Course Title: Information Retrieval (IR)

Course Code: IT550

Course Instructor: Dr. Parth Mehta, Parmonic AI

Credit Structure (L-T-P-Cr): (3-0-2-4) [three hrs lecture and two hrs Lab per week]

Course Placement: M.TECH (ICT) III ML Elective. Technical elective for B.Tech VII (ICT and CS, CPI criterion 7.4 for enrollment)

Prerequisites/desired skill set: Understanding of Linear algebra, basic probability and statistics, basic algorithms and data structures, Python programming. I

A good Information Retrieval system is as much about good engineering as about the underlying search algorithms, perhaps even more. As such the course will be programming-heavy and the participants are expected to have fair programming skills.

Course objective: This course focused on designing efficient Information Retrieval systems. The first half of this course focuses on the traditional Information Retrieval system which includes preprocessing, indexing, ranking, retrieval and evaluation. However, search has evolved a lot in the post-LLM era, replacing conventional search engines with conversational agents. The second half of the course focuses on Neural Information Retrieval, prompting, semantic search, word embeddings, vector retrieval and Retrieval Augmented generation. Participants of the course should expect a fast-paced course, with a lot of hands-on experimentation.

The course involves a course project, which is a major component of the final evaluation. Participants are expected to choose their desired topic early during the course and work on it throughout the semester. Periodic evaluation will be carried out.

Course content:

Module 1: Basics of IR.

- Handling unstructured text
- Preprocessing pipeline - Tokenization, stopword removal, stemming, lemmatization
- Inverted Indexing - posting lists
- Boolean Retrieval

Module 2: Term weighting and Ranking

- Term Weighting - TF-IDF (term frequency/inverse document frequency) weighting
- Text-similarity metrics - word overlap; cosine similarity
- Ranked retrieval: Vector-space retrieval models, Okapi BM25

Module 3: Evaluation

- Benchmark text collections
- Performance metrics: recall, precision, and F-measure
- Advanced metric: NDCG, MAP, MRR

Module 4: Query Operations

- Relevance feedback, Pseudo Relevance feedback, Rocchio Algorithm
- Query expansion

Module 5: Probabilistic IR

- The binary independence model
- Probabilistic Language modelling - Query likelihod, Heimstras Language model
- The query likelihood model

Module 6: Text Classification and Clustering [4 Lecture converted to two lab sessions]

- Text Classification - Naive Bayes text classification, KNN Classification
- Clustering - Flat clustering, Hierarchical clustering,

Module 7: Web search

- Web Search engine
- Crawling and Link Analysis
- Page Rank Algorithm

Module 8: Distributed word representations for IR

- Latent Semantic Indexing
- Word Embeddings - Word2Vec
- Transformer Model

Module 9: Neural Information Retrieval

- Learning to rank,
- Neural Language Model
- Deep neural methods for rankings

Module 10: Search in Post LLM Era

- The Query - Prompting equivalence, Prompting Techniques
- Vector Retrieval

- Retrieval Augmented Generation (RAG)

Module 11:Domain Specific Applications [tentative]

- Summarization
- Legal Information Access
- Noisy Retrieval

Lab Only Topics

- Distributed IR - search in real life
- Optimizing IR systems
- Elastic Search and Kibana

Course Outcome: At the end of the course, students will understand the basic concepts like indexing,query processing, similarity measures, and ranking that are important to the design of the search system. They are able to evaluate the performance of Information retrieval systems using a set of metrics. This course will help them learn about different text representation schemes that are used to transform the unstructured text into a lower-dimensional vector space. Students are able to use state-of-the-art machine learning methods in the Information Retrieval field.

Assessment/ Evaluation:

- 2 Exams: One In-sem and Final End-Sem examination
- 8-10 Lab Assignments and course project
- Grading scheme is relative
- In-Sem -20%; End-Sem-20%, Lab Assignments-20%, Course project-40%

Suggested textbooks/references

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press, 2008. ISBN-13: 978-0521865715
2. Bhaskar Mitra and Nick Craswell, An Introduction to Neural Information Retrieval, Now publishers Inc
3. Jure Leskovec, Anand Rajaraman , Jeffrey D. Ullman. Mining of Massive Datasets, Cambridge University Press, 2011. ISBN: 978-1107077232.