****

**SECOND SEMESTER 2023-2024**

# Course Handout Part II

09-01-2024

In addition to part-I (General Handout for all courses appended to the time table) this portion gives further specific details regarding the course.

Course No. : CS F469

Course Title : INFORMATION RETREIVAL

Instructor-in-Charge : Dr. Prajna Devi Upadhyay ([prajna.u@hyderabad.bits-pilani.ac.in](mailto:prajna.u@hyderabad.bits-pilani.ac.in))

### 1. Scope and Objectives

This course studies the theory, design, and implementation of text-based information systems. The Information Retrieval core components of the course include statistical characteristics of text, representation of information needs and documents, several important retrieval models and their evaluation measures. The student should also

The student should be able:

* To understand the architecture of information retrieval systems – crawling, indexing, and retrieval
* To analyze data structures for indexing large collections
* To compare and implement different retrieval models – Boolean, Vector-based, Probabilistic, Learning to Rank, Neural, and LLM-based, understand topic models such as LDA and LSA
* To get familiar with the design of test collections (TREC, crowd-sourcing) and evaluation measures (precision, recall, micro-/macro-F measure, nDCG)
* To understand and model Knowledge Graphs for retrieval
* To understand ethical issues related to Information Retrieval

**2. Pre requisites:**

* Programming in Python and/or Java
* Knowledge of core data structures and algorithms.

##### 3.a. Text Book

* **T**. C. D. Manning, P. Raghavan and H. Schutze. Introduction to Information Retrieval, Cambridge University Press, 2008. The entire book is available at <http://nlp.stanford.edu/IR-book/>

**3.b. Reference Books and Other Resources**

* **R1:** Modern Information Retrieval, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley, 2000. <http://people.ischool.berkeley.edu/~hearst/irbook/>
* **R2**:Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, Cambridge University Press
* **R3**: [Domain-Specific Knowledge Graph Construction](https://link.springer.com/book/10.1007/978-3-030-12375-8), Mayank Kejriwal, Springer
* **R4**: Entity based Retrieval Models,  <https://dl.acm.org/doi/pdf/10.1145/2970398.2970423>
* **R5**: Learning to Rank for Information Retrieval, [https://link.springer.com/book/10.1007/978-3-642-14267-3](https://link.springer.com/book/10.1007/978-3-642-14267-3%20)
* **R6**: An Introduction to Neural Information Retrieval, Bhaskar Mitra and Nick Craswell, 2018. <https://www.microsoft.com/en-us/research/uploads/prod/2017/06/fntir2018-neuralir-mitra.pdf>
* **R7:**FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval. [https://sigir.org/wp-content/uploads/2019/december/p020.pdf](https://sigir.org/wp-content/uploads/2019/december/p020.pdf%C2%A0)
* **R8:** Deep Learning. Ian Goodfellow, Yoshua Bengio, Aaron Courville. MIT Press.

**4. Course Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| **Lecture No** | **Learning Outcomes** | **List of Topic Title**  **(from content structure in Part A)** | **Text/Ref Book/external resource** |
| 1 | List the course objectives and define the vocabulary used in IR | Introduction   * Basic Concepts * The retrieval process | R1 Ch1, Ch2 |
| 2 - 4 | Boolean Retrieval   * Inverted index * Processing Boolean queries * Boolean Vs Ranked retrieval * Term vocabulary and postings lists * Phrase queries * Exercises | T Ch2 |
| 5 - 7 | Evaluate and apply wild card queries and spelling correction | Dictionary and Tolerant Retrieval   * Search Structures for dictionaries * Wildcard queries * Phonetic Correction | T Ch3 |
| 8 - 10 | Understand techniques to construct and compress indexes that do not fit in memory | Index Construction and Compression   * Blocked sort-based Indexing * Single pass in-memory indexing * Distributed and dynamic indexing * Dictionary comparison * Postings file compression * Exercises | T Ch4, Ch5 |
| 11 - 12 | Apply tf-idf and cosine score to score documents against a query | Vector Space Model   * Term frequency and weighting * The vector space model for scoring * Tf-idf functions | T Ch6 |
| 13 – 15 | Get familiar with the design of test collections (TREC, crowd-sourcing) and evaluation measures (micro-/macro-F measure, nDCG) | Evaluation in IR   * TREC Collections * Evaluation of ranked results * Evaluation of unranked results * Relevance Feedback * Exercises | T Ch8, Ch9 |
| 16 – 19 | Formulate IR problem using Probabilistic approach, model documents as language models, model relevance as a query generation process | Probabilistic Retrieval and Language Models   * The Binary Independence Model * BM25 * Language Models as Multinomials * Query Likelihood | T Ch11, Ch12 |
| 20 – 22 | Formulate document collections as mixture of latent models | * LSI, LDA, LSA | T Ch18 |
| 23 - 26 | Formulate Information Retrieval as Learning Tasks | Learning to Rank   * Pointwise * Pairwise * Listwise | R5 Ch1, Ch2, Ch3, Ch4 |
| 27 – 32 | Understand Neural approaches to IR – shallow unsupervised neural algorithms such as Word2Vec and Document Autoencoders, supervised approaches such as Siamese Networks | Neural Information Retrieval   * Neural and Deep Neural Networks * Deep Neural Networks for IR * Large Language Models for IR | R8 Ch6, R6 Ch6, Ch7 |
| 33 | Get familiar with Knowledge Graphs and their storage models | Knowledge Graphs   * Introduction to Entities, Relations, and Triples * RDF and PG Data Model * DBPedia, YAGO, Google Knowledge Graph, Wikidata | R3 Ch1, Ch2 |
| 34 | Understand how knowledge bases can improve existing retrieval models | Entity-based Retrieval Models | R4 |
| 35-37 | Formulate Google’s Page Ranks algorithm | Link Analysis   * The web as a graph * Google’s page rank * Hub and Authorities (HITS) | R2 Ch3, T Ch21 |
| 38-40 | Understand ethical issues related to Information Retrieval | Responsible IR | R7 |

**5. Evaluation Scheme**

**5.a Major Components**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Component** | **Duration** | **Weightage** | **Date&Time** | **Mode** |
| Two Programming Assignments | Take Home | **40%** | TBA | Open Book |
| Mid-Term exam | 90 mins | **25%** | 11/03 - 9.30 - 11.00AM | Closed Book |
| Comprehensive exam | 3 hours | **35%** | 06/05 FN | Closed Book |

**\*Note:** 40% of the evaluation will be completed by mid semester grading

**6. Chamber Consultation:** TBA

**7. Notices:** All notices related to the course will be displayed on the **CMS**.

**8. Make-up Policy:**

Make-ups for Mid Sem and Comprehensive examination tests shall be granted by the I/C on prior permission and only to genuine cases in case of hospitalization. Permission will be granted only if the candidate has applied makeup for all other registered courses.

**9.Academic Honesty and Integrity Policy:**Academic honesty and integrity are to be maintained by all the students throughout the semester and no type of academic dishonesty is acceptable.

**Instructor-in-charge**

**CS F469**