IT-550 Lecture Notes (Lectures 1-13)

Instructor: Parth Mehta

Lecture 1 - Introduction - (23.07.2024)

Reading Material

- PPT

Topics Covered

1. Introduction to the course
2. Course Plan and Evaluation

Lecture 2 - Basics of Search (24.07.2024)

Reading Material

- Chapter 1 - Boolean Retrieval
- Chapter 2, section 2.2 - Text Preprocessing
- Chapter 2, section 2.3 - Skip pointers and Fast posting list intersection. Not discussed in class, but some of you might find it interesting and useful.

Topics Covered

1. Basics of search - Dataset, Information need, Retrieval Mechanism and Evaluation
2. Simple Search - Regex and Grep
3. Boolean Indexing and Querying
4. Issues with sparse Term-Document Matrix
5. Inverted Index using posting list
6. Query matching in posting list
7. Text Pre-processing Steps

Lecture 3 and 4 - Vector Space model (26/29.07.2024)

Reading Material

- Chapter 6 - sections 6.2 to 6.4
- Pivoted Document Normalization (SIGIR 1996)
- Chapter 6 - Section 6.1 How to assign different wights to words in different parts of the document (e.g. title, metadata, etc). Not discussed in class, but interesting and useful.

Topics Covered

1. Term and Document Frequencies
2. Vector space model for retrieval
3. Pivoted Document Normalization

Lecture 5 and 6 - Efficient Retrieval (31.07/05.08.2024)

Reading Material

- Chapter 7 - sections 7.1

Topics Covered

1. Term at a time (TAAT) and Document at a time (DAAT) scoring
2. Inexact Retrieval
3. Index Elimination
4. Champion List
5. Static Document Scores
6. Impact Ordering
7. Cluster Pruning

Lecture 7 and 8 - Evaluation of IR Systems (07/09.08.2024)

Reading Material

- Chapter 8 - All sections. Sections 8.5-8.7 were briefly touched upon in the class and are self-read.

Topics Covered

1. Test Collections and the Cranfield Experiment
2. Manual relevance judgments and document pooling
3. Basic Evaluation Metrics - Recall, Precision, F-score
4. R-P Curve
5. Advanced evaluation metrics - Mean Average Precision, R-Precision, ROC Curve
6. Evaluation metrics for Ranked Retrieval - Cumulative gain (CG), Discounted CG, Normalized Discounted CG

Lecture 9 - Relevance Feedback and Query Expansion (12.08.2024)

Reading Material

- Chapter 9 - All sections

Topics Covered

1. Methods of Query expansion - Dictionary and Synonym lookup, Spelling normalization, etc
2. Relevance feedback and Pseudo Relevance feedback
3. Optimal query and Query reformulation using relevance feedback
4. Probabilistic Relevance feedback
5. Evaluation after query expansion
6. Limitations of Query expansion

Lecture 10 and 11 - Probabilistic IR (14/21.08.2024)

Reading Material

- Chapter 11 - All sections

Topics Covered

1. Basics of Probability Theory
2. Probability Ranking Principle
3. Binary Independence model
4. Probabilistic Relevance feedback
5. Okapi BM25

Lecture 12 and 13 - Language Models (23/27.08.2024)

Reading Material

- Chapter 12 - All sections
- Using language models for information retrieval - 2002.

  Trivia: The popular Hiemstra's language model was proposed by Djoerd Hiemstra as a part of his PhD. The thesis (link above) would make an exciting read for anyone looking to explore Language models for IR in detail. For this course, we will focus only on chapter 4.

Topics Covered

1. Language Model
2. Query Likelihood Model
3. Hiemstra's Language Model
