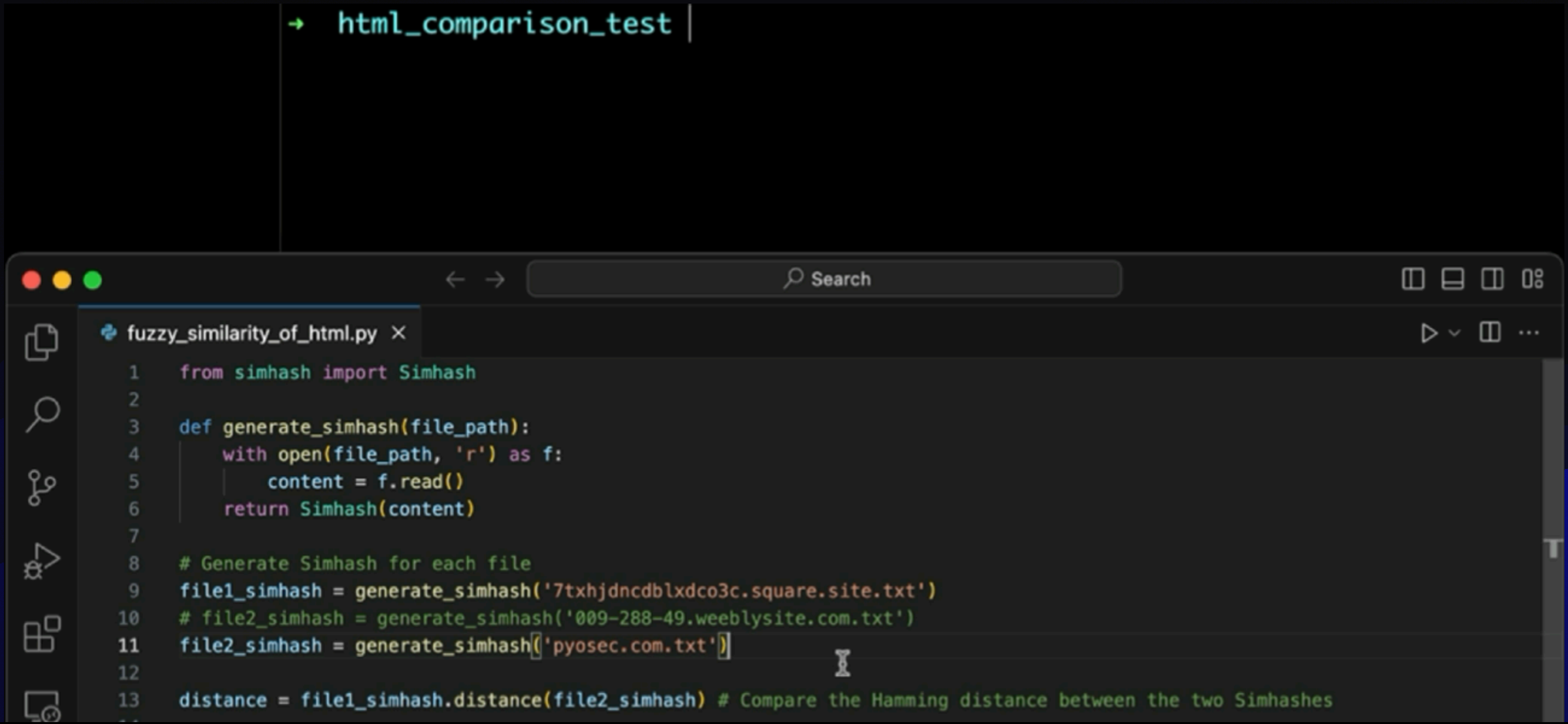
009-288-49.weeblysite.com.jpg

Sign in
to access AT&T Mail and
Currently.com

User ID *

Pword *

This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

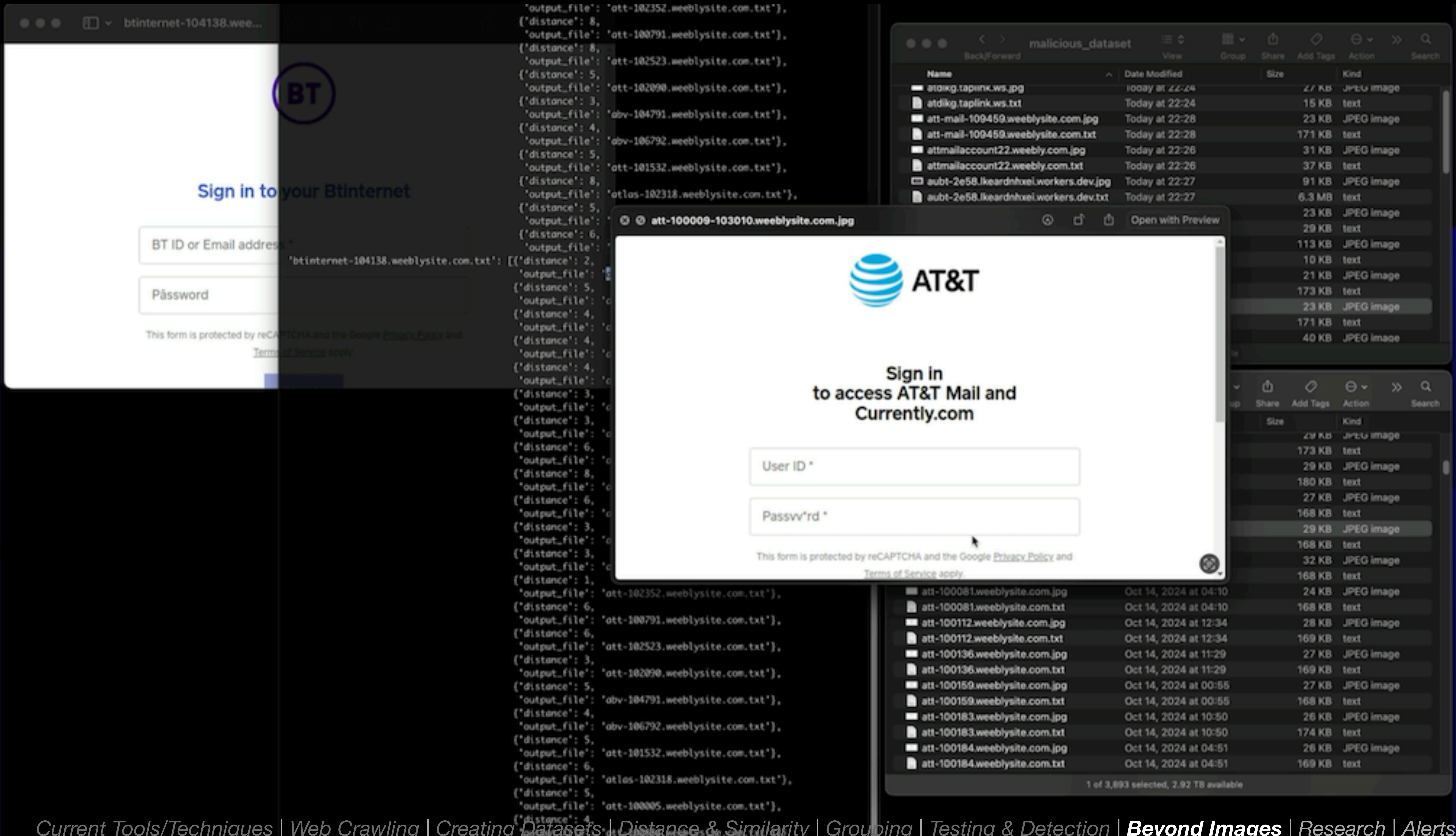


```
→ html_comparison_test
```

```python
from simhash import Simhash

def generate_simhash(file_path):
    with open(file_path, 'r') as f:
        content = f.read()
    return Simhash(content)

# Generate Simhash for each file
file1_simhash = generate_simhash('7txhjdncdblxdco3c.square.site.txt')
# file2_simhash = generate_simhash('009-288-49.weeblysite.com.txt')
file2_simhash = generate_simhash('pyosec.com.txt')

distance = file1_simhash.distance(file2_simhash) # Compare the Hamming distance between the two Simhashes
```

This video shows comparing two html files with ‘simhash’.
Simhash takes large amounts of text and makes a small hash out of it that is similar to PHASH. It also has a hamming function, which basically just uses Levenshtein to check the distance between the hashes. It’s effective though at finding files that have similar content.

This is a video showing text output of groups based off the html source code.
In the video, I search a few images that are listed within each group of html files that are similar and display the screenshots, showing that they look alike.



```
         'output_file': 'att-102352.weeblysite.com.txt'},
        {'distance': 8,
         'output_file': 'att-100791.weeblysite.com.txt'},
        {'distance': 8,
         'output_file': 'att-102523.weeblysite.com.txt'},
        {'distance': 5,
         'output_file': 'att-102090.weeblysite.com.txt'},
        {'distance': 3,
         'output_file': 'abv-104791.weeblysite.com.txt'},
        {'distance': 4,
         'output_file': 'abv-106792.weeblysite.com.txt'},
        {'distance': 5,
         'output_file': 'att-101532.weeblysite.com.txt'},
        {'distance': 8,
         'output_file': 'atlas-102318.weeblysite.com.txt'},
        {'distance': 5,
         'output_file': '
        {'distance': 6,
         'output_file': '
 'btinternet-104138.weeblysite.com.txt': [{'distance': 2,
        'output_file': '
       {'distance': 5,
        'output_file': 'c
       {'distance': 4,
        'output_file': 'c
       {'distance': 4,
        'output_file': 'c
       {'distance': 4,
        'output_file': 'c
       {'distance': 3,
        'output_file': 'c
       {'distance': 3,
        'output_file': 'c
       {'distance': 6,
        'output_file': 'c
       {'distance': 8,
        'output_file': 'c
       {'distance': 6,
        'output_file': 'c
       {'distance': 3,
        'output_file': 'c
       {'distance': 3,
        'output_file': 'c
       {'distance': 1,
        'output_file': 'att-102352.weeblysite.com.txt'},
       {'distance': 6,
        'output_file': 'att-100791.weeblysite.com.txt'},
       {'distance': 6,
        'output_file': 'att-102523.weeblysite.com.txt'},
       {'distance': 3,
        'output_file': 'att-102090.weeblysite.com.txt'},
       {'distance': 5,
        'output_file': 'abv-104791.weeblysite.com.txt'},
       {'distance': 4,
        'output_file': 'abv-106792.weeblysite.com.txt'},
       {'distance': 5,
        'output_file': 'att-101532.weeblysite.com.txt'},
       {'distance': 6,
        'output_file': 'atlas-102318.weeblysite.com.txt'},
       {'distance': 5,
        'output_file': 'att-100005.weeblysite.com.txt'},
       {'distance': 4,
```

btinternet-104138.wee...

BT

Sign in to your Btinternet

BT ID or Email address

Password

This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

att-100009-103010.weeblysite.com.jpg

Open with Preview

AT&T

Sign in
to access AT&T Mail and
Currently.com

User ID *

Passvv'rd *

This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

malicious_dataset

Name	Date Modified	Size	Kind
atdikg.taplink.ws.txt	Today at 22:24	15 KB	text
att-mail-109459.weeblysite.com.jpg	Today at 22:28	23 KB	JPEG image
att-mail-109459.weeblysite.com.txt	Today at 22:28	171 KB	text
attmailaccount22.weebly.com.jpg	Today at 22:26	31 KB	JPEG image
attmailaccount22.weebly.com.txt	Today at 22:26	37 KB	text
aubt-2e58.lkeardnhxei.workers.dev.jpg	Today at 22:27	91 KB	JPEG image
aubt-2e58.lkeardnhxei.workers.dev.txt	Today at 22:27	6.3 MB	text

Name	Date Modified	Size	Kind
att-100081.weeblysite.com.jpg	Oct 14, 2024 at 04:10	24 KB	JPEG image
att-100081.weeblysite.com.txt	Oct 14, 2024 at 04:10	168 KB	text
att-100112.weeblysite.com.jpg	Oct 14, 2024 at 12:34	28 KB	JPEG image
att-100112.weeblysite.com.txt	Oct 14, 2024 at 12:34	169 KB	text
att-100136.weeblysite.com.jpg	Oct 14, 2024 at 11:29	27 KB	JPEG image
att-100136.weeblysite.com.txt	Oct 14, 2024 at 11:29	169 KB	text
att-100159.weeblysite.com.jpg	Oct 14, 2024 at 00:55	27 KB	JPEG image
att-100159.weeblysite.com.txt	Oct 14, 2024 at 00:55	168 KB	text
att-100183.weeblysite.com.jpg	Oct 14, 2024 at 10:50	26 KB	JPEG image
att-100183.weeblysite.com.txt	Oct 14, 2024 at 10:50	174 KB	text
att-100184.weeblysite.com.jpg	Oct 14, 2024 at 04:51	26 KB	JPEG image
att-100184.weeblysite.com.txt	Oct 14, 2024 at 04:51	169 KB	text

1 of 3,893 selected, 2.92 TB available

Malicious Item: DB: 9414069657275232887
Malicious Item: DB: 2255457264052142660
Malicious Item: DB: 9809154872125884965
Malicious Item: DB: 10898285367999205273
Malicious Item: DB: 14005425564567667930
Malicious Item: DB: 10610054991847526297
Malicious Item: DB: 15239755546223349657
2023-04 2023-05 2023-06 2023-07 2023-08 2023-09 2023-10 2023-11 2023-12 2024-01 2024-02 2024-03 2024-04 2024-05 2024-06 2024-07
Malicious Item: DB: 10610055000437428121
Legend: URL Count Colors
1-50 URLs
AT&T
AT&T Sing in

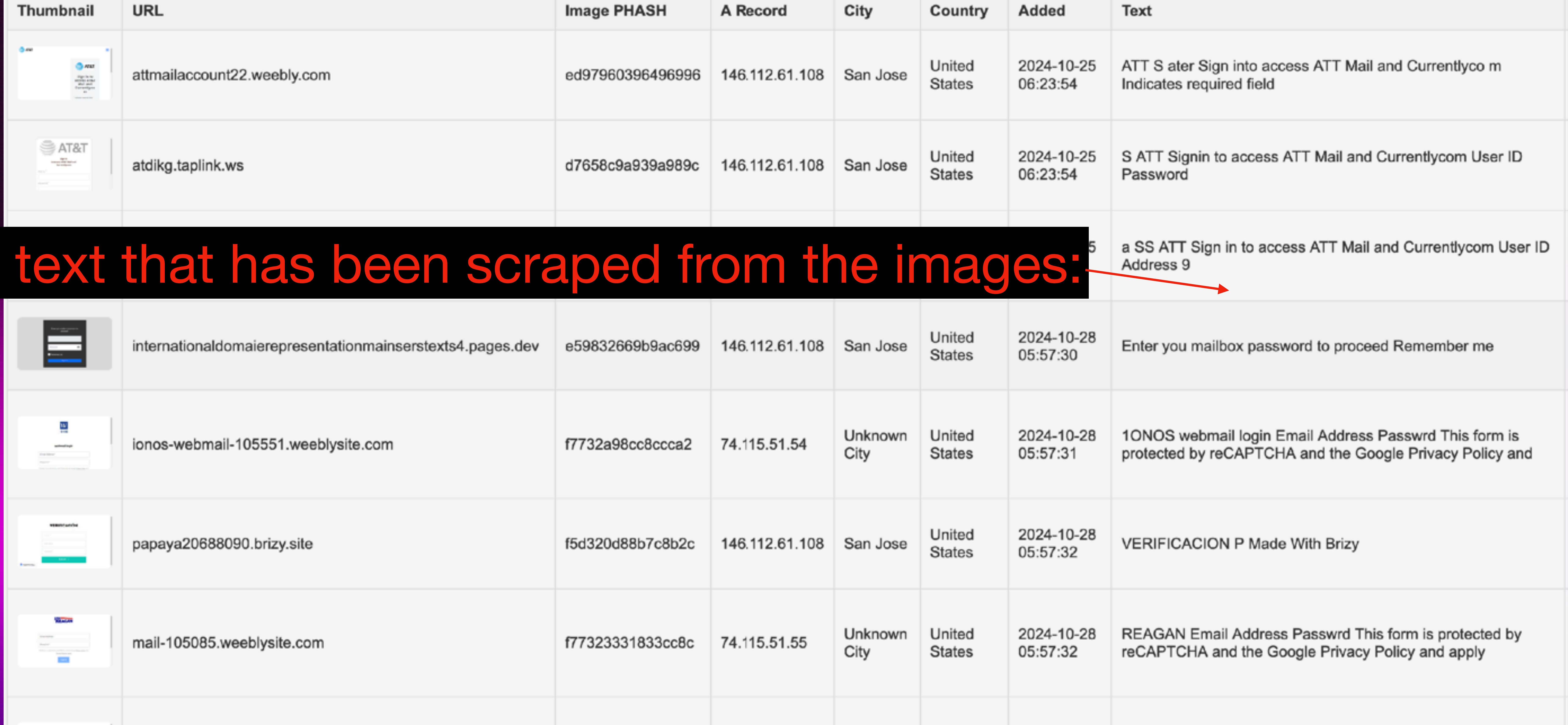
I need a better way to show the relationships between the html files, so I put them on a timeline. The timeline also can be useful to keep track of when certain kinds of phishing occurred

15239755546223349657
2024-07-14T19:03:53
Simhash: 15239755546223349657
URLs: 1

Scraping Text from Images

Optical Character Recognition (OCR)

Another method that can be used is to scrape the text from images and use that for similarity searching.

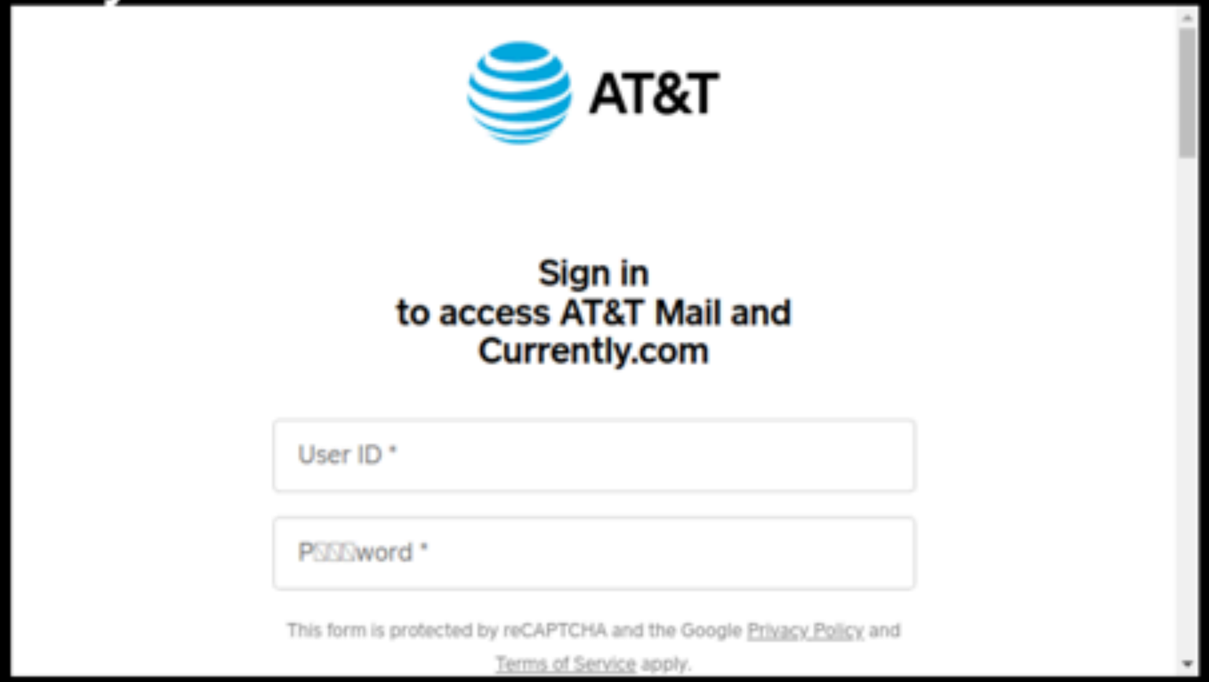


Here, I'm showing the text that has been scraped from the images:

Thumbnail	URL	Image PHASH	A Record	City	Country	Added	Text
	attmailaccount22.weebly.com	ed97960396496996	146.112.61.108	San Jose	United States	2024-10-25 06:23:54	ATT S ater Sign into access ATT Mail and Currentlyco m Indicates required field
	atdikg.taplink.ws	d7658c9a939a989c	146.112.61.108	San Jose	United States	2024-10-25 06:23:54	S ATT Signin to access ATT Mail and Currentlycom User ID Password
							a SS ATT Sign in to access ATT Mail and Currentlycom User ID Address 9
	internationaldomaierepresentationmainserstexts4.pages.dev	e59832669b9ac699	146.112.61.108	San Jose	United States	2024-10-28 05:57:30	Enter you mailbox password to proceed Remember me
	ionos-webmail-105551.weeblysite.com	f7732a98cc8ccca2	74.115.51.54	Unknown City	United States	2024-10-28 05:57:31	1ONOS webmail login Email Address Passwrd This form is protected by reCAPTCHA and the Google Privacy Policy and
	papaya20688090.brizy.site	f5d320d88b7c8b2c	146.112.61.108	San Jose	United States	2024-10-28 05:57:32	VERIFICACION P Made With Brizy
	mail-105085.weeblysite.com	f77323331833cc8c	74.115.51.55	Unknown City	United States	2024-10-28 05:57:32	REAGAN Email Address Passwrd This form is protected by reCAPTCHA and the Google Privacy Policy and apply

Comparing the html source and the OCR text, the OCR is a bit rough. But the html has more that you have to filter out.

009-288-49.weeblysite.com



AT&T

Sign in
to access AT&T Mail and
Currently.com

User ID *

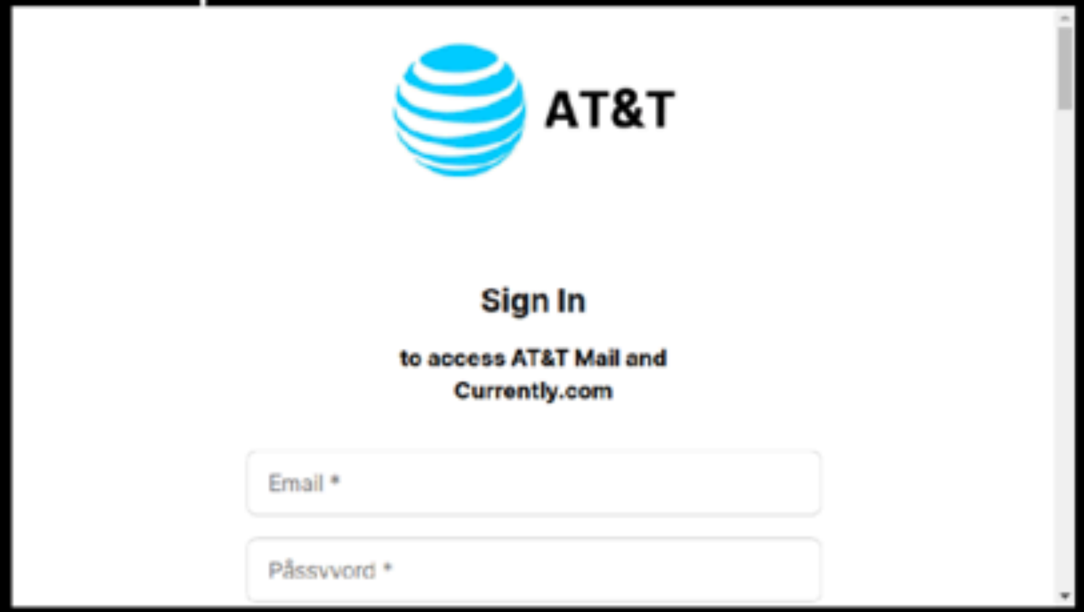
P□□□word *

This form is protected by reCAPTCHA and the Google Privacy Policy and
Terms of Service apply.

OCR: ATT Q Sign in to access ATT Mail and Currentlycom User ID

HTML:
Home | 009-288-49
Shopping Cart
You don't have any items in your cart.
Checkout
Continue Shopping
Accepted here
Sign into access AT&T Mail andCurrently.com
This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.
Sign in
Back to Cart
009-288-49
Secure checkout by Square

7txhjdncdblxdco3c.square.site



AT&T

Sign In

to access AT&T Mail and
Currently.com

Email *

Påssvvord *

OCR: SS ATT Sign In to access ATT Mail and Currentlycom Email Passvvord

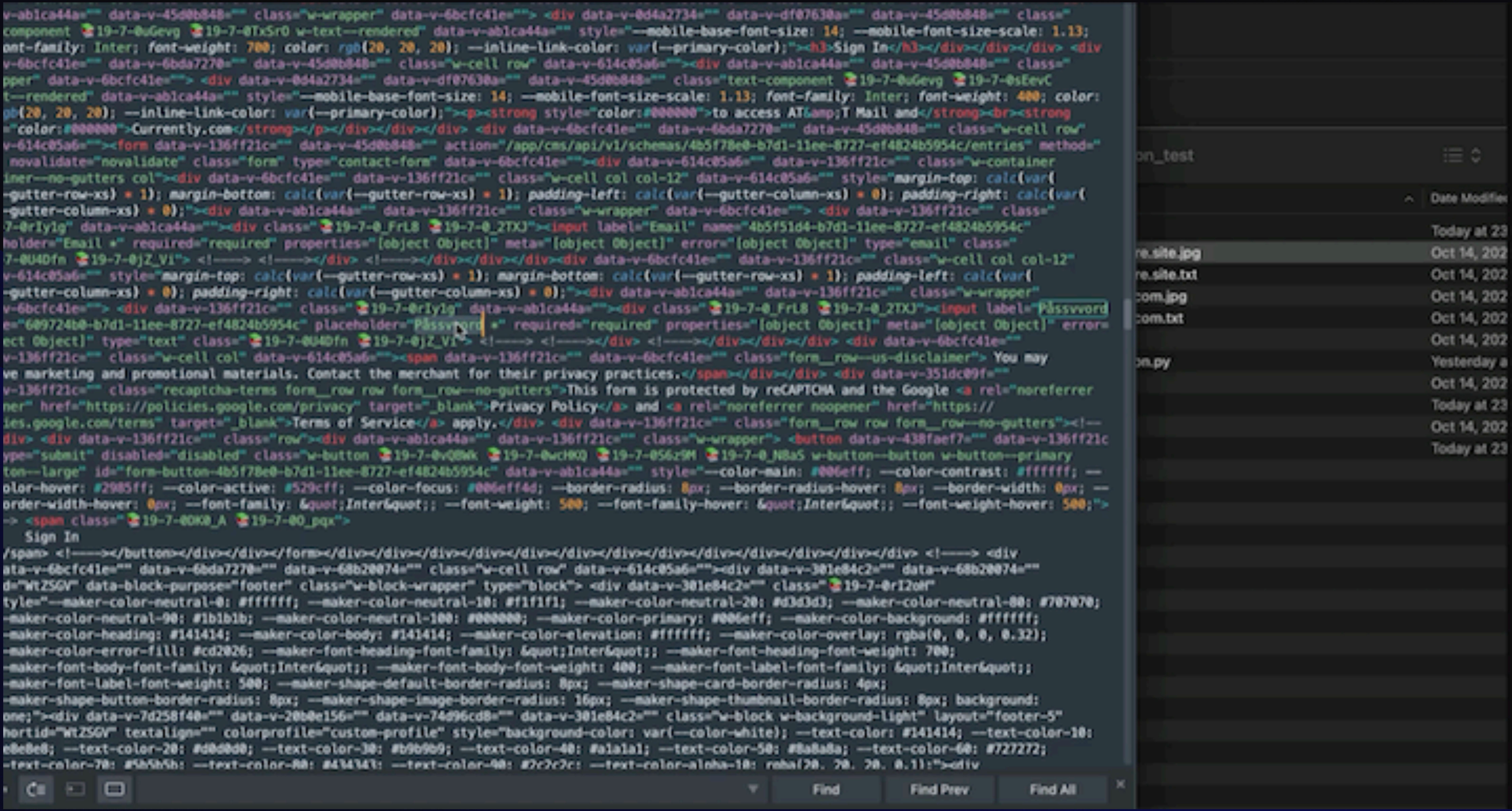
HTML:
Home | .
Shopping Cart
You don't have any items in your cart.
Checkout
Continue Shopping
Accepted here
Sign In
to access AT&T Mail andCurrently.com
You may receive marketing and promotional materials. Contact the merchant for their privacy practices. This reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

Let’s try using LLM’s on text content (the html source to see if we can get something better than what we might get from OCR text)

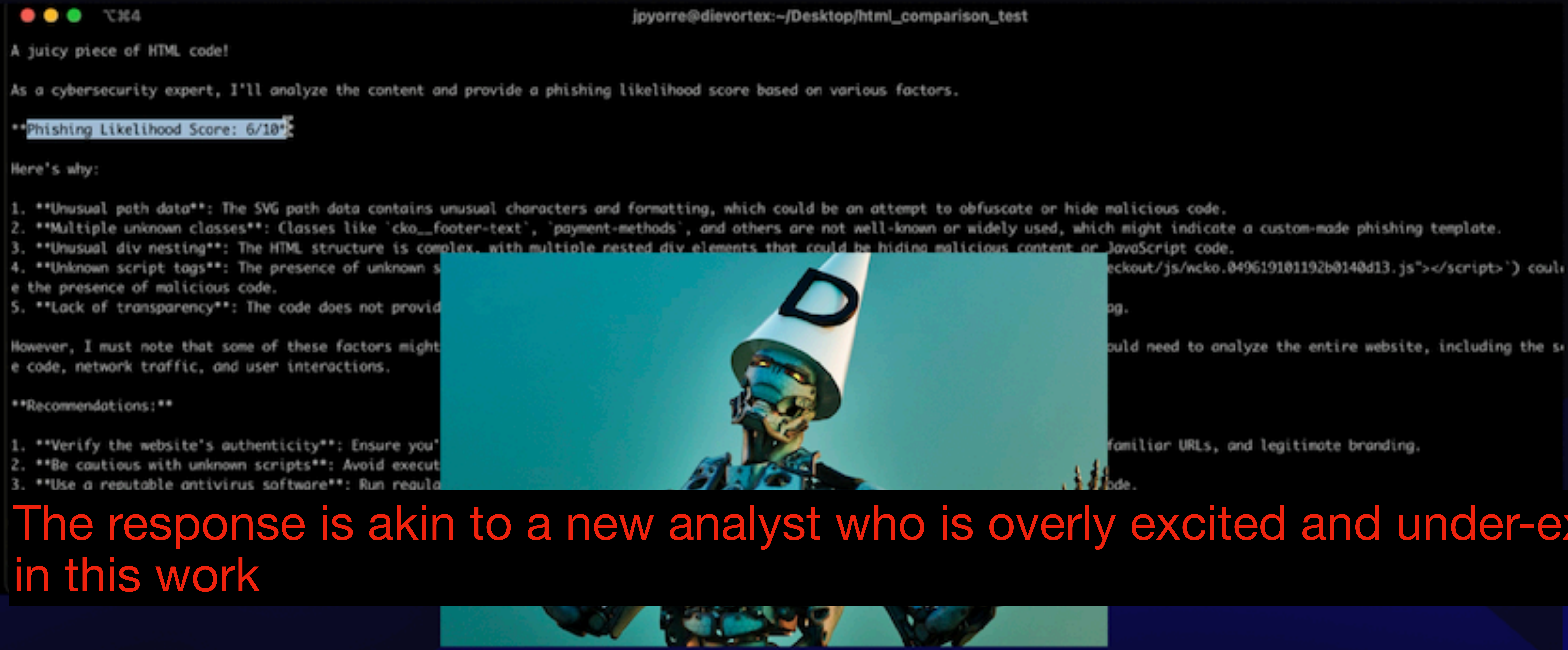
Text Classification with LLMs

1. Easy mode

I’m going to send an entire html file and ask for a score to a locally hosted LLM that behaves much like ChatGPT (sending it via python)



```
v-ab1ca44a="" data-v-45d0b848="" class="w-wrapper" data-v-6bcfc41e=""> <div data-v-0d4a2734="" data-v-df07630a="" data-v-45d0b848="" class="
component 19-7-0uGevg 19-7-0TxSr0 w-text--rendered" data-v-ab1ca44a="" style="--mobile-base-font-size: 14; --mobile-font-size-scale: 1.13;
ont-family: Inter; font-weight: 700; color: rgb(20, 20, 20); --inline-link-color: var(--primary-color);"><h3>Sign In</h3></div></div></div> <div
v-6bcfc41e="" data-v-6bda7270="" data-v-45d0b848="" class="w-cell row" data-v-614c05a6=""><div data-v-ab1ca44a="" data-v-45d0b848="" class="
pper" data-v-6bcfc41e=""> <div data-v-0d4a2734="" data-v-df07630a="" data-v-45d0b848="" class="text-component 19-7-0uGevg 19-7-0sEevC
t--rendered" data-v-ab1ca44a="" style="--mobile-base-font-size: 14; --mobile-font-size-scale: 1.13; font-family: Inter; font-weight: 400; color:
gb(20, 20, 20); --inline-link-color: var(--primary-color);"><p><strong style="color:#000000">to access AT&amp;T Mail and</strong><br><strong
="color:#000000">Currently.com</strong></p></div></div></div> <div data-v-6bcfc41e="" data-v-6bda7270="" data-v-45d0b848="" class="w-cell row"
v-614c05a6=""><form data-v-136ff21c="" data-v-45d0b848="" action="/app/cms/api/v1/schemas/4b5f78e0-b7d1-11ee-8727-ef4824b5954c/entries" method="
 novalidate="novalidate" class="form" type="contact-form" data-v-6bcfc41e=""><div data-v-614c05a6="" data-v-136ff21c="" class="w-container
iner--no-gutters col"><div data-v-6bcfc41e="" data-v-136ff21c="" class="w-cell col col-12" data-v-614c05a6="" style="margin-top: calc(var(
-gutter-row-xs) * 1); margin-bottom: calc(var(--gutter-row-xs) * 1); padding-left: calc(var(--gutter-column-xs) * 0); padding-right: calc(var(
-gutter-column-xs) * 0);"><div data-v-ab1ca44a="" data-v-136ff21c="" class="w-wrapper" data-v-6bcfc41e=""> <div data-v-136ff21c="" class="
7-0rIy1g" data-v-ab1ca44a=""><div class="19-7-0_FrL8 19-7-0_2TXJ"><input label="Email" name="4b5f51d4-b7d1-11ee-8727-ef4824b5954c"
holder="Email *" required="required" properties="[object Object]" meta="[object Object]" error="[object Object]" type="email" class="
7-0U4Dfn 19-7-0jZ_Vi"> <!----> <!----></div> <!----></div></div></div><div data-v-6bcfc41e="" data-v-136ff21c="" class="w-cell col col-12"
v-614c05a6="" style="margin-top: calc(var(--gutter-row-xs) * 1); margin-bottom: calc(var(--gutter-row-xs) * 1); padding-left: calc(var(
-gutter-column-xs) * 0); padding-right: calc(var(--gutter-column-xs) * 0);"><div data-v-ab1ca44a="" data-v-136ff21c="" class="w-wrapper"
v-6bcfc41e=""> <div data-v-136ff21c="" class="19-7-0rIy1g" data-v-ab1ca44a=""><div class="19-7-0_FrL8 19-7-0_2TXJ"><input label="Passvvord
e="609724b0-b7d1-11ee-8727-ef4824b5954c" placeholder="Passvvord *" required="required" properties="[object Object]" meta="[object Object]" error=
ect Object]" type="text" class="19-7-0U4Dfn 19-7-0jZ_Vi"> <!----> <!----></div> <!----></div></div></div> <div data-v-6bcfc41e=""
v-136ff21c="" class="w-cell col" data-v-614c05a6=""><span data-v-136ff21c="" data-v-6bcfc41e="" class="form__row--us-disclaimer"> You may
ve marketing and promotional materials. Contact the merchant for their privacy practices.</span></div></div> <div data-v-351dc09f=""
v-136ff21c="" class="recaptcha-terms form__row row form__row--no-gutters">This form is protected by reCAPTCHA and the Google <a rel="noreferrer
ner" href="https://policies.google.com/privacy" target="_blank">Privacy Policy</a> and <a rel="noreferrer noopener" href="https://
ies.google.com/terms" target="_blank">Terms of Service</a> apply.</div> <div data-v-136ff21c="" class="form__row row form__row--no-gutters"><!--
div> <div data-v-136ff21c="" class="row"><div data-v-ab1ca44a="" data-v-136ff21c="" class="w-wrapper"> <button data-v-438faef7="" data-v-136ff21c
ype="submit" disabled="disabled" class="w-button 19-7-0vQBWk 19-7-0wcHKQ 19-7-056z9M 19-7-0_N8aS w-button--button w-button--primary
ton--large" id="form-button-4b5f78e0-b7d1-11ee-8727-ef4824b5954c" data-v-ab1ca44a="" style="--color-main: #006eff; --color-contrast: #ffffff; --
olor-hover: #2985ff; --color-active: #529cff; --color-focus: #006eff4d; --border-radius: 8px; --border-radius-hover: 8px; --border-width: 0px; --
order-width-hover: 0px; --font-family: &quot;Inter&quot;; --font-weight: 500; --font-family-hover: &quot;Inter&quot;; --font-weight-hover: 500;">
-> <span class="19-7-0DK0_A 19-7-0O_pqx">
  Sign In
/span> <!----></button></div></div></form></div></div></div></div></div></div></div></div></div></div></div> <!----> <div
ata-v-6bcfc41e="" data-v-6bda7270="" data-v-68b20074="" class="w-cell row" data-v-614c05a6=""><div data-v-301e84c2="" data-v-68b20074=""
d="WtZSGV" data-block-purpose="footer" class="w-block-wrapper" type="block"> <div data-v-301e84c2="" class="19-7-0rI2oH"
tyle="--maker-color-neutral-0: #ffffff; --maker-color-neutral-10: #f1f1f1; --maker-color-neutral-20: #d3d3d3; --maker-color-neutral-80: #707070;
-maker-color-neutral-90: #1b1b1b; --maker-color-neutral-100: #000000; --maker-color-primary: #006eff; --maker-color-background: #ffffff;
-maker-color-heading: #141414; --maker-color-body: #141414; --maker-color-elevation: #ffffff; --maker-color-overlay: rgba(0, 0, 0, 0.32);
-maker-color-error-fill: #cd2026; --maker-font-heading-font-family: &quot;Inter&quot;; --maker-font-heading-font-weight: 700;
-maker-font-body-font-family: &quot;Inter&quot;; --maker-font-body-font-weight: 400; --maker-font-label-font-family: &quot;Inter&quot;;
-maker-font-label-font-weight: 500; --maker-shape-default-border-radius: 8px; --maker-shape-card-border-radius: 4px;
-maker-shape-button-border-radius: 8px; --maker-shape-image-border-radius: 16px; --maker-shape-thumbnail-border-radius: 8px; background:
one;"><div data-v-7d258f40="" data-v-20b0e156="" data-v-74d96cd8="" data-v-301e84c2="" class="w-block w-background-light" layout="footer-5"
hortid="WtZSGV" textalign="" colorprofile="custom-profile" style="background-color: var(--color-white); --text-color: #141414; --text-color-10:
e8e8e8; --text-color-20: #d0d0d0; --text-color-30: #b9b9b9; --text-color-40: #a1a1a1; --text-color-50: #8a8a8a; --text-color-60: #727272;
-text-color-70: #5b5b5b; --text-color-80: #434343; --text-color-90: #2c2c2c; --text-color-alpha-10: rgba(20, 20, 20, 0.1);"><div
```



```
A juicy piece of HTML code!

As a cybersecurity expert, I'll analyze the content and provide a phishing likelihood score based on various factors.

**Phishing Likelihood Score: 6/10**

Here's why:

1. **Unusual path data**: The SVG path data contains unusual characters and formatting, which could be an attempt to obfuscate or hide malicious code.
2. **Multiple unknown classes**: Classes like `cko__footer-text`, `payment-methods`, and others are not well-known or widely used, which might indicate a custom-made phishing template.
3. **Unusual div nesting**: The HTML structure is complex, with multiple nested div elements that could be hiding malicious content or JavaScript code.
4. **Unknown script tags**: The presence of unknown s                                                  eckout/js/wcko.0496191011192b0140d13.js"></script>`) coul
e the presence of malicious code.
5. **Lack of transparency**: The code does not provid                                                  og.

However, I must note that some of these factors might                                                  ould need to analyze the entire website, including the s
e code, network traffic, and user interactions.

**Recommendations:**

1. **Verify the website's authenticity**: Ensure you'                                                  familiar URLs, and legitimate branding.
2. **Be cautious with unknown scripts**: Avoid execut
3. **Use a reputable antivirus software**: Run regula                                                  ode.
```

The response is akin to a new analyst who is overly excited and under-experienced in this work

When I remove the superfluous text commentary to keep just the score and run it again, I get a different score. This LLM is too generalized and can't produce consistent results.

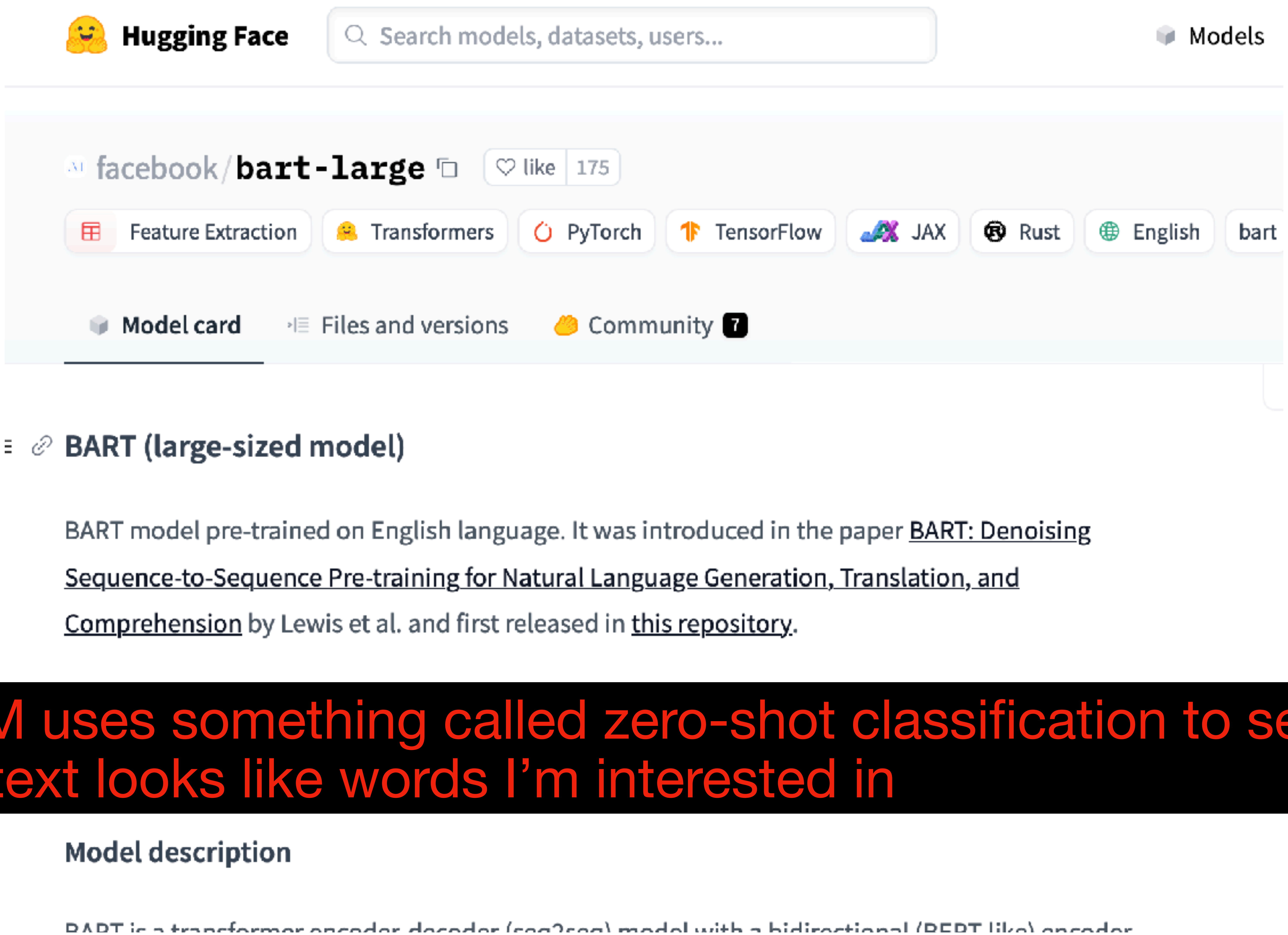
I need to try using a more specific LLM

I’m going to demo zero-shot classification and a tool called OmniParser

Text Classification with LLMs
2. Better modes

Zero-Shot Classification

OmniParser



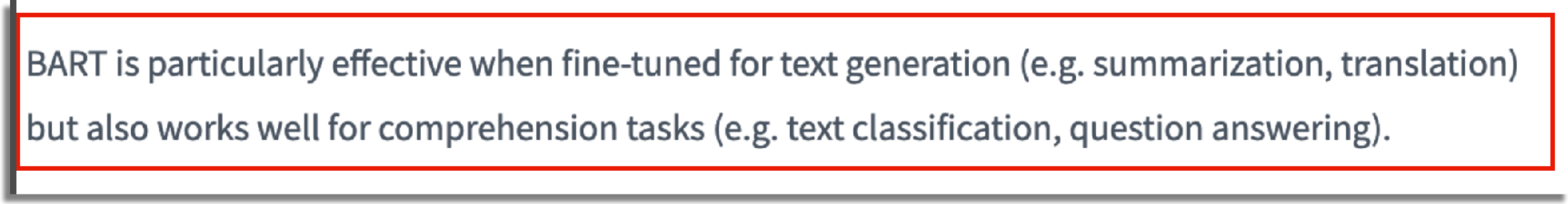
BART (large-sized model)

BART model pre-trained on English language. It was introduced in the paper BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension by Lewis et al. and first released in this repository.

This LLM uses something called zero-shot classification to see how much a blob of text looks like words I'm interested in

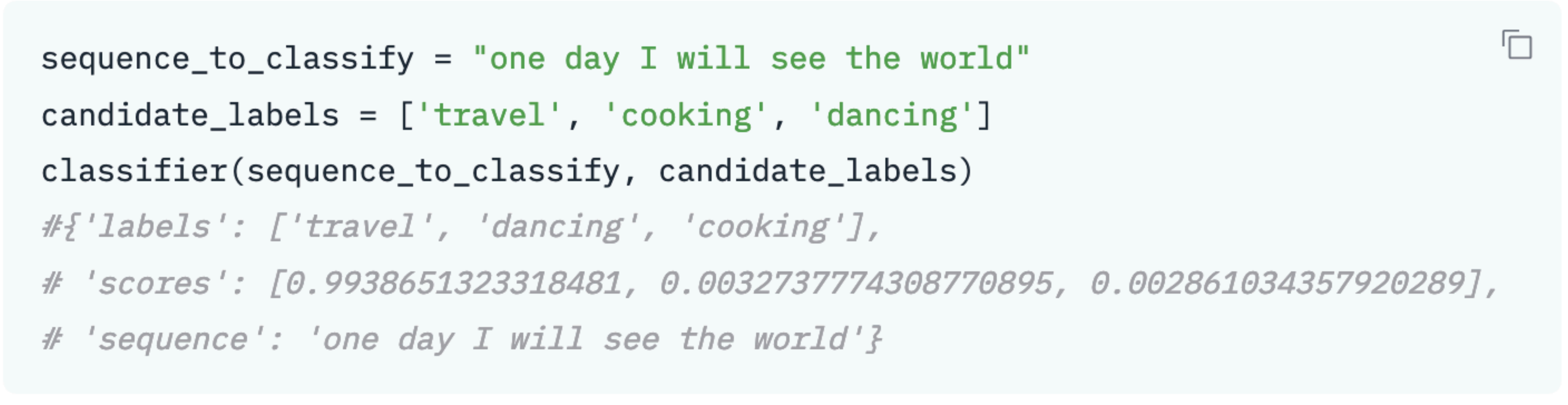
Model description

BART is a transformer encoder-decoder (seq2seq) model with a bidirectional (BERT-like) encoder



BART is particularly effective when fine-tuned for text generation (e.g. summarization, translation) but also works well for comprehension tasks (e.g. text classification, question answering).

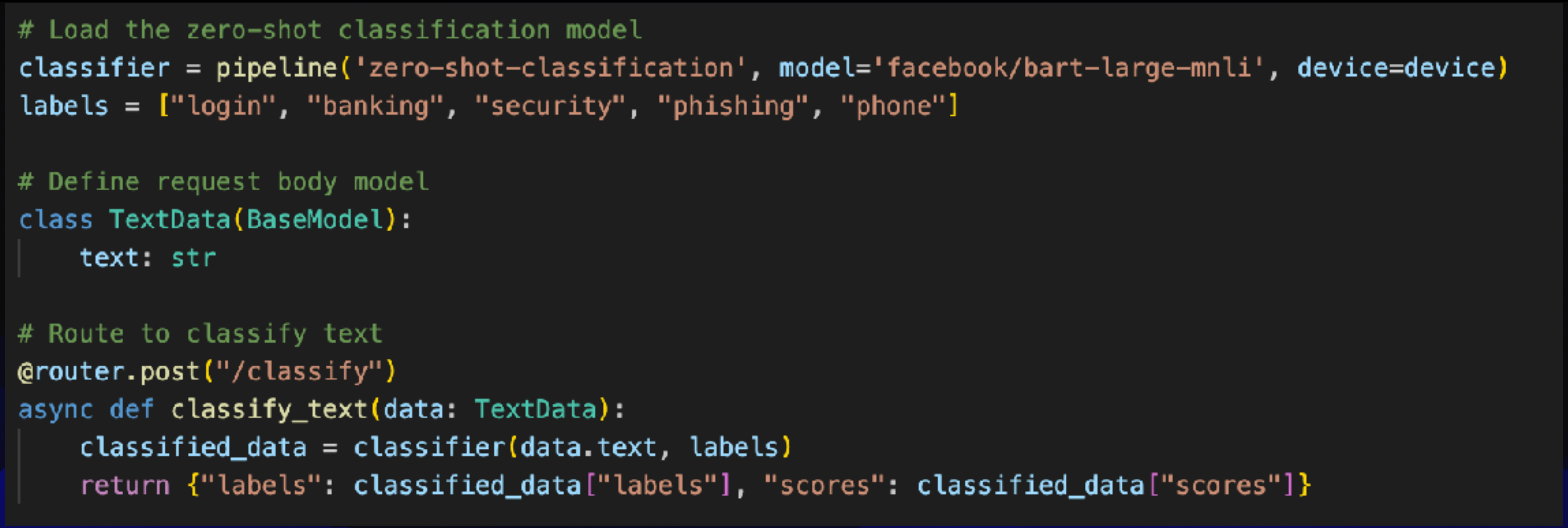
Zero-shot-classification



```
sequence_to_classify = "one day I will see the world"
candidate_labels = ['travel', 'cooking', 'dancing']
classifier(sequence_to_classify, candidate_labels)
#{'labels': ['travel', 'dancing', 'cooking'],
# 'scores': [0.9938651323318481, 0.0032737774308770895, 0.002861034357920289],
# 'sequence': 'one day I will see the world'}
```

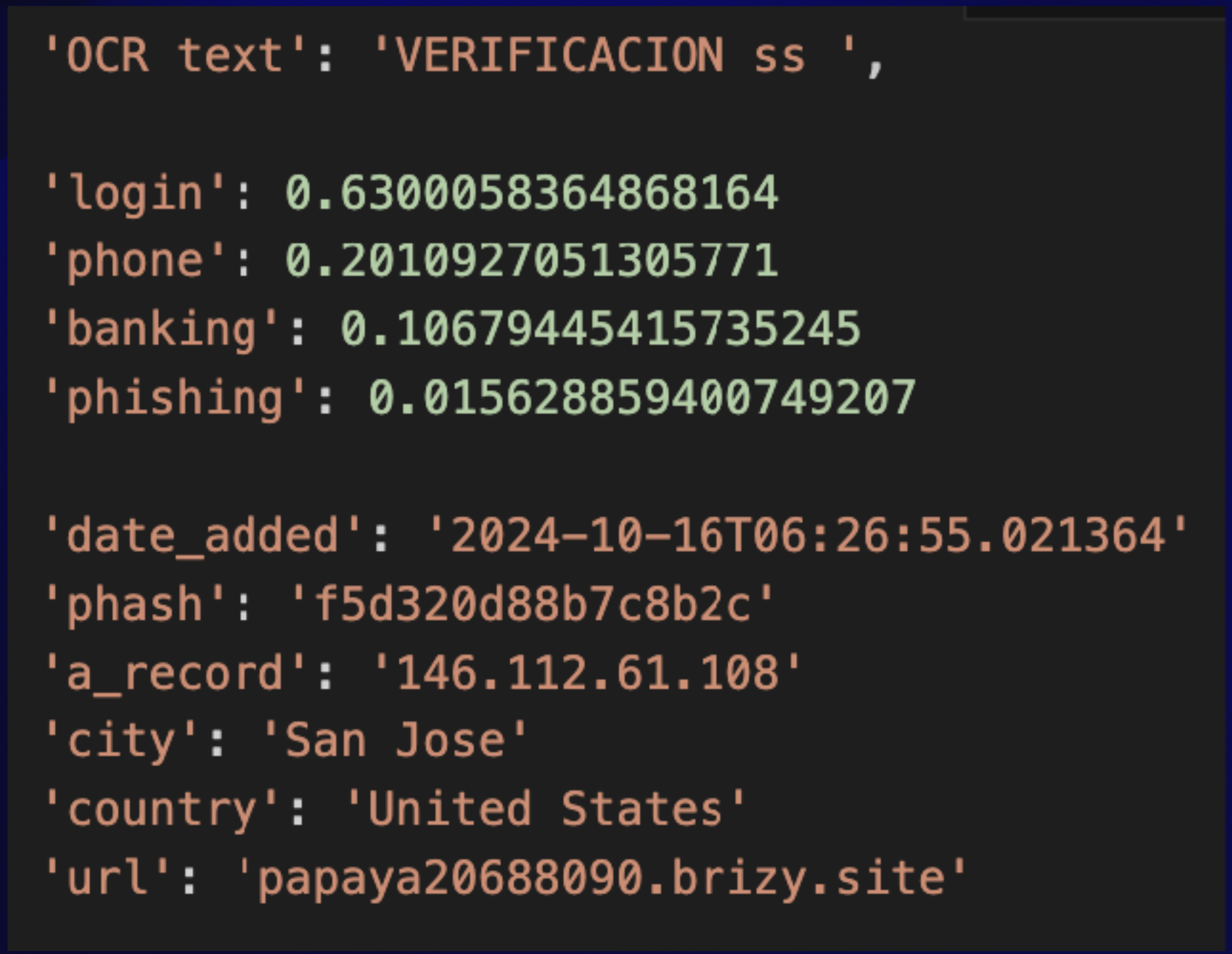
This demo, from hugging face shows three keywords (‘travel’, ‘cooking’, ‘dancing’). It receives a sentence (‘one day I will see the world’) and then it analyzes the text, resulting in a score for each keyword. The score represents how close the text is to each keyword.

For my test, I gave it 5 keywords. They aren't super specific, but just to use against any text I send to it. You can change this in the code (at end of presentation)



```
# Load the zero-shot classification model
classifier = pipeline('zero-shot-classification', model='facebook/bart-large-mnli', device=device)
labels = ["login", "banking", "security", "phishing", "phone"]

# Define request body model
class TextData(BaseModel):
    text: str

# Route to classify text
@router.post("/classify")
async def classify_text(data: TextData):
    classified_data = classifier(data.text, labels)
    return {"labels": classified_data["labels"], "scores": classified_data["scores"]}
```

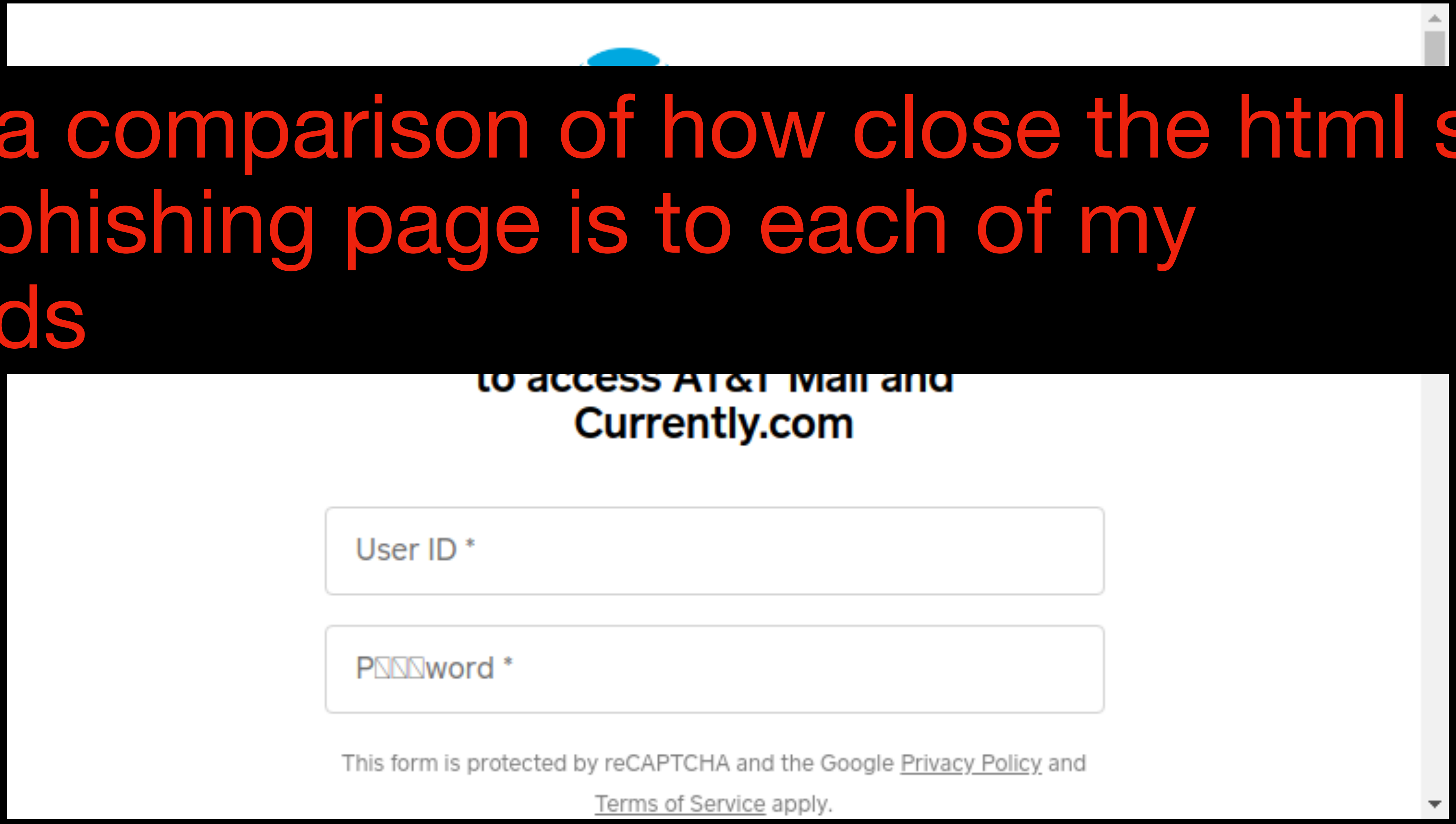
Zero-shot-classification results



```
'OCR text': 'VERIFICACION ss ',

'login': 0.6300058364868164
'phone': 0.2010927051305771
'banking': 0.10679445415735245
'phishing': 0.015628859400749207

'date_added': '2024-10-16T06:26:55.021364'
'phash': 'f5d320d88b7c8b2c'
'a_record': '146.112.61.108'
'city': 'San Jose'
'country': 'United States'
'url': 'papaya20688090.brizy.site'
```

009-288-49.weeblysite.com

Here, I’m seeing a comparison of how close the html source and the OCR text of this phishing page is to each of my specified keywords



to access AT&T Mail and
Currently.com

User ID *

P⧅⧅⧅word *

This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

OCR Classification:

login: 0.8000938296318054
security: 0.12101650983095169
phone: 0.05189700424671173
banking: 0.016330692917108536
phishing: 0.010662009939551353

HTML Source Classification:

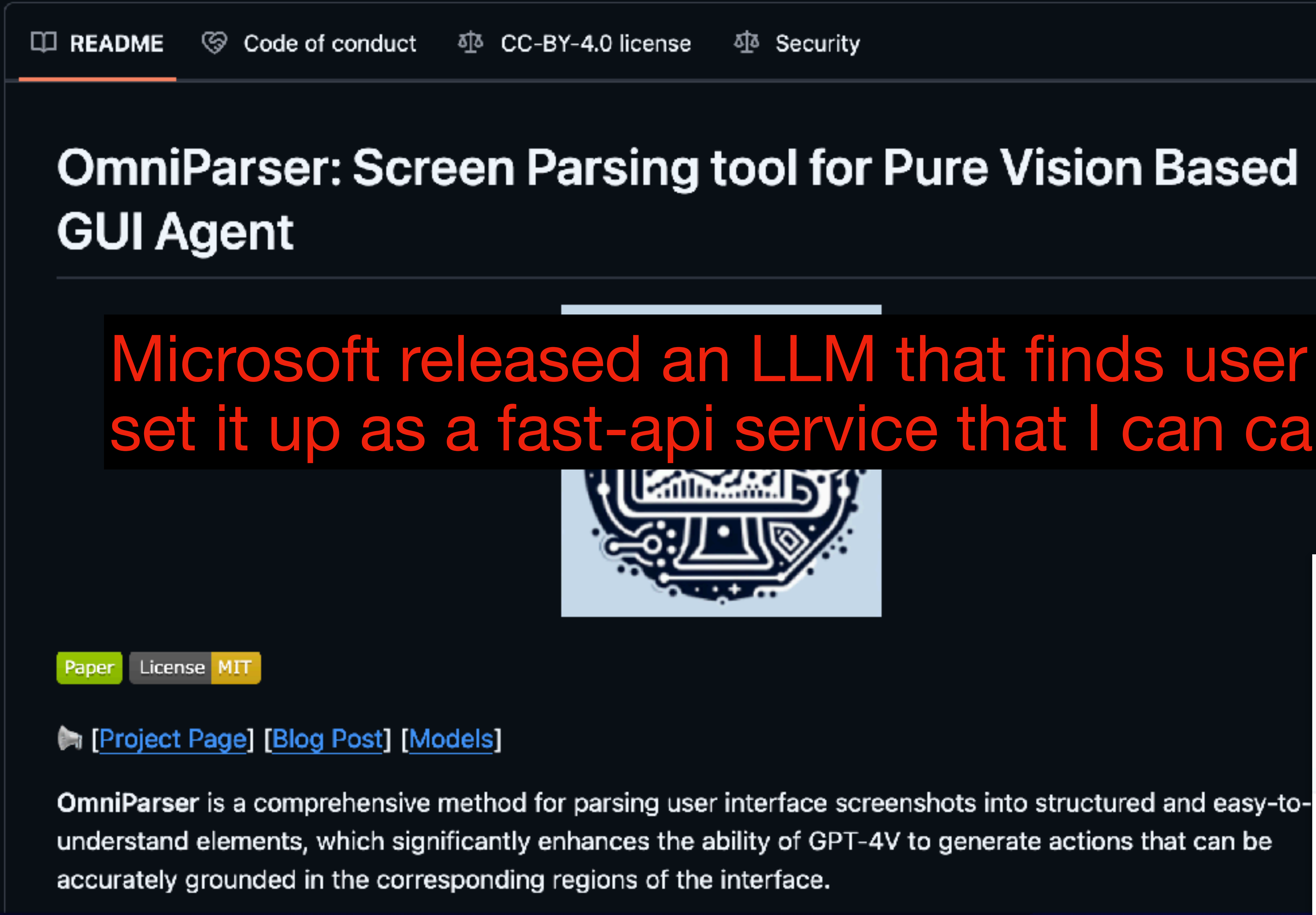
login: 0.8146952986717224
security: 0.09933193027973175
phone: 0.05374485993457794
banking: 0.017227377742528915
phishing: 0.015000510029494762

I want a better text scraping method. Thankfully, I found an LLM just for this.

Parsing the UI

And getting better OCR Results

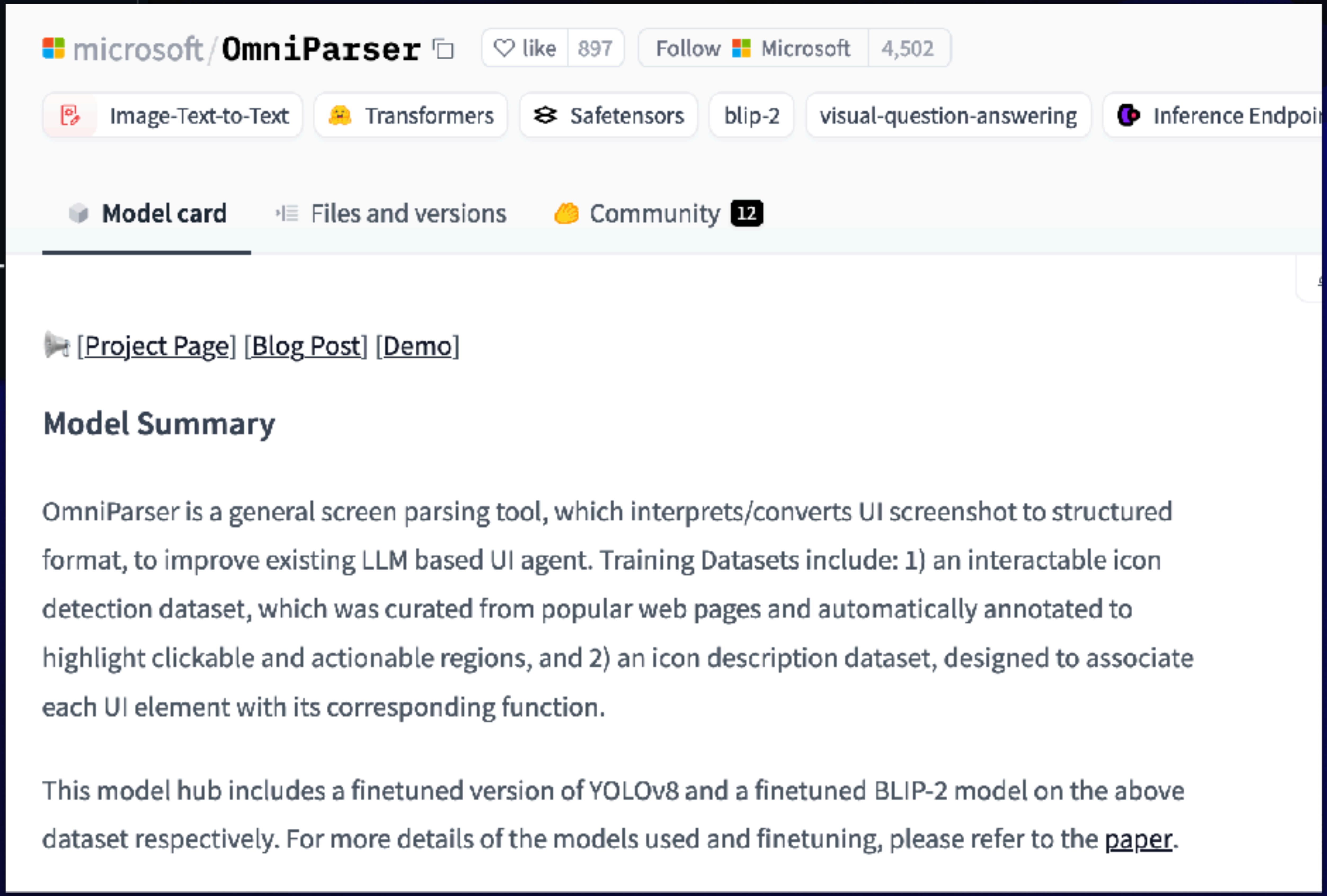
Microsoft released an LLM that finds user interface elements of a screenshot. I set it up as a fast-api service that I can call from my web app.



OmniParser: Screen Parsing tool for Pure Vision Based GUI Agent

[Project Page] [Blog Post] [Models]

OmniParser is a comprehensive method for parsing user interface screenshots into structured and easy-to-understand elements, which significantly enhances the ability of GPT-4V to generate actions that can be accurately grounded in the corresponding regions of the interface.



[Project Page] [Blog Post] [Demo]

Model Summary

OmniParser is a general screen parsing tool, which interprets/converts UI screenshot to structured format, to improve existing LLM based UI agent. Training Datasets include: 1) an interactable icon detection dataset, which was curated from popular web pages and automatically annotated to highlight clickable and actionable regions, and 2) an icon description dataset, designed to associate each UI element with its corresponding function.

This model hub includes a finetuned version of YOLOv8 and a finetuned BLIP-2 model on the above dataset respectively. For more details of the models used and finetuning, please refer to the paper.

This is OmniParsers demo of how this works. You upload an image (left) and get a result (right).

OmniParser for Pure Vision Based General GUI Agent 🔥

OmniParser is a screen parsing tool to convert general GUI screen to structured elements.

Upload image
AT&T
Sign in
Complete the process by sign-in to AT&T Mail Any form of incorrect email/passcode will be marked as spam and email will be deleted
Email *
Password
Image Output
0
1
2
3
4
5
6
Submit

Box Threshold 0.05 (0.01 – 1)

IOU Threshold 0.1 (0.01 – 1)

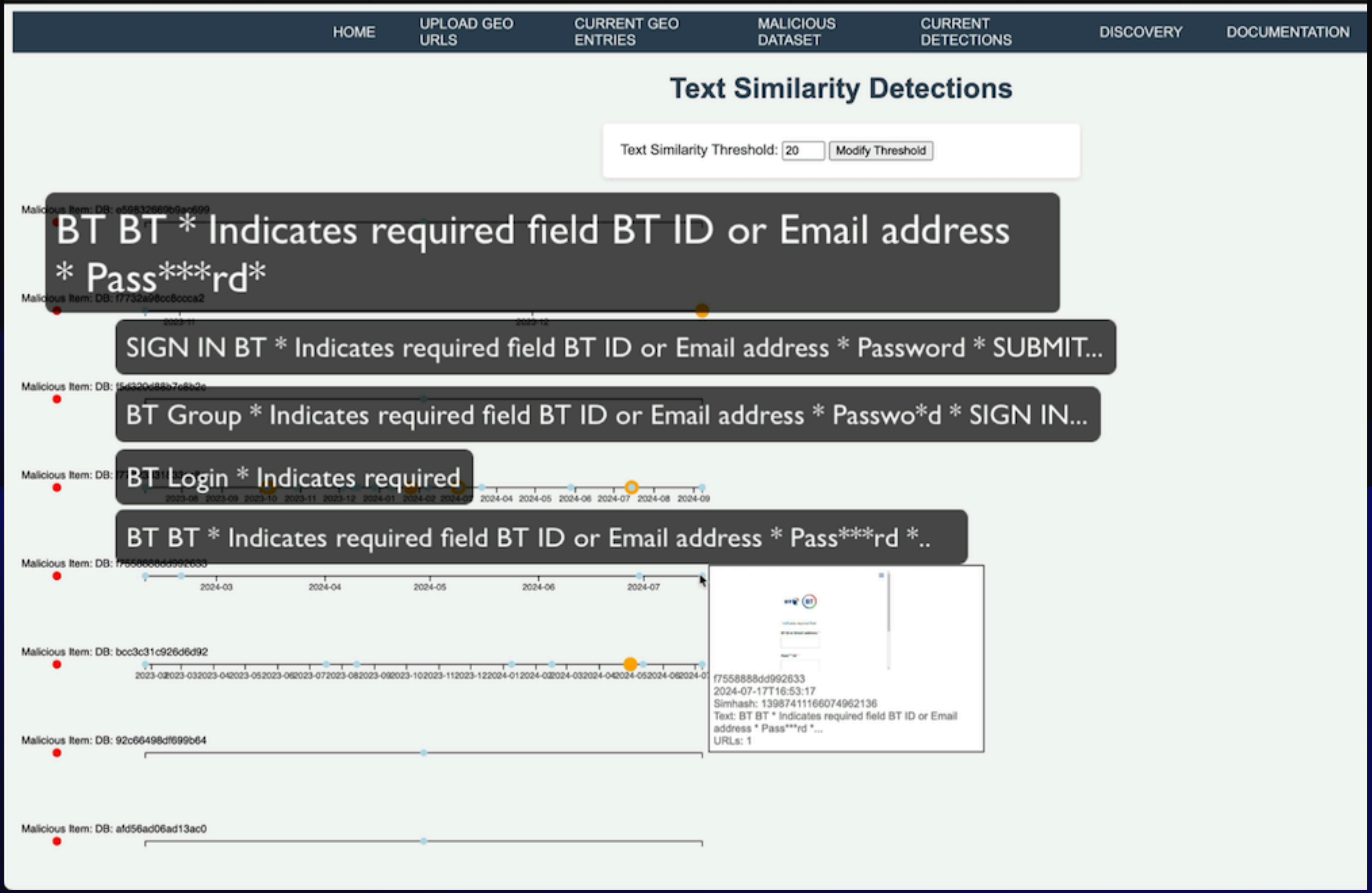
Parsed screen elements

Text Box ID 0: AT&T
Text Box ID 1: Sign in
Text Box ID 2: Complete the process by sign-in to AT&T Mail Any form
Text Box ID 3: of incorrect email/passcode will be marked as spam and
Text Box ID 4: emall will be deleted
Text Box ID 5: Emall
Text Box ID 6: Password

Let’s compare plain OCR vs OmniParser Classification.

URL	OCR text	OmniParse	Notable Differences
attmailaccount22.weebly.com	login (0.551) security (0.262) phone (0.115)	login (0.667) phone (0.226) security (0.070)	Higher login and phone scores in OmniParse
atdikg.taplink.ws	login (0.706) security (0.206) phone (0.058)	login (0.574) security (0.236) phone (0.162)	Lower login, higher security and phone in OmniParse
att-mail-109459.weeblysite.com	login (0.641) security (0.263) phone (0.062)	login (0.638) phone (0.171) security (0.163)	Similar login, higher phone and security in OmniParse

Text similarity from the OCR text



Text Similarity Detections

Text Similarity Threshold: 20 Modify Threshold

Malicious Item: DB: e59832669b9ac699

BT BT * Indicates required field BT ID or Email address * Pass***rd*

Malicious Item: DB: f7732a98cc8ccca2

SIGN IN BT * Indicates required field BT ID or Email address * Password * SUBMIT...

Malicious Item: DB: f5d320d88b7c8b2c

BT Group * Indicates required field BT ID or Email address * Passwo*d * SIGN IN...

Malicious Item: DB:

BT Login * Indicates required

BT BT * Indicates required field BT ID or Email address * Pass***rd *..

Malicious Item: DB: f7558888dd992633

f7558888dd992633
2024-07-17T16:53:17
Simhash: 13987411166074962136
Text: BT BT * Indicates required field BT ID or Email address * Pass***rd *...
URLs: 1

Malicious Item: DB: bcc3c31c926d6d92

Malicious Item: DB: 92c66498df699b64

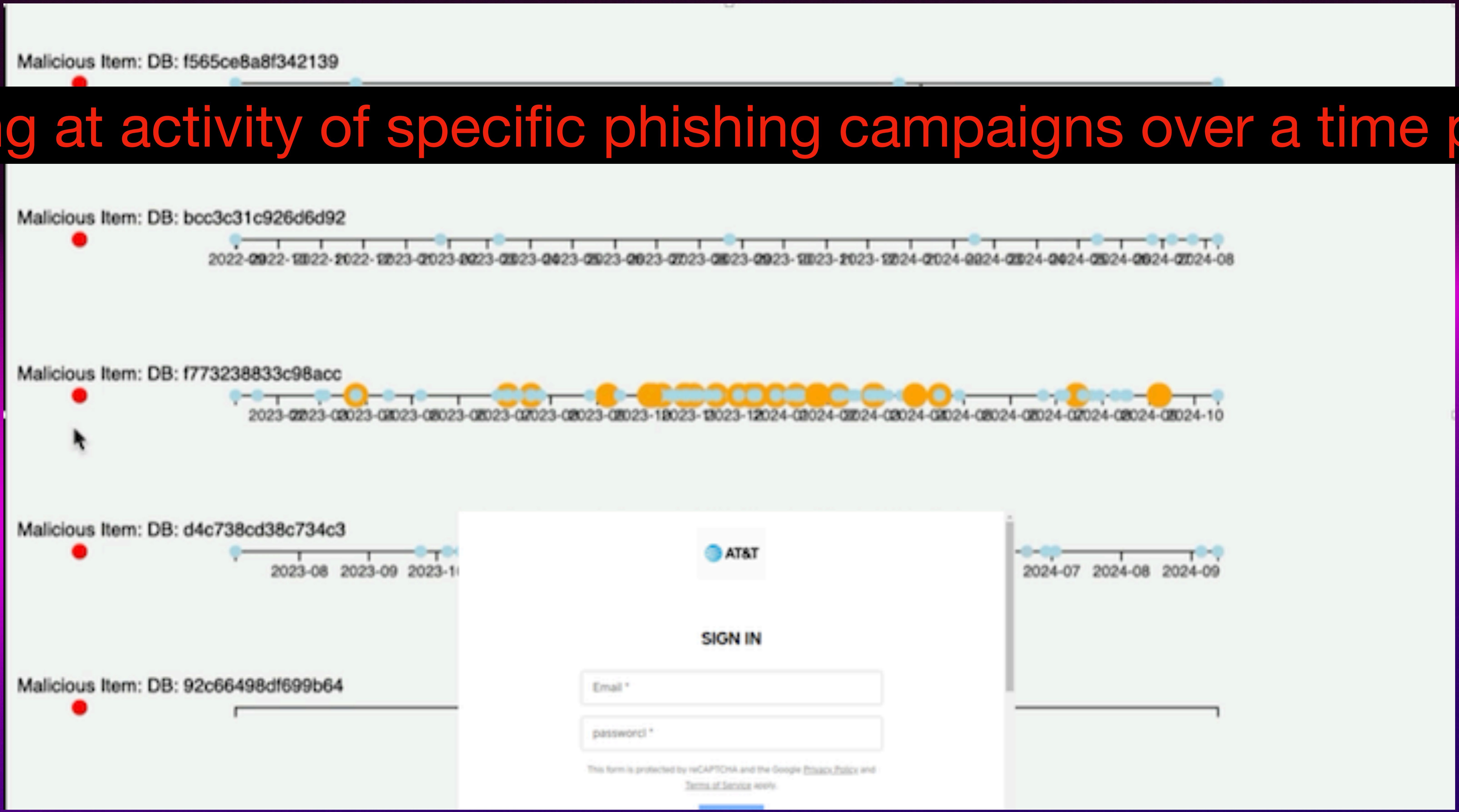
Malicious Item: DB: afd56ad06ad13ac0

Showing text similarity of scraped text on a timeline (timeline shows the screenshot, but is grouped from the OCR text)

Research

Campaigns over Time

I started building this as an inline detection tool, but it actually might have more use in researching phishing campaigns.

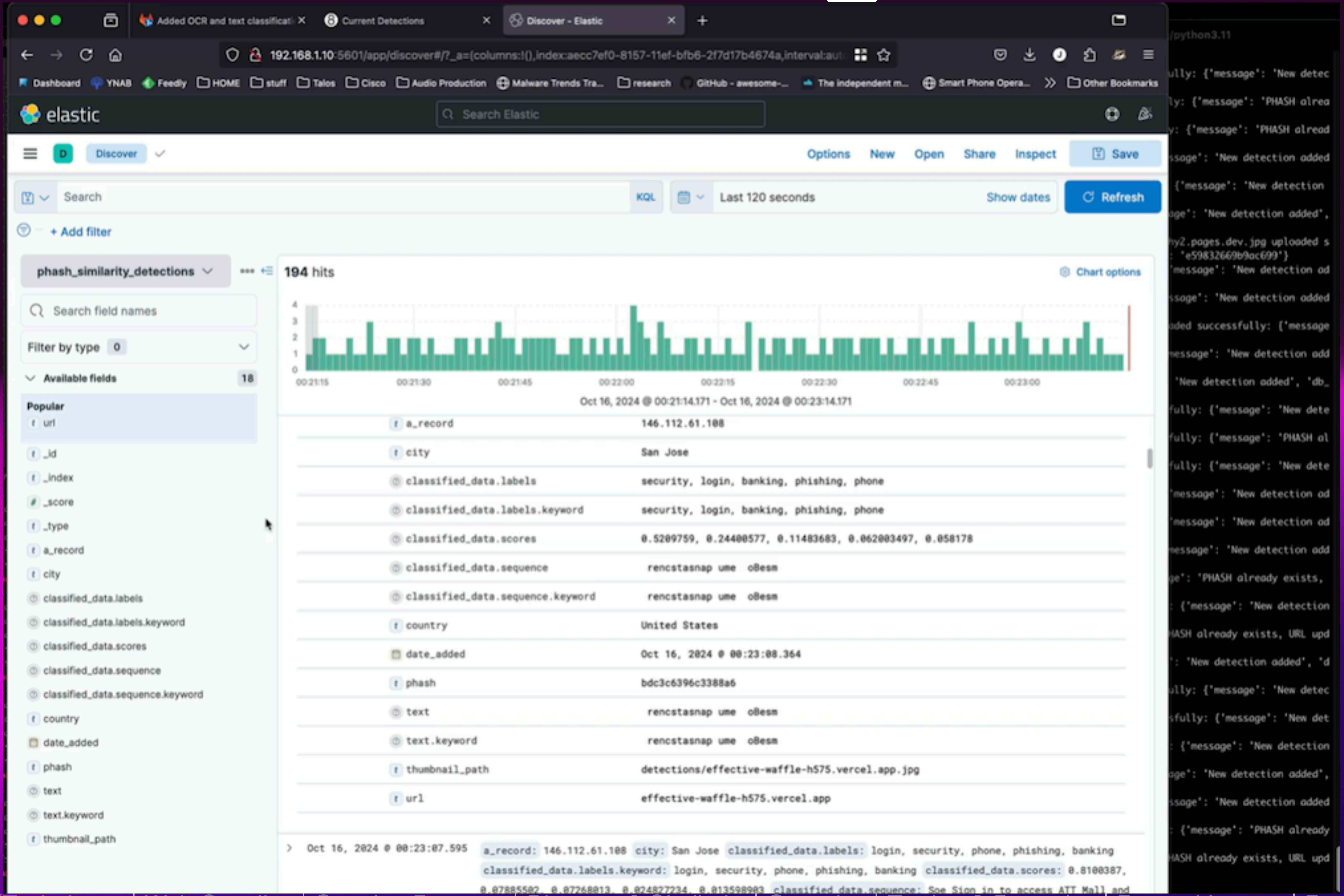


Looking at activity of specific phishing campaigns over a time period

Malicious Item: DB: f565ce8a8f342139
Malicious Item: DB: bcc3c31c926d6d92
Malicious Item: DB: f773238833c98acc
2024-10
Malicious Item: DB: d4c738cd38c734c3
2023-08
2023-09
2024-07
2024-08
2024-09
Malicious Item: DB: 92c66498df699b64
AT&T
SIGN IN
Email *
password *

Another portion of the interface and process is to send alerts to a SIEM or some system. I chose Elasticsearch and Kibana

Alerting



a_record	146.112.61.108
city	San Jose
classified_data.labels	security, login, banking, phishing, phone
classified_data.labels.keyword	security, login, banking, phishing, phone
classified_data.scores	0.5209759, 0.24400577, 0.11483683, 0.062003497, 0.058178
classified_data.sequence	rencstasnap ume oBesm
classified_data.sequence.keyword	rencstasnap ume oBesm
country	United States
date_added	Oct 16, 2024 @ 00:23:08.364
phash	bdc3c6396c3388a6
text	rencstasnap ume oBesm
text.keyword	rencstasnap ume oBesm
thumbnail_path	detections/effective-waffle-h575.vercel.app.jpg
url	effective-waffle-h575.vercel.app

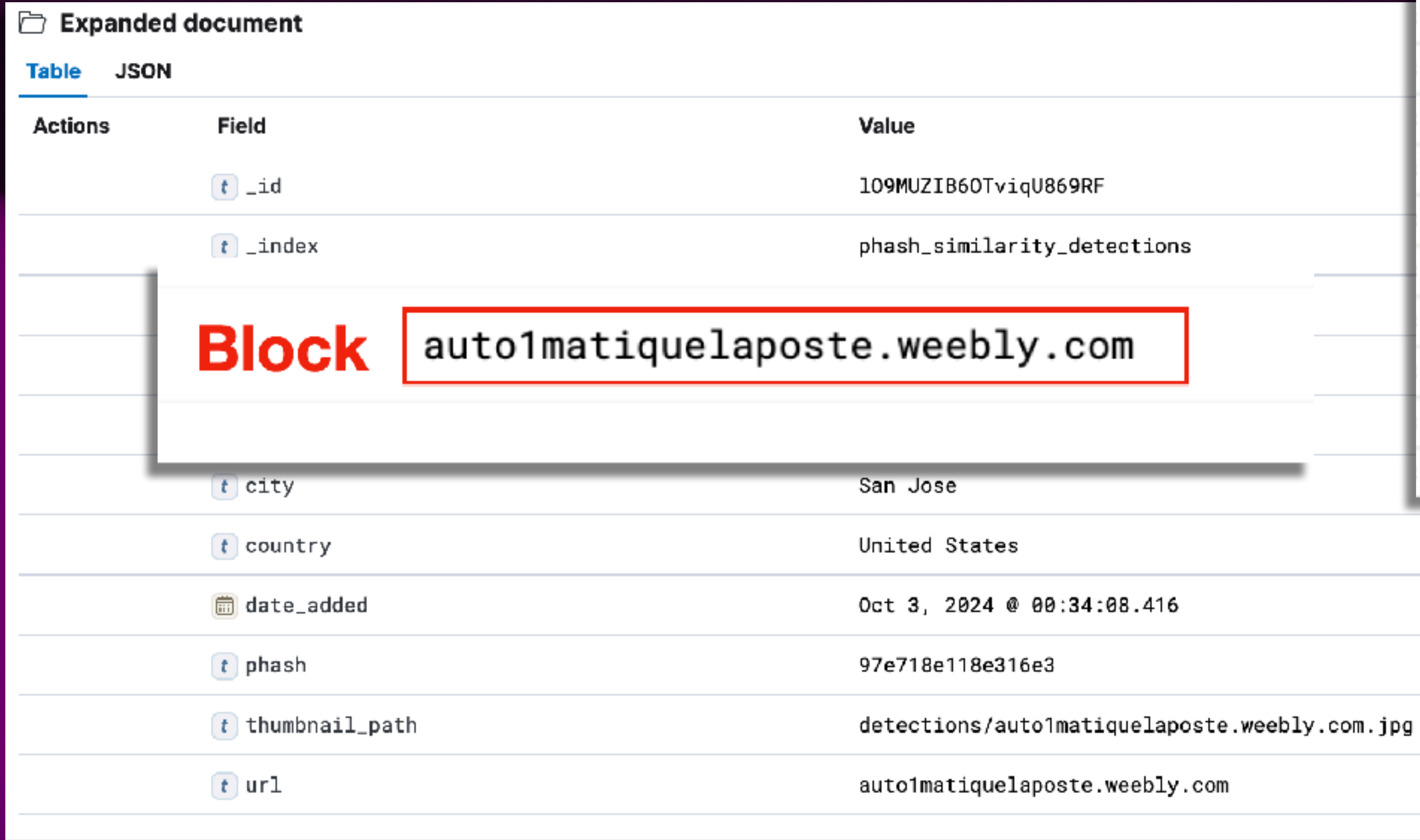
Oct 16, 2024 @ 00:23:07.595 a_record: 146.112.61.108 city: San Jose classified_data.labels: login, security, phone, phishing, banking classified_data.labels.keyword: login, security, phone, phishing, banking classified_data.scores: 0.8100387, 0.07885502, 0.07268013, 0.024827234, 0.013598903 classified_data.sequence: Soe Sign in to access ATT Mall and

```
ully: {'message': 'New detec
ly: {'message': 'PHASH alrea
y: {'message': 'PHASH alread
ssage': 'New detection added
 {'message': 'New detection
ge': 'New detection added',
hy2.pages.dev.jpg uploaded s
: 'e59832669b9ac699'}
'message': 'New detection ad
ssage': 'New detection added
aded successfully: {'message
essage': 'New detection add
 'New detection added', 'db_
fully: {'message': 'New dete
fully: {'message': 'PHASH al
fully: {'message': 'New dete
'message': 'New detection ad
'message': 'New detection ad
essage': 'New detection add
ge': 'PHASH already exists,
: {'message': 'New detection
HASH already exists, URL upd
': 'New detection added', 'd
ully: {'message': 'New detec
sfully: {'message': 'New det
: {'message': 'New detection
ge': 'New detection added',
ssage': 'New detection added
: {'message': 'PHASH already
HASH already exists, URL upd
```

Taking Action

If using for detection, you’re probably going to take some sort of action when a new phishing page is seen

And that action is typically to block the domain or URL



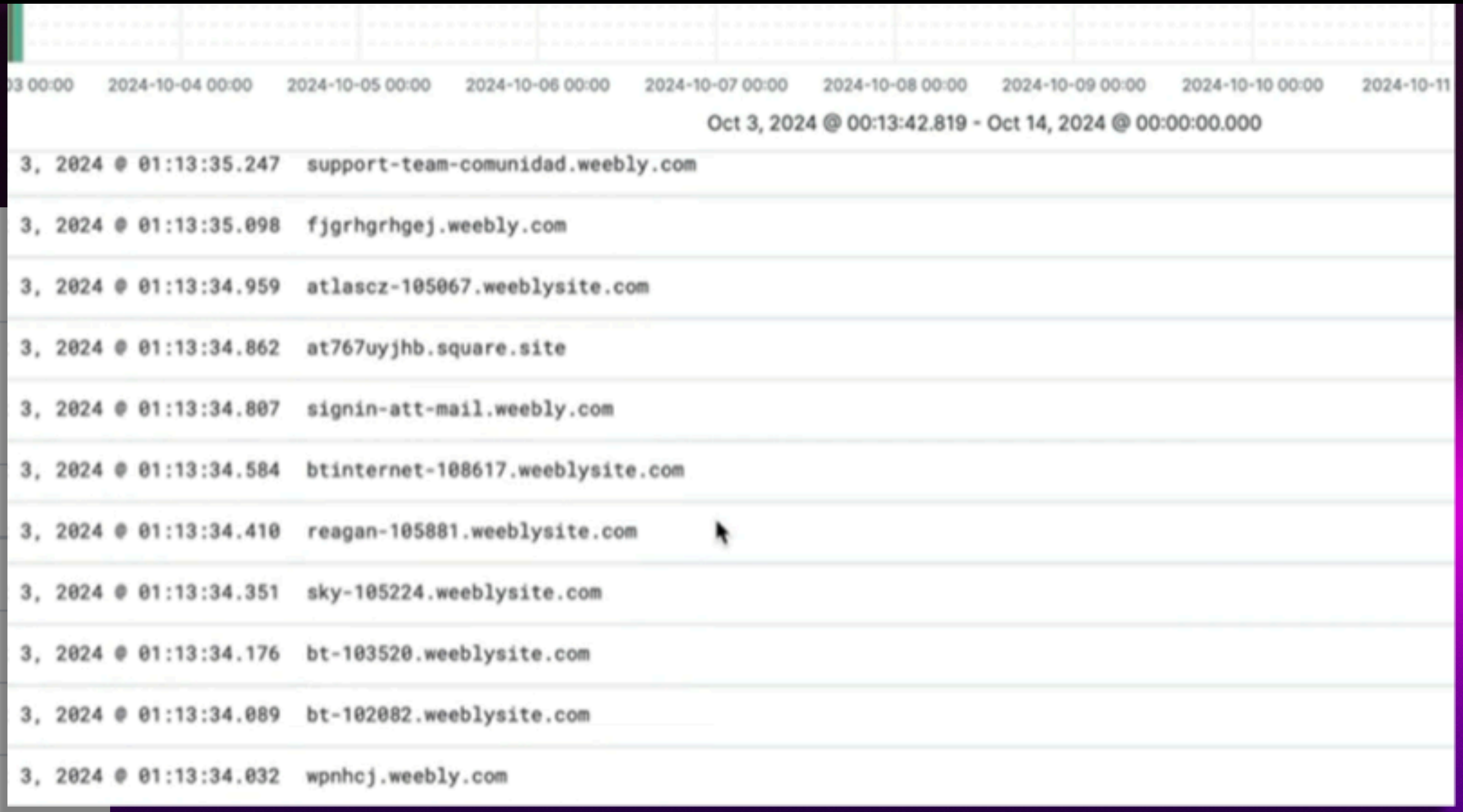
Expanded document

Table JSON

Actions	Field	Value
	_id	1O9MUZIB6OTviqU869RF
	_index	phash_similarity_detections
	city	San Jose
	country	United States
	date_added	Oct 3, 2024 @ 00:34:08.416
	phash	97e718e118e316e3
	thumbnail_path	detections/auto1matiquelaposte.weebly.com.jpg
	url	auto1matiquelaposte.weebly.com

Block auto1matiquelaposte.weebly.com

With alerts going to your SIEM, it’s easy to automate the blocking process



Oct 3, 2024 @ 00:13:42.819 - Oct 14, 2024 @ 00:00:00.000

```
3, 2024 @ 01:13:35.247   support-team-comunidad.weebly.com
3, 2024 @ 01:13:35.098   fjgrhgrhgej.weebly.com
3, 2024 @ 01:13:34.959   atlascz-105067.weeblysite.com
3, 2024 @ 01:13:34.862   at767uyjhb.square.site
3, 2024 @ 01:13:34.807   signin-att-mail.weebly.com
3, 2024 @ 01:13:34.584   btinternet-108617.weeblysite.com
3, 2024 @ 01:13:34.410   reagan-105881.weeblysite.com
3, 2024 @ 01:13:34.351   sky-105224.weeblysite.com
3, 2024 @ 01:13:34.176   bt-103520.weeblysite.com
3, 2024 @ 01:13:34.089   bt-102082.weeblysite.com
3, 2024 @ 01:13:34.032   wpnhcj.weebly.com
```

The next several slides will show various aspects of the interface at the time these slides were created (the code is still in development, so it might change in the future)

Current Version of Interface

Detections page

HOME
UPLOAD GEO URLS
CURRENT GEO ENTRIES
MALICIOUS DATASET
CURRENT DETECTIONS
DISCOVERY
DOCUMENTATION

Current Detections

Added	Image PHASH	URL	A Record	Location	Extracted Text	Classification	HTML Source Simhash
2024-07-09 22:13:34	b86fc59090c73b3a	rackspace1883999400488595.weebly.com	146.112.61.108	San Jose, United States	Webmail Login SUSPICIOUS Email address EMAIL Learn to quickly identify and report P@ssword A	**login:** 0.45 **phishing:** 0.37 **security:** 0.17 **banking:** 0.01 **phone:** 0.00	10337228898236707554
2024-01-11 00:42:47	f464239a99999b39	my-site-101155.weeblysite.com	146.112.61.108	San Jose, United States	A6B Q BJIE3TE, 3A AA IPObJD+KUTE A6B norpeбuren * Hapona This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.	**security:** 0.76 **login:** 0.10 **phone:** 0.08 **phishing:** 0.03 **banking:** 0.02	9931658834035383472
2024-07-02 21:04:28	e59832669b9ac699	inshaoiuyfldtcfyguijicdowiuhmkjnbhvgcfdxsedrftgyuhjikopoiu.pages.dev	146.112.61.108	San Jose, United States	Enter you mailbox password to proceed. Password Remember me Sign In	**security:** 0.64 **login:** 0.30 **phone:** 0.02 **banking:** 0.02 **phishing:** 0.02	14423626393392064642
2024-02-08 16:02:53	f777cc883164c998	worker-home-att-767e.rjolo7229.workers.dev	146.112.61.108	San Jose, United States	AT&T Sign In to access AT&T Mail and Currently.com User ID Forgot user ID? Password	**login:** 0.51 **security:** 0.33 **phone:** 0.13 **phishing:** 0.02 **banking:** 0.01	3930323897056612998
2024-07-05 12:10:01	b7f58c8c603333d8	wefvbjhwevbhiwcdghwghiwerhrfhghbvhcdvd.weebly.com	146.112.61.108	San Jose, United States	SNetZero" Welcome! Please sign-in below. * Indicates required field Email * Passvvord *	**login:** 0.82 **security:** 0.12 **phishing:** 0.02 **phone:** 0.02 **banking:** 0.02	9944788745945115358

Hovering over an OmniParser result screenshot

HOME
UPLOAD GEO URLS
CURRENT GEO ENTRIES
MALICIOUS DATASET
CURRENT DETECTIONS
DISCOVERY
DOCUMENTATION
Upload

Malicious Dataset

Added	Image PHASH	URL	A Record	Location	Extracted Text	Classified Data	HTML Source Simhash
2024-11-07 09:07:42	e59832669b9ac699	internationaldomaierepresе				**security:** 0.64 **login:** 0.30 **phone:** 0.02 **banking:** 0.02 **phishing:** 0.02	1442362639392064642
2024-11-07 09:07:43	f7732a98cc8ccca2	ionos-webmail-105551.wee				**security:** 0.58 **login:** 0.39 **phishing:** 0.02 **phone:** 0.01 **banking:** 0.01	3024539918263099024
2024-11-07 09:07:43	f5d320d88b7c8b2c	papaya20688090.brizy.site				**security:** 0.51 **login:** 0.39 **phone:** 0.05 **phishing:** 0.04 **banking:** 0.02	1949342313809748612
2024-11-07 09:07:44	f77323331833cc8c	mail-105085.weeblysite.com	74.115.51.55	Unknown City, United States	REAGAN Email Address * Passw*rd * This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply. Log In	**security:** 0.81 **login:** 0.16 **phishing:** 0.02 **phone:** 0.01 **banking:** 0.01	3160703433817627825
2024-11-07 09:07:45		mazulyta.pages.dev	146.112.61.108	San Jose, United States		**login:** 0.84 **security:** 0.11 **phishing:** 0.03	6441788642322681002

HOME
UPLOAD GEO URLS
CURRENT GEO ENTRIES
MALICIOUS DATASET
CURRENT DETECTIONS
DISCOVERY
DOCUMENTATION

Hovering over the html

Malicious Dataset

Upload ZIP File: Choose File No file chosen Upload

Added	Image PHASH	URL	A Record	Location	Extracted Text	Classified Data	HTML Source Simhash
2024-11-07 09:07:42	e59832669b9ac699	internationaldomaierepresentationmainserstexts4.pages.dev	146.112.61.108	San Jose, United States	Enter passw Passw m	security: 0.64	1442382639392064642
2024-11-07 09:07:43	f7732a98cc8ccca2	ionos-webmail-105551.weeblysite.com	74.115.51.54	Unknown City, United States	1&1 IO login E- Passw is p reCAPTCHA and the Google Privacy Policy and		5399182630990024
2024-11-07 09:07:43	f5d320d88b7c8b2c	papaya20688090.brizy.site	146.112.61.108	San Jose, United States	VERIFICACION Email: User name:. Password: Submit - Made With Brizye	**security:** 0.51 **login:** 0.39 **phone:** 0.05 **phishing:** 0.04 **banking:** 0.02	1949342313809748612
2024-11-07 09:07:44	f77323331833cc8c	mail-105085.weeblysite.com	74.115.51.55	Unknown City, United States	REAGAN Email Address * Passw*rd * This form is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply. Log In	**security:** 0.81 **login:** 0.16 **phishing:** 0.02 **phone:** 0.01 **banking:** 0.01	3160703433817627825
2024-11-07 09:07:45		[illegible]	146.112.61.108	San Jose, United States		login: 0.84	[illegible]

```
<html><head>

<title>Login - </title>
<meta name="viewport" content="width=device-width, initial-
scale=1.0">
<link
href="https://cdn.jsdelivr.net/npm/bootstrap@5.0.2/dist/css/bootstrap.n
rel="stylesheet" integrity="sha384-
EVSTQN3/azprG1Anm3QDgpJLIm9Nao0Yz1ztcQTwFspd3yD65Vohh
crossorigin="anonymous">
<link rel="stylesheet"
href="https://cdnjs.cloudflare.com/ajax/libs/font-
awesome/6.1.1/css/all.min.css" integrity="sha512-
KfkfwYDsLkIlwQp6LFnl8zNdLGxu9YAA1QvwINks4PhcElQSvqcyVLLD
crossorigin="anonymous" referrerpolicy="no-referrer">

<style>
@import url('https://fonts.googleapis.com/css2?
```

HOME
UPLOAD GEO URLS
CURRENT GEO ENTRIES
MALICIOUS DATASET
CURRENT DETECTIONS
DISCOVERY
DOCUMENTATION

PHASH Detections

PHASH Threshold: 8

Modify Threshold

Orange Circles are overlapping items
Click to expand

▨ 1-50 URLs

Malicious Item: DB: e59832669b9ac699

2024-03
2024-04
2024-05
2024-06
2024-07

Malicious Item: DB: f7732a98cc8ccca2

2023-06 2023-07 2023-08 2023-09 2023-10 2023-11 2023-12 2024-01 2024-02 2024-03 2024-04 2024-05 2024-06 2024-07 2024-08 2024-09

Malicious Item: DB: f5d320d88b7c8b2c

2023-08 2023-09 2023-10 2023-11 2023-12 2024-01 2024-02 2024-03 2024-04 2024-05 2024-06 2024-07

Malicious Item: DB: f77323331833cc8c

2023-04 2023-05 2023-06 2023-07 2023-08 2023-09 2023-10 2023-11 2023-12 2024-01 2024-02 2024-03 2024-04 2024-05 2024-06 2024-07 2024-08 2024-09

Malicious Item: DB: f5c70ae627d022d5

Malicious Item: DB: f7558888dd992633

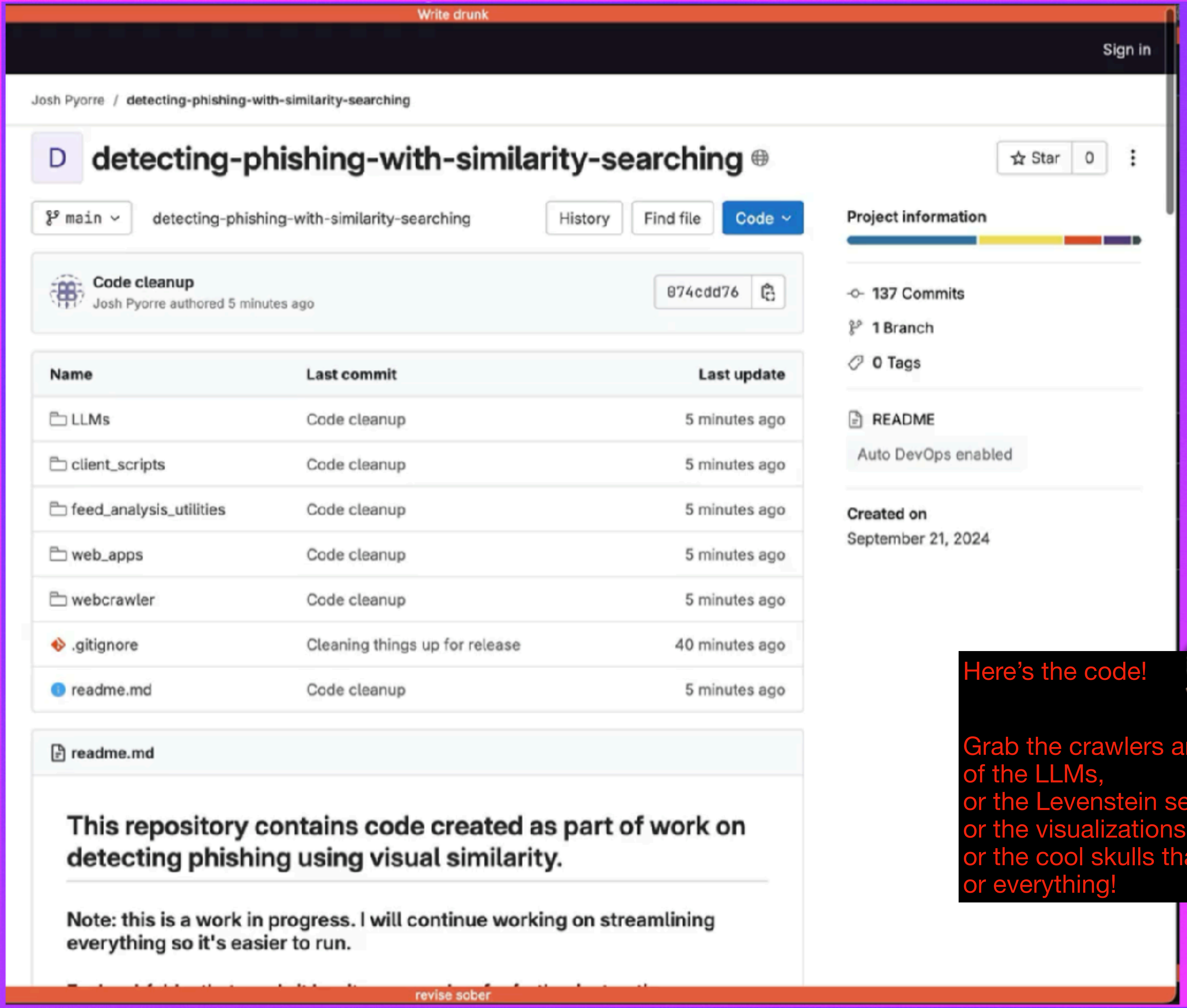
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09

Malicious Item: DB: bcc3c31c926d6d92

2023-02 2023-03 2023-04 2023-05 2023-06 2023-07 2023-08 2023-09 2023-10 2023-11 2023-12 2024-01 2024-02 2024-03 2024-04 2024-05 2024-06 2024-07 2024-08

Timeline based off screenshot similarities

https://gitlab.pyosec.com



Josh Pyorre / detecting-phishing-with-similarity-searching

detecting-phishing-with-similarity-searching

Code cleanup
Josh Pyorre authored 5 minutes ago
874cdd76

Name	Last commit	Last update
LLMs	Code cleanup	5 minutes ago
client_scripts	Code cleanup	5 minutes ago
feed_analysis_utilities	Code cleanup	5 minutes ago
web_apps	Code cleanup	5 minutes ago
webcrawler	Code cleanup	5 minutes ago
.gitignore	Cleaning things up for release	40 minutes ago
readme.md	Code cleanup	5 minutes ago

Project information

- 137 Commits
- 1 Branch
- 0 Tags

README
Auto DevOps enabled

Created on
September 21, 2024

readme.md

This repository contains code created as part of work on detecting phishing using visual similarity.

Note: this is a work in progress. I will continue working on streamlining everything so it's easier to run.

Here's the code!

Grab the crawlers and crawl things,
of the LLMs,
or the Levenstein searching,
or the visualizations,
or the cool skulls that are used in the maps,
or everything!

Detecting Phishing using Visual Similarity

https://gitlab.pyosec.com/

Thank you!

Find me:

- pyosec.com
- bsky.app/profile/joshpyorre.bsky.social
- infosec.exchange/@joshpyorre

Music (dark electronic): dievortex.com DIE VORTEX