Web Information Retrieval syllabus

A.A. 2020/2021

Intro to IR: Boolean Retrieval

(Prof. Luca Becchetti)

1. What is Information Retrieval (IR)?
2. A first simple example
3. Boolean Queries

[1] Chapter 1

Term vocabulary and posting list

(Prof. Luca Becchetti)

1. Terms, tokenization lemmatization and stemming
2. Skip pointers
3. Positional indexes and phrase queries
4. Index, tolerant retrieval and spell correction

[1] Chapter 2 and Chapter 3, Sections 3.1 - 3.3

Index construction

(Prof. Luca Becchetti)

1. Algorithms for index construction and sorting
2. BSBI and SPIMI algorithms
3. Distributed indexing
4. Dynamic indexing

[1] Sections 4.1 to 4.5

Scoring and vector space model

(Prof. Luca Becchetti)

- Term and document frequency
- tf-idf weighing
- Vector model
- Efficient query answering - techniques and architectural considerations (reading below-mentioned sections on this part is up to students)

[1] Chapter 6, Sections 7.1 and 7.2, Jupyter notebook made available by instructor

Language models for IR

- Probabilistic information retrieval and Bayes' rule
- The Probabilistic Ranking Principle (PRP)
- The Binary Independence Model (BIM)
- Generative language models: the multinomial model

For intro, PRP and BIM: [1], Sections 11.1 - 11.3

For generative language models: 12.1 - 12.3

Link Analysis

- Basic notions on Markov chains
- Pagerank and its computation

[1] Sections 21.1 and 21.2

[2] Section 5.1

References

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press. 2008. [2] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2014. Slides and book chapters available on http://web.stanford.edu/class/cs246/ [3] Notes by the instructors [4] Jacob Eisenstein. *Introduction to Natural Language Processing*. The MIT Press, 2020.