

Join GitHub today

GitHub is home to over 20 million developers working together to host and review code, manage projects, and build software together.

[Sign up](#)

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

[#forensic-analysis](#)

22 commits

1 branch

0 releases

Fetching contributors

MIT

master ▾

[New pull request](#)[Find file](#)[Clone or download ▾](#)

Fetching latest commit...

.gitignore	gitignore	Jan 2, 2018
BRIT_SPELLS	Added more unique country spellings	Jan 22, 2018
LICENSE	Initial commit	Dec 31, 2017
README.md	Update README.md	Jan 3, 2018
TestFile.txt	Fixed printing issue	Jan 3, 2018
TestFile.txt.safe	Fixed printing issue	Jan 3, 2018
US_SPELLS	Added more unique country spellings	Jan 22, 2018
characters_safetext.py	Proper coding for file	Jan 7, 2018
safetext.py	Added more unique country spellings	Jan 22, 2018

[README.md](#)

SafeText

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

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 HOMOGYPHIC 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 HOMOGYPHIC CHARACTER GREEK_B ON LINE 5
[!] FOUND HOMOGYPHIC 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.