HOMOGLYPH DETECTION TOOL

Exploration and Proof of concept (POC)

Prathamesh Arun Kamble

04.08.2025
Intern Id-195

🔍 Tool : Homoglyph detection tool

📌 Objective:

Detect lookalike (homoglyph) domain names that use Unicode characters resembling Latin characters (e.g., gοogle.com with Cyrillic "ο") to identify phishing, spoofing, and impersonation attempts.

🧠 Core Concepts Behind This Tool

1. What Are Homoglyphs?

Homoglyphs are visually similar characters from different scripts.

Example:

Latin	Cyrillic	Lookalike?
o	о (U+043E)	Yes
e	е (U+0435)	Yes
a	а (U+0430)	Yes
B	Ь (U+042C)	Yes

Attackers register fake domains using homoglyphs:

google.com → gοogle.com (Cyrillic "ο")

2. Why It's a Threat?

Users can't visually distinguish such domains.

Used in phishing, malware delivery, identity spoofing.

Difficult to detect using basic string comparisons.

3. Unicode Normalization

Python’s unicodedata.normalize() can transform visually similar characters to their canonical equivalents using:

NFKC (Normalization Form Compatibility Composition)

So gоogle.com and google.com become identical after normalization.

4. Tool Workflow (Functionality)

➤ Input:

Text file with list of domain names/usernames.

➤ Process:

1. Normalize each string using Unicode NFKC.

2. Track normalized strings in a dictionary.

3. Compare current domain’s normalized form:

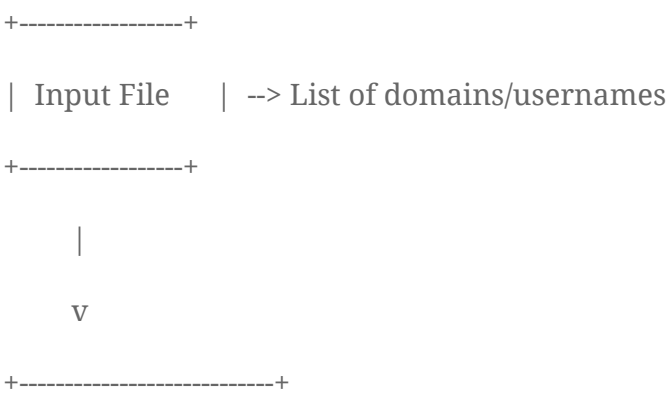
If already seen → homoglyph detected.

If new → store and continue.

➤ Output:

Warnings for each visually similar (homoglyph) domain found.

🛠️ Tool Architecture Overview



| Unicode Normalization |

| (unicodedata.normalize) |

+-------------------------+

|

v

+----------------------+

| Compare Normalized |

| Entries in Dictionary |

+----------------------+

|

v

+-----------------+

| Show Detections | --> CLI Output

+-----------------

Examples for output

1.google.com

2.ɡoogle.com

💻 Sample Output

1.Enter a domain name to check : google.com

Original: google.com

Normalized: google.com

✅ Safe. No suspicious characters or homoglyph detected.

2.Enter a domain name to check: ɡoogle.com

Original: ɡoogle.com

Normalized: google.com

⚠️ Suspicious! Look like : google.com

🛡️ MITRE ATT&CK Mapping

Field	Value
Tactic	Reconnaissance (TA0043)
Technique	Acquire Infrastructure → Domain Registration (T1583.001)
Detection	Lookalike Domain Use
Tool Purpose	Identify phishing infrastructure before attack execution

📂 Use Cases

Use Case	Description
Red Team Recon	Check registered domains to avoid overlap
Blue Team Threat Detection	Scan suspicious domain lists
OSINT Investigations	Verify social media handle lookalikes
Brand Protection	Find spoofed versions of business domains

🔐 Limitations

Doesn’t check for real-time DNS resolution.

False positives possible if domains are legitimate multilingual brands.

Only compares visually — no DNS or WHOIS context.

🚀 Expansion Ideas

Add web interface (Flask).

Add DNS record lookup.

Auto-check against known TLD phishing lists (e.g., openphish, abuse.ch).

Integration with whois, dig, VirusTotal.

✅ Summary

Feature	Description
Goal	Detect homoglyph lookalike domains
Language	Python
Runs on	Termux, Linux, Windows
Input	whitelist
Output	CLI alert of potential homoglyph matches
MITRE Technique	T1583.001 – Lookalike Domains