Requirements Analysis Doc
Tech Sites WebCrawler

1.Introduction

1. Purpose of the System

    1.1.Gather email addresses from the web

2. Scope of the System

    2.1.Will be used once by a user to get a list of emails

3. Core System Functionalities

    3.1.Given a set of seeds, we collect a list of emails for people related to tech in the shenandoah valley

4. Objectives and Success Criteria

    4.1. We get a list of email addresses

5. Definitions, Acronyms, Abbreviations

    5.1. seed — the initial site given to the crawler

6. References

    6.1.https://en.wikipedia.org/wiki/Web_crawler

2.Current System

None

3.Proposed System

1. Overview

    1.1.System will be given a few sites to start with and then it will collect email addresses from sites and sites that they link to and return them in a file

2. Functional Requirements

    (The requirements should be documented, actionable, measurable, testable, traceable, related to identified business needs or opportunities, and defined to a level of detail sufficient for system design.)

Program has options to be terminated based on time, # of emails, # of pages, or until out of pages

Doesn't record duplicate emails and doesn't visit sites more than once

Visit html sites ending in {.com, .edu, .gov, .org, .net}

Addresses outputted to file

Will obey robots.txt and other website directions

will be multithreaded to visit multiple sites at once and will not ping sites too often to not give them unnecessary traffic load

3. Non-Functional Requirements
    - 3.1. Usability
        - 3.1.1. User knows how to run python
    - 3.2. Reliability
        - 3.2.1. Handles errors
    - 3.3. Performance
        - 3.3.1. No efficiency requirements but should be within reason
    - 3.4. Supportability
        - 3.4.1. Will use as few packages as possible to increase compatibility
        - 3.4.2. Code will be well commented
    - 3.5. Implementation
        - 3.5.1. Implemented in Python

4. System Models
    - 4.1. Use-Case Diagram
        - 4.1.1. User —->address -—> program —-> addresses —-> emails —-> file -—>User
    - 4.2. Object Model (Omit)
    - 4.3. User-Interface

        Will likely be run from the command line unless a GUI seems like a useful feature

4.References

https://en.wikipedia.org/wiki/Web_crawler

https://en.wikipedia.org/wiki/Email_address_harvesting