Detecting Duplicates

Web Crawling produces Duplicates

- Exact Duplicates
  - Easy to check
- Near Duplicates
  - Hard to check
  - Web pages can have same text content but different layout/style

Techniques

- Checksumming
- Cyclic Redundancy Check (CRC)
- Near-duplicates defined by some threshold of a similarity measure between two documents, ex 90%

Fingerprinting

- N-grams and hashing
- Simhashing