CSCI 572 - Information Retrieval

Comprehensive Midterm Study Guide

Study Guide Summary:

Total Lectures Covered: 10

Key Concepts: 867

Important Definitions: 609

Formulas & Algorithms: 267

Examples & Case Studies: 382

Lectures Covered:

- Deduplication (30 slides analyzed)
- Info Retrieval (45 slides analyzed)
- Inverted Indexing (38 slides analyzed)
- Querying (52 slides analyzed)
- Se-Basics (31 slides analyzed)
- Se-Evaluation (31 slides analyzed)
- Text Processing (27 slides analyzed)
- Web Crawling (44 slides analyzed)
- Web Serving Basics (37 slides analyzed)
- Youtube (39 slides analyzed)

Quick Reference - Key Concepts:

- 1. **De-Duplication** - (process of identifying and avoiding essentially identi...
- 2. **Locker Storage** - (strategy where only single copy of file is stored with...
- 1. **Deduplication**: - The process of removing duplicate copies of data or URL...
- 2. **Virtual hosts**: - A feature that allows multiple hostnames to share the s...
- 3. **URL structure**: - The components of a URL (protocol, hostname, path, page...
- Deduplication: removing duplicate data or records to improve efficiency and redu...
- Data Duplication
- Deduplication (concept of removing duplicate data)
- Similarity between web pages (differing slightly)
- Deduplication (not explicitly mentioned in this slide, but relevant to the topic...
- * Mirroring ()
- Apache mirrors
- Deduplication (implied by the context of the slide)
- **Deduplication**: Not explicitly mentioned in the text, but implied to be relat...
- * **Deduplication**: Avoiding or minimizing duplicate results in crawling
- * **Smarter Crawling**: Optimizing crawling to reduce resources and increase po...
- * **Better Connectivity Analysis**: Combining in-links from multiple mirror sit...
- **Deduplication**: The process of identifying and removing duplicate or near-dup...
- * Duplicate Problem: Exact match vs. Near-Duplicate Problem: Approximate match
- * Cryptographic hashing for exact match detection

... and 847 more concepts