Review
Introduction

• Data and information
• Information retrieval: concept, content, applications, and tasks
• Relevance, retrieval models, and evaluation
• Information needs and queries
• Search engines
Search Engine Architecture

• Primary goals of search engines
• Two major functions: the index process and the query process
• The index process: text acquisition, text transformation and index creation
• The query process: user interaction, ranking and evaluation
Web Crawling

- Web crawling and must-have/should-have features
- Basic crawler architecture
- Politeness policy
- Freshness measure
- Distributed crawlers
- Compression
- Big table
- Duplicate/near-duplicate detection: fingerprints, shingling, simhash, …
Processing Text

- Word frequencies and Zipf’s law
- Vocabulary growth and Heap’s law
- Estimating result size
- Tokenization
- Stop words
- Stemming and lemmatization
- Spelling correction
- Edit distance
Indexing and Ranking

• Inverted index
• Blocked sort-based indexing (BSBI) and single pass in-memory indexing
• MapReduce: concepts and basic pseudocode
• Incremental maintenance of inverted index
• Result merging and using multiple indexes
• Vocabulary and storage
Indexing and Ranking

• Space and time tradeoff and compression
• Delta decoding, Elias-$\gamma$ codes and Elias-$\delta$ codes
• Skipping and efficient implementation
• Scoring: fields and zones
• Parametric indexes, zone indexes
• Learning weights
• TF/IDF
• Document vectors, cosine similarity
• Document/Query Weighting, TF normalization
Indexing and Ranking

• Document-at-a-time evaluation and term-at-a-time evaluation
• Efficient scoring
• Phrase queries
• Threshold adjustment algorithm
Queries and Interfaces

- Vocabulary look up
- Wildcard queries
- Term association and word co-occurrence measures
- Noisy-channel model and spell checking
- Relevance feedback, the Rocchio algorithm, pseudo-relevance feedback
- Result snippet generation
- Advertising and sponsored search
- Clustering search results
Information Retrieval Models

- Information needs and queries
- Relevance, precision and recall
- Retrieval models
- The Boolean retrieval model
- Probabilistic retrieval model
- Probabilistic ranking principle (PRP)
- Binary independence model (BIM)
- Ranking by odds and retrieval status value (RSV)
- Language models
- Maximum likelihood estimation and smoothing
- Ranking based on model comparison
Evaluation

- Text collections
- Relevance to information needs
- Subjective versus objective judgements
- Effectiveness versus efficiency
- Relevance judgements, Kappa statistics
- Precision, recall and tradeoff, type I and type II errors
- Evaluation of ranked lists: precision at rank $p$, precision-recall curve, average precision, mean-average precision, discounted cumulative gain (DCG), normalized DCG
- Efficiency measures
- Significance tests
- A/B tests
Link Analysis and Web Search

- Hyperlinks and anchor text
- Random walks and PageRank, topic-based and personalized PageRank
- Hubs and authorities
- Search engine index size estimation
Exam Questions

- All multiple choice questions
- Questions about concepts, principles, and applications
  - Section 1: each question has only one correct answer
  - Section 2: each question has one or multiple questions
- Questions about algorithms and calculation
- In total about 40-50 questions for 180 minutes