NATIONAL UNIVERSITY
of Computer & Emerging Sciences, Lahore

Department of Computer Science

CS-4051 Information Retrieval
Spring 2024

Instructor Name: Dr. Asma Naseer **TA Name:**
Email address: asma.naseer@nu.edu.pk **Email address:** @lhr.nu.edu.pk
Office Location/Number: New Building - 11
Office Hours: **Thu 10 pm – 12 pm**

Course Information

Program: BS (CS) **Credit Hours:** 3 **Course Type:** Elective

Pre-requisites (if any): Core Programming and Algorithm skills, Probability and Statistics

Class Meeting Time: (Mon & Wed) (8:30 am to 10:00 AM – 10:00 am 11:30 am)
Class Venue: CS-5

Course Description/Objectives/Goals:

This course provides broad coverage of the important issues in information retrieval. It is designed to help you to understand how search engines work, how to build your own search engine, evaluate its performance, and modify it for specific applications. A number of advanced topics will be covered to address more recent developments in IR such as collaborative filtering and Topic Modeling. Students will furthermore acquire practical experience in the construction of IR systems by a series of programming assignments. Mathematical experience including basic probability is strongly desirable.

Course Learning Outcomes (CLOs):

At the end of the course students will be able to:	**Domain**	**BT* Level**
understand the common algorithms and techniques for information retrieval (document indexing and retrieval, query processing, etc.)	C	2
use the quantitative evaluation methods for the IR systems and data mining techniques	C	6
implement a basic textual information retrieval system using Java or Python	C	3
Learn the popular probabilistic retrieval methods and ranking principles	C	3
apply information retrieval techniques to the problems of text clustering, recommendation systems, text classification etc.	C	3
understand the common algorithms and techniques for information retrieval (document indexing and retrieval, query processing, etc.)	A	5
use the quantitative evaluation methods for the IR systems and data mining techniques	P	6
* BT= Bloom's Taxonomy, C=Cognitive domain, P=Psychomotor domain, A= Affective domain **Bloom's taxonomy Levels:** 1. Knowledge, 2. Comprehension, 3. Application, 4. Analysis, 5. Synthesis, 6. Evaluation		

Course Textbook

[MRS] Introduction to Information Retrieval by Manning, Raghavan, and Schütze - available free online.

Additional references and books related to the course:

- (MG) Managing Gigabytes, by I. Witten, A. Moffat, and T. Bell.
- (IRAH) Information Retrieval: Algorithms and Heuristics, by D. Grossman and O. Frieder.
- (MIR) Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto.
- (FSNLP) Foundations of Statistical Natural Language Processing, by C. Manning and H. Schütze.
- (SE) Search Engines: Information Retrieval in Practice, by B. Croft, D. Metzler, and T. Strohman.
- (IRIE) Information Retrieval: Implementing and Evaluating Search Engines, by S. Büttcher, C. Clarke, and G. Cormack.
- (TDMA) Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, C. Zhai and S. Massung, ACM Book Series, Morgan & Claypool Publishers, 2016.

Tentative Weekly Schedule

Week	Topics	Text Book Sections
1	Key problems, Information need, Queries and documents, Matching scores	[MRS] Chapter 1
2	**Text Preprocessing** Tokenization Stopping, stemming	[MRS] Chapter 2 Section 2.2
3	**Inverted Index Construction (**Posting Lists, Dictionary, Distributed indexing, dynamic indexing) ● Term frequency (TF) ● Document frequency (DF) and inverse document frequency (IDF) ● TF transformation ● BM25 ● Inverted index and postings ● Binary coding, unary coding, gamma-coding, and d-gap *Recommended Readings:* ⇒ C. Zhai and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapter 6 - Section 6.3, and Chapter 8**. ⇒ Ian H. Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition. Morgan Kaufmann, 1999.	[MRS] Chapter 4 (from 4.2 till 4.5)
4	**Zipf's Law, Heap's Law, Index Compression**	[MRS] Chapter 5 Section 5.1 and 5.3
5	**Retrieval Models (Vector Space Models) (**Vector-space model, Cosine Similarity, Tf-Idf, BM25) ● Query likelihood ● Statistical and unigram language models ● Maximum likelihood estimate ● Background, collection, and document language models ● Smoothing of unigram language models	[MRS] Chapter 6 Section 6.2 and 6.3

	● Relation between query likelihood and TF-IDF weighting ● Linear interpolation smoothing ● Feedback in the Vector Space Model ● Feedback in Language Model *Recommended Readings:* ● C. Zhai and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapter 6 - Section 6.4**	
6	**Word Vectors (Word Embeddings)**	
7	**Retrieval Models (Language Models)** Smoothing Methods	[MRS] Chapter 12 12.1 to 12.3
8	**IR Evaluation/ Measures (**Ranking measures: R-prec, Mean Average Precision, nDCG, Reciprocal Rank) ● Evaluation methodology ● Precision and recall ● Average precision, mean average precision (MAP), and geometric mean average precision (gMAP) ● Reciprocal rank and mean reciprocal rank ● F-measure ● Normalized Discounted Cumulative Gain (nDCG) ● Statistical significance test *Recommended Readings:* ⇒ Mark Sanderson. Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval 4, 4 (2010), 247- 375. ⇒ C. Zhai and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapter 9**	[MRS] Chapter 8 Section 8.1 to 8.4 (Interpolated precision is not included)
9	**Web Retrieval (**Link analysis, Markov Chains, PageRank) ● Relevance feedback ● Pseudo-relevance feedback ● Implicit feedback ● Rocchio feedback ● Scalability and efficiency ● Spams ● Crawler, focused crawling, and incremental crawling ● Google File System (GFS) ● MapReduce ● Link analysis and anchor text	[MRS] Chapter 21 21.1, 21.2.1, 21.2.2 [CMS]4.5.2

	● PageRank and HITS *Recommended Readings:* ⇒ C. Zhai and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapters 7 & 10**	
10	**Logistic Regression** ● Learning to rank, features, and logistic regression ● Content-based filtering ● Collaborative filtering ● Beta-Gamma threshold learning ● User profile ● Exploration-exploitation tradeoff	
11	**Document Clustering (**K-means clustering, Evaluation of clustering) ● Overview of Clustering Techniques ● Document Clustering ● Term Clustering ● Evaluation of Text Clustering	[MRS] Chapter 16 Section 16.3, 16.4
12	**Clustering (**Hierarchal Agglomerative Clustering)	[MRS] Chapter 17 Section 17.1 till 17.4
13	**Text Classification (**Naive Bayes, Sentiment Analysis) ● Overview of Text Categorization/Classification Methods ● Text Categorization Problem ● Features for Text Categorization ● Classification Algorithms ● Evaluation of Text Categorization	[MRS] Chapter 13 Section 13.1 till 13.3
14	**Project Presentations**	

(Tentative) Grading Criteria:

1. Assignments + Class Exercises (10%)
2. Quizzes (10%)
3. Project and research paper (10+5%)
4. Midterm Exam (25%)
5. Final Exam (40%)

- Grading scheme for this course is **Absolute** under application of CS department's grading policies.
- Minimum requirement to pass this course is to obtain at least **50%** absolute marks

Course Policies:

- All assignments and homework must be done individually, until specified as a group task.
- Quizzes will be announced.
- No makeup for missed quiz or assignment.
- Late Submissions of assignments will not be accepted.
- Minimum 80% attendance is required for appearing in the Final exams.

Plagiarism in Assignments

You are not allowed to copy code for programming assignments from internet or any other student. Penalty of plagiarism in programming assignments will be from one of the following depending on severity of case:

- -1 absolute from final grade
- Final grade is lowered
- F in course