WebHandyTool
User Guide : newAGEtech, Group H

Genevieve Okon (Okong), Abraham Omorogbe(Omorogoa),
Eric Le Forti(Leforte)

December 7, 2015

Contents

List of Tables

Revision History

Revision	Revision Date	Description of Change	Author
1	20-10-15	Initiate Test Plan Document and Introduction	Eric Le Fort
2	26-11-2015	Finalize Outline	Abraham Omorogbe
3	26-11-2015	Functional system tests	Abraham Omorogbe
4	26-11-2015	Functional system tests	Eric Le Fort
5	26-11-2015	Non-Functional system tests	Eric Le Fort
6	26-11-2015	Usability Testing	Eric Le Fort
7	26-11-2015	Requirements Traceability	Eric Le Fort
8	27-11-2015	Testing Summary	Genevieve Okon
9	27-11-2015	Code Coverage	Genevieve Okon

Table 1: Revision History

1 Introduction

1.1 Purpose

WebHandyTool was created in an attempt to ease the difficult and tedious process of verifying a website's current status and state of availability as well as to assist researchers in scouring the web efficiently for potentially relevant sources of data. Some specific tasks expected to be completed using this software include: ensuring all of a website's pages are functioning correctly and resources that should be available are, data mining for certain research topics and for security professionals ensuring that only the webpages that should be visible are.

WebHandyTool will allow any user to search for any query on a website, crawl a website, check a website for error and download all the website's resources.

1.2 Tables of Acronyms, Abbreviations & Definitions

Term	Definition
PIP	A package management system used to install and manage software packages written in Python.
UNIX	Computer operating system
Beautiful Soup	An existing framework that breaks a webpage down into its components.
WGET	A computer program that retrieves content from web servers.
HTML Status Code	A three digit number that corresponds to various states of a website.

Table 2: Definitions

2 Installation Instructions

2.1 Safety and Precaution

Setting the depth above 2 may result in extremely long processing times depending on the website in question, use it with caution and be patient when you select this option.

2.2 OS X

1. Make sure Python is installed on your machine. If it is not, install it from here: `https://www.python.org/downloads/`
2. Install pip on your machine from here: `https://pip.pypa.io/en/stable/installing/`
3. Install Beautiful Soup 4, go the terminal and run the command: *pip install beautifulsoup4*

**If you are using the downloading option, your must have wget installed.*

4. If not installed already, install wget from here: `http://coolestguidesontheplanet.com/install-and-conf`

2.3 Linux

1. Make sure Python is installed on your machine. If it is not install it from here: `https://www.python.org/downloads/`
2. Install pip on your machine, from here: `https://pip.pypa.io/en/stable/installing/`
3. Install Beautiful Soup 4, go the terminal and run the command: *pip install beautifulsoup4*

**If you are using the downloading option, your must have wget installed.*

4. If not installed already, install wget from here `http://www.ehowstuff.com/how-to-install-wget-on-linux`

2.4 Windows

**Windows is not recommended, try to use a UNIX based machine.*

1. Make sure Python is installed on your machine. If it is not install Python `https://www.python.org/downloads/`
2. Install pip on your machine, `https://pip.pypa.io/en/stable/installing/`
3. Install Beautiful Soup 4, go the terminal and run the command *pip install beautifulsoup4*

**If you are using the downloading option, your must have wget installed.*

4. If not installed already, install wget `http://gnuwin32.sourceforge.net/packages/wget.htm`

3 Getting Started

1. Go the terminal and make sure you are in the correct directory, it should be named WebHandyTool
2. Run the command: *python Web_Crawler.py*
3. Select your options. (option description below)
4. Enter the website, you want to use.
5. Watch WebHandyTool work its magic. The tool will constantly update the screen with the links it is currently working on.

3.1 Options

3.1.1 Select 1 for Download Resources

This option will download all the data from the website you entered. The file type that is downloaded is set in the config.py file (Refer to the configurations section). The resources are stored in the source folder under the domain name of the selected website.

3.1.2 Select 2 for Check for Errors

This option will return a status code from website crawled. The depth is set in the config.py file (Refer to the configurations section)

3.1.3 Select 3 for Search for Query

This option will return the index of where the query was found, and 30 characters before and after the search query. The option also shows you the page where the query was found. The search type (exact or similar) is set in the config.py file (Refer to the configurations section)

3.1.4 Select 4 for Just Crawl

This option will return a list of all the links that were found on the website specified by the user. The depth is set in the config.py file (Refer to the configurations section)

4 Configurations

A user can configure several aspects of WebHandyTool.

4.1 String Search

To change search type, use the following options, and change "type" variable
Exact Search = 1
Similar Search = 2
To change similar search proximity, use the following options, and change "proximity" variable (Refer to the FAQ's to understand similar search better)

4.2 Depth

Change the "depth" variable to increase or reduce the depth the crawl traverses.

4.3 Download

You can specify the downloader to only retrieve file type you specify (None will download all types). Change "file_type" to any thing you wish (*.pdf,*.png, etc.)
To alter the downloaders option you change the "option" variable. All the options are from wget library, however we highly recommend you only use the options specified below.
-r : recursive, downloads every link on that domain.
-np : no-parent, only downloads children links. It will NOT download from links that are not connected to the specified site.
**To join both commands simple write "-rnp"*

5 Troubleshooting

If you run into the following error:
raise URLError(err) urllib2.URLError: <urlopen error [Errno 8] nodename nor servname provided, or not known> You must have entered an invalid url.

6 Frequently Asked Questions

How do I stop the crawl in the middle of execution?
If you press Ctrl+C it should stop the crawler.

What does similar search do?
Searches through a String for a certain phrase or term. Returns results that are close to the query as well. (i.e. "ap ple" or "bpple" would be noted for "apple") Returns the starting index for all occurrences of Strings sufficiently close to the query. If the query is not located, it will return an empty array.