Day 3 Notes

$$y = mx + b$$

{#eq:description}

Web Crawling

- Systematic browsing of web for purpose of indexing and updating
- BFS: unformly explore from entry
- DFS: explore by branch

Focused Crawling

- Prioritize visiting sequence of the web
- Size of web too large for a crawler to cover
- Emphasize high quality documents. Maximize weighted coverage.
- –add equation–

Prioritize by in-degree

- page with highest incoming hyperlinks from previously crawled pages is crawled next

Prioritize by page-rank

- BFS at eary stage
- Compute pagerank periodically
- More consistent with search relevance

Prioritize by topical relevance

- Vertical search. Only crawl relevant pages
- Esimate similarity to current by by anchor or text near anchor

Avoid duplicates

- Check using a trie or hash table

Politness policy

- Robots.txt: Policy that allows crawler to download

Basic Text Analysis

- Extract informative content. Analyze and index the cralwed web pages

HTML parsing

- Shallow parsing : remove html tags. only keep text between title and paragraph tags
- Visual parsing: frequent pattern mining of visually similar HTML Blocks
- Issue with representing by string? No semantic meaning
- Represent list of sentences.. sentence short doc
- Represent by list of words. Tokenize first

Full text index

- Represent words vs documents as adjacency matrix.