

# **PROOF OF CONCEPT (PoC) REPORT**

**Tool Name:** Homoglyph Detection Tool

**Author:** Sristi Dutta

**Intern Id:** 387

---

## **Executive Summary**

The **Homoglyph Detection Tool** is a Python-based command-line utility designed to identify and flag suspicious domain names containing homoglyph characters. Homoglyphs are visually similar characters from different scripts (Latin, Cyrillic, Greek, etc.) that attackers use in phishing campaigns to mimic trusted domains.

This PoC demonstrates how the tool detects such domains, explains the differences, and assists cybersecurity analysts in mitigating phishing risks.

---

## **Objective**

The goal of this PoC is to validate the functionality of the Homoglyph Detection Tool by:

- Detecting homoglyph-based phishing domains.
  - Comparing suspicious domains against a whitelist of legitimate domains.
  - Highlighting the exact character differences.
  - Demonstrating a realistic use case scenario in a cybersecurity context.
- 

## **Scope**

- **In Scope:**
    - Detection of homoglyphs in domain names.
    - Comparison against known safe domains (`safe_domains.txt`).
    - Identification of suspicious characters and their Unicode details.
    - Highlighting and explaining exact character substitutions.
  - **Out of Scope:**
    - URL reputation analysis.
    - Automated blocking of detected URLs.
    - Integration with external threat intelligence platforms.
    - Detection of non-homoglyph phishing techniques such as typosquatting without character replacements.
-

## Tool Overview

The tool processes user-provided URLs/domains, normalizes them using **Unicode NFKC normalization**, and compares them against a whitelist using **string similarity matching**. Suspicious domains are flagged, with replaced characters highlighted and explained.

### Key Features:

- Unicode homoglyph detection.
- Visual highlighting of suspicious characters.
- Detailed reasoning for each flagged character.
- Lightweight, CLI-based, and easy to integrate.

---

## Requirements

### Software Requirements:

- Python latest version
- Built-in modules: `unicodedata`, `difflib`

### Data Requirements:

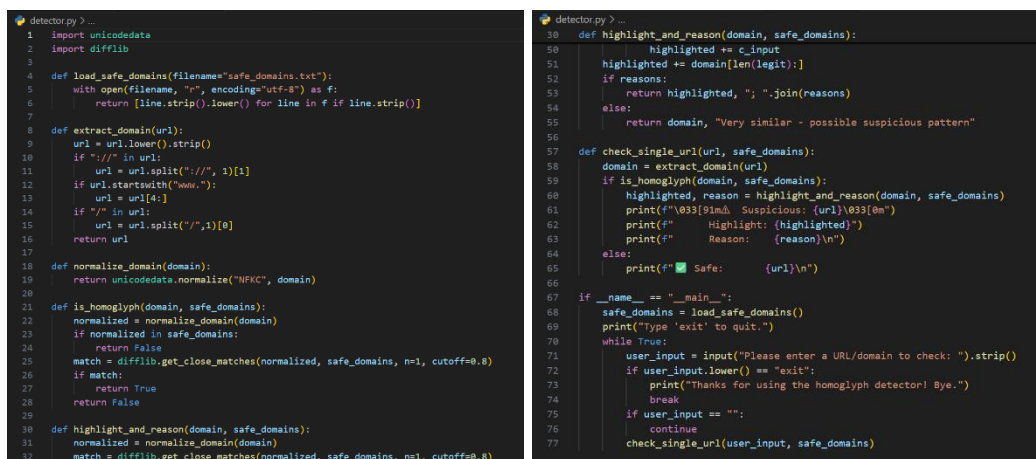
- `safe_domains.txt` — whitelist of legitimate domains.
- `sample_input.txt` — test URLs for batch processing.
- `detector.py` — python script for running the tool.

---

## Steps to Run the Tool

### 1. Install Python Latest Version

Ensure Python is installed and accessible via the terminal or command prompt, and prepare the `detector.py` script with all required commands and formatting so the tool runs correctly.



```
1 import unicodedata
2 import difflib
3
4 def load_safe_domains(filename="safe_domains.txt"):
5     with open(filename, "r", encoding="utf-8") as f:
6         return [line.strip().lower() for line in f if line.strip()]
7
8 def extract_domain(url):
9     url = url.lower().strip()
10    if "://" in url:
11        url = url.split("://", 1)[1]
12    if url.startswith("www."):
13        url = url[4:]
14    if "/" in url:
15        url = url.split("/", 1)[0]
16    return url
17
18 def normalize_domain(domain):
19     return unicodedata.normalize("NFKC", domain)
20
21 def is_homoglyph(domain, safe_domains):
22     normalized = normalize_domain(domain)
23     if normalized in safe_domains:
24         return False
25     match = difflib.get_close_matches(normalized, safe_domains, n=1, cutoff=0.8)
26     if match:
27         return True
28     return False
29
30 def highlight_and_reason(domain, safe_domains):
31     normalized = normalize_domain(domain)
32     match = difflib.get_close_matches(normalized, safe_domains, n=1, cutoff=0.8)
33
34 def highlight_and_reason(domain, safe_domains):
35     highlighted = ""
36     highlighted += c_input
37     highlighted += domain[len(legit):]
38     if reasons:
39         return highlighted, "; ".join(reasons)
40     else:
41         return domain, "Very similar - possible suspicious pattern"
42
43 def check_single_url(url, safe_domains):
44     domain = extract_domain(url)
45     if is_homoglyph(domain, safe_domains):
46         highlighted, reason = highlight_and_reason(domain, safe_domains)
47         print(f"\033[91mA Suspicious: {url}\033[0m")
48         print(f"    Highlight: {highlighted}")
49         print(f"    Reason: {reason}\n")
50     else:
51         print(f"\033[92mSafe: {url}\n")
52
53 if __name__ == "__main__":
54     safe_domains = load_safe_domains()
55     print("Type 'exit' to quit.")
56     while True:
57         user_input = input("Please enter a URL/domain to check: ").strip()
58         if user_input.lower() == "exit":
59             print("Thanks for using the homoglyph detector! Bye.")
60             break
61         if user_input == "":
62             continue
63         check_single_url(user_input, safe_domains)
```

## 2. Prepare the Safe Domains List

Create a file named `safe_domains.txt` containing legitimate domains, one per line:

```
safe_domains.txt
1 google.com
2 github.com
3 microsoft.com
4 apple.com
5 facebook.com
6 twitter.com
7 instagram.com
```

## 3. Prepare Test URLs

Create `sample_input.txt` with URLs to check:

```
sample_input.txt
1 https://google.com
2 https://www.github.com
3 https://www.microsoft.com
4
5 https://google.com
6 https://www.microsoft.com
7 http://apple.com
8 https://facebook.com
9 https://twitter.com
10 https://www.instagram.com
```

## 4. Run Interactive Mode

```
PS C:\Users\USER\OneDrive\Desktop\Homoglyph> Python detector.py
```

- Enter URLs one by one.
- Type exit to quit.

---

## PoC Test Execution

```
PS C:\Users\USER\OneDrive\Desktop\Homoglyph> Python detector.py
Type 'exit' to quit.
Please enter a URL/domain to check: https://google.com
⚠ Suspicious: https://google.com
Highlight: g[o][o]gle.com
Reason: position 2: 'o' is CYRILLIC SMALL LETTER O, should be 'o' (LATIN SMALL LETTER O); position 3: 'o' is CYRIL
LIC SMALL LETTER O, should be 'o' (LATIN SMALL LETTER O)

Please enter a URL/domain to check: https://twitter.com
⚠ Suspicious: https://twitter.com
Highlight: twitt[e]r.com
Reason: position 6: 'e' is CYRILLIC SMALL LETTER IE, should be 'e' (LATIN SMALL LETTER E)

Please enter a URL/domain to check: https://facebook.com
✅ Safe: https://facebook.com

Please enter a URL/domain to check: https://faceb00k.com
⚠ Suspicious: https://faceb00k.com
Highlight: faceb[0][0]k.com
Reason: position 6: '0' is DIGIT ZERO, should be 'o' (LATIN SMALL LETTER O); position 7: '0' is DIGIT ZERO, should
be 'o' (LATIN SMALL LETTER O)
```

```
Please enter a URL/domain to check: https://google.com
✓ Safe:      https://google.com

Please enter a URL/domain to check: https://github.com
✓ Safe:      https://github.com

Please enter a URL/domain to check: https://microsoft.com
✓ Safe:      https://microsoft.com

Please enter a URL/domain to check: https://twitter.com
✓ Safe:      https://twitter.com

Please enter a URL/domain to check: exit
Thanks for using the homoglyph detector! Bye.
PS C:\Users\USER\OneDrive\Desktop\Homoglyph> |
```

### Use Case Scenario

A SOC analyst is reviewing email phishing logs and identifies multiple suspicious domains. Instead of manually checking each one, the analyst runs them through the Homoglyph Detection Tool. Within seconds, the tool highlights which domains are malicious lookalikes and specifies the exact deceptive characters used.

---

### Advantages

- **Accurate Detection:** Finds even subtle character replacements.
- **Lightweight:** Minimal dependencies, easy to run anywhere.
- **Customizable:** Safe domains list can be expanded for different organizations.
- **Informative Output:** Detailed reasons make investigation faster.

---

### Threat Impact Analysis

#### **Threat Addressed:**

- Homoglyph attacks in phishing campaigns.
- Credential theft via visually deceptive domains.
- Malicious redirects to attacker-controlled websites.

#### **Potential Risks Without This Tool:**

- Users unknowingly visiting malicious websites.
- Compromise of login credentials.
- Malware delivery via fake login portals.

- Loss of trust in brand/domain.

### **Impact Reduction:**

This tool enables **early detection**, allowing security teams to block suspicious domains before phishing emails or malicious ads reach end-users.

---

### **Future Enhancements**

- **Integration with Threat Intelligence Feeds** – automatically check detected domains against blacklists.
  - **Real-Time Browser Extension** – warn users before visiting a suspicious link.
  - **Email Gateway Integration** – scan incoming emails for homoglyph-based URLs.
  - **Machine Learning Models** – improve detection accuracy and adapt to new homoglyph attack techniques.
  - **Automated Blocking** – link with firewall or proxy rules for immediate protection.
- 

### **Conclusion**

This PoC confirms the Homoglyph Detection Tool's effectiveness in identifying visually deceptive domains. By combining **Unicode normalization** with **similarity analysis**, it offers a lightweight yet powerful defense against phishing. With further integration into real-time monitoring systems, it can significantly reduce phishing success rates.