Script to remove homoglyphs and zero-width characters to allow for safe distribution of documents from anonymous sources.

forensic-analysis

README.md

SafeText

Tool to sanitize text to allow for safe distribution of documents from anonymous sources by removing zero-width characters and homoglpyhs.

Individuals attempting to leak an email or other text file face the risk of identification through fingerprinting. Fingerprinting often occurs when the original distributor of the document has embedded some form of a canary. For example, Elon Musk's email in 2008 in response to leaks featured slightly different wording for each employee. This tactic was realized by the employees, and failed. An easier tactic that is also employed, is the presence of nearly invisible changes to the text. SafeText is designed to identify and remove these changes. Specifically this tool will remove homoglyphs, zero-width characters, and other subtle characters. This tool will also attempt to identify unique spelling of words that could give away an individual's location.

Usage

To use SafeText, call:

```
python safetext.py inputfile
```

Example output is:

```
λ python safetext.py TestFile.txt
[*] Cleaning TestFile.txt to TestFile.txt.safe ...
[!] FOUND HOMOGLYPHIC CHARACTER CYRILLIC_large_H ON LINE 1
The message said: "(H)ey, let's hang out!"
[!] FOUND a SPACE ON LINE # 2
Lorem*Ipsum*Dolor*Sit
[!] WARNING - Use of spelling (colour) that identifies country on line 3
[!] FOUND HOMOGLYPHIC CHARACTER GREEK_B ON LINE 5
[!] FOUND HOMOGLYPHIC CHARACTER GREEK_C ON LINE 5
Subject: (B)udget (C)uts
[*] Output file closed
```

Note: The relevant characters will be underlined - not enclosed by parentheses. SafeText will output to infile.safe.