

State of the Quarter

Introduction to Information Retrieval

CS 221

Donald J. Patterson



Course Evaluation

- https://eee.uci.edu/toolbox/eval/form/preview_form.php?preview_id=40723&preview_type_id=56&preview_ccode=35220
- Card Evaluation
 - What did we spend too much time on?
 - What did we spend too little time on?
 - What are you most excited to have learned?



What did we do this quarter?

- 29 lectures
- 3 quizzes
- 7 assignments
- 1 field trip
- 1 guest lecture



From the text book we covered

- Chapter 19 - Web search basics
- Chapter 20 - Web crawling and indexes
- Chapter 4 - Index construction
- Chapter 1 - Information Retrieval Basics
- Chapter 6 - Scoring, term weighting and vector space model
- Chapter 18 - Matrix decomposition and Latent Semantic Indexing
- Chapter 21 - Link analysis
- Chapter 8 - Evaluation in information retrieval



Supplementary Readings included

- The background on Vannevar Bush and the Memex
- Looking at the web as a graph
 - Statistics about how it is connected.
 - How to compress a web graph so you can work with it in memory.
- The first publication about Google's architecture



Assignments

- Asked for information from and about you for context.
- You wrote a web crawler.
 - You searched for specific information
 - You searched for specific paths in the web graph
- You created a web-search U/I
 - To be embedded in firefox
- You created an index of your web crawl
- You implemented a ranked relevance query engine
- Built (will build !) an embedded search engine



Web Search Basics

- HTML
 - Basics of tagging and how HTML translates into a web graph
 - Meta tag keywords
 - Context around links for various IR uses



Web Search Basics

- Behavior around web search
 - Search engine usage
 - The role that search plays in scaling the internet
- Ads and search
 - History
 - Incentives
 - Business Models



Web Search Basics

- Terminology
 - Corpus
 - Relevance
- Differences between classic IR and web IR
- History of web IR
 - business model development
- The web corpus
 - Characteristics of it.



Web Search Basics

- Dynamic pages
 - How does it work
- The web as a graph
 - Construction
 - Characteristics
 - How big is it
 - Rate of change



Web Search Basics

- User needs
- Expectations of users
- The web as a graph
 - Construction
 - Characteristics
 - How big is it
 - Rate of change



Web Crawling Basics

- URL Frontier
- Basic Crawl Algorithm
- Crawling in reality
 - Politeness
- Robust Crawling
 - DNS caching
 - Other stages in process
 - what do they do? what are the concerns?
- Desired characteristics of a web crawler



Web Crawling Basics

- Mercator implementation
 - Front and back queues
 - issues associated with that.



Web indices

- What are we indexing?
- Vector Space Model
 - Term Document Matrix
- WebGraph compression
 - How does it work?



MapReduce

- Architecture
- How you might use it for creating posting lists
- Google just announced MapReduce enable them to sort
 - 20,000,000,000,000,000 bytes in six hours (20 PB)
 - (3 MB for every person in the world)



LSI

- What is it?
- How do you use it to capture “semantics”?
- Demo of how to prototype your own solutions



Spam

- Characteristics
- Reasons that it exists
- Different ways that it occurs



Helping the user

- Information needs
- query shortcuts
- implicit context
 - types of context
- aggregation of results



Index details

- Term document pairs
- Posting lists
 - Construction
- Index scaling
- implicit context
 - types of context
- BSBI SPMI

