Project notes

2019-05-06 22:55
1a0f14a

Contents

Synopsis

webgc

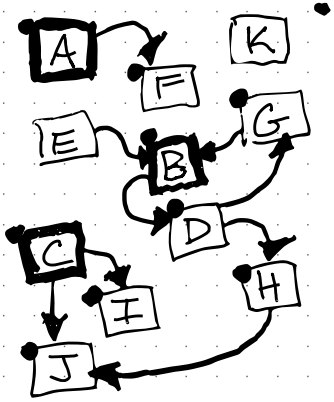
Wed. 20 March 2019

- Web pages contain links to "assets":
  - image files, • style sheets, • scripts
  - arbitrary files to download; AND
  - other web pages.
- As site evolves, two things can happen to links:
  1. Link becomes broken (aka "dangling") when asset it referenced is moved/renamed/deleted.
  2. An asset might no longer be referenced by any link (aka "unreachable" or "garbage"), perhaps because it's no longer needed or relevant.
- There exist many tools that can help us find dangling links – our tool (also) finds unreachable assets.

GC = Garbage Collector / Collection

Garbage Collection Overview

One popular "batch" technique is called "mark-sweep".



- It starts with a "root set" — nodes which are considered to be the entry points.
  Suppose
  roots = {A, B, C}
- Then it traverses & MARKS "live set" - nodes that are reachable from roots by following arrows.

live = {A, B, C, F, D, G, H, J, I}

- Finally, it SWEEPS away (aka deletes, collects, recycles) any nodes not marked as live.
  In this example, {E, K} would be collected.