IPSW - Modelling Change of Website Archives

Group 4

1 Introduction

The goal is to construct and compare different metrics to quantify domain changes over time. We aim to determine a single quantitative measure at each time that encapsulates the changes the magnitude of the change in the domain since the previous time-step.

$$\sigma(t) = (\text{change in links})w_1 + (\text{change in text})w_2 + (\text{change in content management server})w_3, \quad (1)$$

where t is time, and w_1, w_2, and w_3 weight the relevant contributions of URL changes, text changes, and CMS changes.

We will compare this with another metric for quantifying change

2 Game plan

- Run code to compare text.
- Do image analysis on thumbnails.
- Take link data and compare lists at different times:
  - Internal vs. external links.
  - Obtain a, b, and c.
  - What is the best timestep?
- Determine whether the content management server (CMS) has changed.
- Look at different weightings - how best to choose these? We don't want to double-count changes.
- Run test cases.
- Look at the variability in change over time. What is the distribution?
- Compare measures for looking at the difference between URLS and text.