pylinkvalidator
Design Document Module Guide : newAGEtech, Group H

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

November 6, 2015

Contents

List of Figures

List of Tables

1 Introduction

2 Anticipated & Likely Changes

3 Module Hierarchy

This section contains the module design structure of our project. Modules are summarized in a hierarchy as shown in Table 1. The modules listed below, most of which are leaves in the hierarchy tree, are the modules that will actually be implemented.

Modules

M1: Option Module
M2: Crawler Module
M3: HTML Corrector Module
M4: Download Resources Module
M5: Exact Query Search Module
M6: Similar Query Search Module
M7: Whitespace Checker Module
M8: Find Links Module
M9: Check Errors Module
M10: Parse Data Module
M11: Query Search Module
M12: Depth Setter Module
M13: Website Structure Modelling Module

Note that M5, M6, M7 are submodules of the larger Query Search Module. M2 is the highest level module and utilizes all others internally.

Hardware hiding

N/A

Behaviour Hiding Modules

- Download Resources Module
- Exact Query Search Module
- Similar Query Search Module
- Whitespace Checker Module
- Find Links Module
- Check Errors Module
- Parse Data Module

- Query Search Module
- Website Structure Modelling Module

Software Decision Modules

- Option Module
- Crawler Module
- HTML Corrector Module
- Depth Setter Module

4 Connection Between Requirements & Design

This design was developed using the requirements document to help guide the decomposition of the project's modules. The requirements were matched to corresponding modules which complete the various tasks. For example, Requirement #2 from the requirements document (The product shall download resources from a website) will be accomplished using module M4.

5 Module Decomposition

5.1 Hardware Hiding Modules

N/A

5.2 Behaviour Hiding Modules

5.2.1 Download Resources

void downloadResources(String: link, String: fileType, String: destination)

Secrets: Parse through the HTML code in the link provided in order to locate all files that match the specified file type. For each file, the result is downloaded into a folder specified by the user.

Services: Writes all resources matching the given file type from the page link to the file specified by destination.

Implemented By: Python

5.2.2 Exact Query Search

int[] searchForString(String: query, String: data)

Secrets: Iterates through the data provided and records every instance matching the query String that was passed in. This will be accomplished using the Knuth-Morris-Pratt String searching algorithm.

Services: Returns a list of all occurrences of a given query in the data provided.

Implemented By: Python

5.2.3 Similar Query Search

int[] searchForSimilarString(String: query, String: data, int: proximity)

Secrets: Iterates through the data provided and records every instance sufficiently close to matching the query String that was passed in. This will be accomplished using a slight deviation from the Knuth-Morris-Pratt String searching algorithm that recognizes fuzzy string searching.

Services: Returns a list of all occurrences within a certain deviation of a given query in the data provided.

Implemented By: Python

5.2.4 Whitespace Checker

boolean isWhitespace(char: character)

Secrets: Checks to see if the character passed in is a tab, a space or a new line character.

Services: Returns whether the character passed in is certain types of whitespace.

Implemented By: Python

5.2.5 Find Links

List¡Links¿ findLinks(BeautifulSoup: data, String: destination) Exceptions: No Data Found

Secrets: Parser that parsers through the HTML code in the data provided

Services: Finds all the links (<a> anchor tags on page) on a page

Implemented By: Python

5.2.6 Check Errors

List¡Errors¿ checkErrors(String: link, Array: List of links) Exceptions: Webpage Unavailable

Secrets: Algorithm to check the header in all the links provided

Services: Checks the all the links and reports the error message associated with all the links inputed

Implemented By: Python

5.2.7 Parse Data

BeautifulSoup parseData(String: link) Exceptions: Invalid Link

Secrets: Converter that converts HTML code link to BeautifulSoup object

Services: Returns Beautiful object for the link given, this will allow modules parse through pages data much faster

Implemented By: Python

5.2.8 Query Search

void querySearch(String: Query, BeautifulSoup: data, String: choice) Exceptions: Invalid Choice

Environment Variables: rawInput: Users keyboard input

Secrets: An algorithm that figures out what query search type to implement

Services: Writes all resources matching the given file type from the page link to the file specified by destination.

Implemented By: Python

5.2.9 Website Structure Modelling

String webStructureModel(String: link, int: depth)

Secrets: An algorithm that uses the seed link and structures the crawled links by depth

Services: It provides a structured model of the website and other site the initial site is connect to. It displays a hierarchy that will show users how crawled link interact with each other.

Implemented By: Python

5.3 Software Decision Modules

5.3.1 Options

void chooseOption()

Secrets: Obtains input from the user to decide whether they would like to download resources, crawl a website, search for a certain String or model a website. Uses searchDepth to allow the user to set their search depth for all but the download resources option as well as allowing the user to set their seed link.

Services: Gets user input to set various program options related to how the user would like to handle crawling a webpage.

Implemented By: Python

5.3.2 Crawler

void crawler()

Secrets: Directly calls the methods shown in the Use Hierarchy (found in section 7 of this document) and consolidates the results.

Services: Delegates various tasks of crawling a webpage to the other methods of this program.

Implemented By: Python

5.3.3 HTML Corrector

String HTMLCorrector(String: link)

Secrets: Tries to fix the given String in various ways in order to create a functioning link. This can include fixing the prefix (http://), adding www., or appending the link to the current page's path.

Services: Fixes the link passed in such that it becomes either a functioning link or is flagged as a broken link.

Implemented By: Python

5.3.4 Depth Setter

int depthSetter(int depth) Exception: Not positive int

Environment Variables
rawInput: Users keyboard input

Secrets Assignor which sets the depth variable

Services Sets the default max depth variable for the web crawler

Implemented By: Python

6 Traceability Matrix

Requirements	Modules
R1	M1, M2, M3, M9, M10
R2	M1, M2, M3, M4, M10
R3	M12
R4	M2
R5	M1, M2, M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, M13
R6	M1, M2, M3, M8, M10, M12, M13
R7	M1, M2, M3, M5, M6, M7, M8, M10, M11, M12

Table 1: Traceback to Requirements

Requirements	Modules
AC1	
AC2	

Table 2: Traceback to Anticipated Changes

7 Use Hierarchy Between Modules

The figure below depicts the uses relationships between all the modules in the project. It can be seen that the graph is a directed acyclic graph (DAG). The facade design pattern is being used to design this system. Higher level modules in relation to the hierarchy are inherently simpler because they delegate work to modules from the lower levels.