NATIONAL UNIVERSITY
of Computer & Emerging Sciences, Lahore

Department of Computer Science

Information Retrieval

SPRING 2022

Instructor Name: Dr. Iqra Safder

TA Name (if any): **Muhammad Adeel** l217258@lhr.nu.edu.pk (Section A)

Haris Ali FastNU l181247@lhr.nu.edu.pk (Section B)

Email address: iqra.safder@nu.edu.pk **Email address:**
Office Location/Number: C-135 **Office Location/Number:**
Office Hours: Tuesday 1pm-2pm**,** Friday 10am – 11:30 am, Saturday 10am - 11:30am,

Course Information

Program: BSCS **Credit Hours:** 3 **Type:** CS Elective

Pre-requisites (if any): Programming competence, Data structures

Course Website (if any) : Google classroom
Class Meeting Time:

Course Description/Objectives/Goals:

Recent years have seen a dramatic growth of natural language text data, including web pages, news articles, scientific literature, emails, enterprise documents, and social media such as blog articles, forum posts, product reviews, and tweets. Text data are unique in that they are usually generated directly by humans rather than a computer system or sensors, and are thus especially valuable for discovering knowledge about people's opinions and preferences, in addition to many other kinds of knowledge that we encode in text.

This course will cover technologies, which play an important role in any data mining applications involving text data for two reasons. First, while the raw data may be large for any particular problem, it is often a relatively small subset of the data that are relevant, and a search engine is an essential tool for quickly discovering a small subset of relevant text data in a large text collection. Second, search engines are needed to help analysts interpret any patterns discovered in the data by allowing them to examine the relevant original text data to make sense of any discovered pattern. You will learn the basic concepts, principles, and the major techniques in text retrieval, which is the underlying science of search engines.

Course Learning Outcomes (CLOs):		
At the end of the course students will be able to:	**Domain**	**BT* Level**
Explain many basic concepts and multiple major algorithms in text retrieval and search engines.		C2
Explain how search engines and recommender systems work and how to quantitatively evaluate a search engine.		C2
Create a test collection, run text retrieval experiments, and experiment with ideas for improving a search engine.		C3
* 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		

Textbook(s) /Supplementary Readings:

The following book will be used as a primary text to guide some of the discussions, but it will be heavily supplemented with lecture notes and reading assignments from other sources.

- 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.

Additional references and books related to the course:

- 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 10 - Section 10.4, Chapter 11.

Tentative Weekly Schedule (Phase wise)

Phase	Topics to be covered	Readings	Assignments /Projects?
1	• Part of Speech tagging, syntactic analysis, semantic analysis, and ambiguity • "Bag of words" representation • Push, pull, querying, browsing • Probability ranking principle • Relevance • Vector space model • Dot product • Bit vector *Recommended Readings:* • N. J. Belkin and W. B. Croft. 1992. Information filtering and information retrieval: Two sides of the same coin? Commun. ACM 35, 12 (Dec. 1992), 29-38. • 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 1 - 6**.		
2	• 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.		
3	• 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**		
4	• p(R=1|q,d), query likelihood, and p(q|d) • Statistical and unigram language models • Maximum likelihood estimate • Background, collection, and document language models • Smoothing of unigram language models • Relation between query likelihood and TF-IDF weighting • Linear interpolation smoothing		

	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**		
5	• 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 • 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**		
6	• Learning to rank, features, and logistic regression • Content-based filtering • Collaborative filtering • Beta-Gamma threshold learning • User profile • Exploration-exploitation tradeoff		

(Tentative) Grading Criteria

Quizzes	10%
Assignments/Homeworks/Project	*15* - 25%
Midterms	25-30%
Final Exam	40 - 45%
Total:	**100 %**

Course Policies

- *Course outline may change 10-20% as we proceed in the semester. We may add and remove a few topics.*
- ***Grading scheme: Relative***
- Depending on the situation of COVID 19, this weightage of midterms can be reduced and added in assignments/homeworks/project.
- *Weightage of other evaluations can also be adjusted if needed.*

- *Assignment deadlines for assignment and Project are hard.*
- *NO Cell Phone usage in class, they must be turned off at all times.*
- *There will be no retake of quizzes or exams.*
- ***Integrity in the assignments/quizzes is expected; otherwise result would be an F grade in the course or may be the case is forwarded to Disciplinary committee.***
- *Attendance MUST be ensured according to the University policy to avoid disqualification.*