String Search Data Set User Guide

Read Me First

Overview

Welcome to the NIST/CFTT String Search Data Set. This is Version 1.1 of test data for string searching using Federated Testing.

This data set is intended to be used with Federated Testing Version 4.0 and later versions for testing string search features.

Do This First:

Copy the folder named, “copy-to-test-computer” to the computer where the search tool you want to test is installed. The folder contains two image files (ss-win-07-25-18.dd and ss-unix-07-25-18.dd) that contain the target strings and two files with target strings to search for. The file string-search-test-cases.html has the strings for each test case (except for test case FT-SS-07-Norm). The strings for FT-SS-07-Norm are in the file ft-ss-07-norm-strings.txt. Each of the text strings has two different representations and each version must be searched for separately. If you do a hex dump of the file you will see that each version of the target strings is different in representation of accents, tilde, umlaut and ligature.

Other Files

The other files can be placed anywhere. These files are for recording test results (expected results by ID and string search data sets meta-data results) and for reference (string_doc and hit-dump) if you want to look deeper at the image files.

File Inventory:

FILE	Content
read-me-first.rtf	This File
Contents of folder to copy to test Computer	
ss-win-07-25-18.dd	Image file for Windows file systems: FAT, ExFAT, NTFS & unallocated space. This file should be attached, imported or added to the search tool as evidence.
ss-unix-07-25-18.dd	Image file for UNIX-like file systems: ext4, HFS+ journaled (OSXJ), HFS+ case-sensitive & APFS. This file should be attached, imported or added to the search tool as evidence
string-search-test-cases.html	Specification of each test case, including search tool settings, the search string and a brief test case description. For non-English search strings (and English too), the string can be copied from the html file and pasted into the search tool.
ft-ss-07-norm-strings.txt	Search Targets for Unicode normalized strings. You must copy and paste the strings into the tool being tested. Each string appears twice, both look the same, but the actual representations are different.
Other Files – Can be Anywhere	

FILE	Content
win_string_doc.txt	Location of target strings in Windows dd file. Listed by string ID.
win_hit-dump.txt	Hex dump around each target string in Windows dd file.
unix_string_doc.txt	Location of target strings in UNIX dd file. Listed by string ID.
unix_hit-dump.txt	Hex dump around each string in UNIX dd file.
Expected Results by ID	List of expected results for each test case. This can be used to make a paper copy of test results to record results for later entry into the Federated Testing DVD test results page. For example, if you have only one computer for both testing and running the Federated Testing DVD you can run tests, recording results to paper and then reboot the NIST DVD to record results and generate a test report.
String Search Data Sets Meta-Data Results	List of expected Meta-Data hits. This can be used to make a paper copy of expected meta-data search hits for later entry in the Federated Testing page.

Creating the Image Files

The dd files were created as follows:

1) A set of base strings was selected.
2) For each base string, a set of files was created with the following characteristics:
    i) Create two files for each partition and one file for unallocated space.
    ii) The file name has the form: **STATUS-AltName-Partition-Encoding.txt**
        (1) STATUS is either LIVE or DELETED reflecting the final status of the file.
        (2) AltName is a word or two words with similar meaning to the base string, e.g., the string WOLF has AltName of AllCap-Lupus. Two files used in the meta-data test case use the base string instead of an AltName (this gets the base string into meta-data, but with no string ID).
        (3) Partition is where the file is located. One of these: fat, exfat, ntfs, unalloc, osxj, osxc, apfs or ext4.
        (4) Encoding is the character encoding of the base string. One of these: ascii, utf-8, utf-16be or utf-16le**.**
    iii) File content begins with: "TESTFILE:(**FILE NAME**)"
    iv) A list of random filler words with a nautical theme, e.g., creek, sea, river, tuna, squid, etc.
    v) The base string (with a unique string ID for each string instance. One string instance in each file), formatted as follows: **encoding ====> STRING ID <==== partition**
    vi) More random nautical filler words.
    vii) The file ends with: Encoding (**FILE NAME**) END-OF-FILE
3) Files are copied to destination partitions on a zero-wiped 2GB drive.
4) Any special case files are copied to the drive, e.g., "lost files."

5) Files with STATUS of DELETED are deleted.
6) The drive is imaged.
7) The image file (.dd) is scanned for each string used as a search target in a test case. The files xxx_string_doc.txt and xxx_hit-dump.txt are created to document the location of each search target.
8) A similar set of steps is followed to create the Unix-like file system image file.

That's all, email CFTT@NIST.GOV if you have questions or need help.