****

Course Title: Information Retrieval and Text Mining

Course Code: IS-669

Department: **Information Systems**

**HSM Vision**

HSM envisions its success in the sustainable contribution that it will make to the industry, academia and research in public and private sector. HSM will lead by providing professionally competent and ethically conscious human resources engaged in the global and local context to foster socio-economic growth and sustainability for the society. HSM envisages having faculty with high research potential and a deep desire for cutting edge research including collaboration with national and international partners.

**HSM Mission**

Being a research-oriented and student-centric business school, we emphasize research publications in impact journals as well as state-of -the-art learning methodologies.  We will prepare our students to become the future ethical business leaders and the guiding post for the society, while equipping them with the knowledge and skills required by world-class professionals.  We will be the leading choice for organizations seeking highly talented human resource. HSM will foster internationalization with key stakeholders and actively work to exchange best practices with business schools across Pakistan through collaborations, workshops, conferences and other means.

**Program Objectives**

The School of Business and Economics at UMT is foreseeing the challenges ahead both at national and international level and the utility of data science.  In Pakistan the multi-dimensional economy integrated with globalization needs a boost assisted by professionally trained and skilled Data Scientists, whom may incorporate and harmonize the unlimited bucket of resources, pouring in from springs of industry, agriculture, business, and human resources etc. in a manner to achieve efficiency to its apex.
In the competitive economy the companies need to adapt data science to gain a competitive advantage in productivity, profitability and sustainable business processes to offer better products and services to their customers. To attain this goal trained and skilled workforce in this area is the need of the hour; who are equipped to manage, understand and model the data, interpret the outcome and communicate the results for business use. Professionals holding a degree in Data Science will be well positioned to help their organizations gain a competitive advantage in a data-driven world.

**Course Objectives**

This course provides a broad introduction to the objective of this course is to learn the Information Retrieval Techniques. The aim is to study the theory, design, and implementation of text-based information system.

Basic and advanced techniques for text-based information systems: efficient text indexing; Boolean and vector space retrieval models; evaluation and interface issues; Web search including crawling, link-based algorithms, and Web metadata; text/Web Probabilistic Model: clustering, classification; text mining; Search Computing, Ranking in Web Search, Diversification, Personalization; Recommender Systems.

A variety of current research topics are also covered, including cross-lingual retrieval, document summarization, topic detection and tracking, and multi-media retrieval. Ranking and diversification in search; text summarization

**Learning Objectives**

* Understanding the general concepts of Information Retrieval and its different terminologies.
* Discuss IR models
* Understanding of technical issues related to Information Retrieval
* Understanding Web search including crawling, link-based algorithms, and Web metadata in IR.
* Detail understanding on advanced level topics of Information Retrieval.

**Pre-requisites:**

* Knowledge of basic computer science principles and skills, at a level sufficient to
* Write a reasonably algorithms.
* Familiarity with the pseudo code and Indexing.

**Teaching Methodology (List methodologies used –example are given below)**

* Lecture
* Interactive Classes
* Case based teaching
* Class activities
* Applied Projects
* Guest Lectures

**STUDENTS ARE REQUIRED TO READ AND UNDERSTAND ALL ITEMS OUTLINED IN THE PARTICIPANT HANDBOOK**

**Class Policy:-**

* Be On Time

You need to be at class at the assigned time. After 10 minutes past the assigned time, you will be marked absent.

* Mobile Policy

**TURN OFF YOUR MOBILE PHONE!**It is unprofessional to be texting or otherwise.

* Email Policy

**READ YOUR EMAILS!** You are responsible if you miss a deadline because you did not read your email.

Participants should regularly check their university emails accounts regularly and respond accordingly.

* Class Attendance Policy

A minimum of 80% attendance is required for a participant to be eligible to sit in the final examination. Being sick and going to wedding sare absences and will not be counted as present. You have the opportunity to use 6 absences out of 30 classes. Participants with less than 80% of attendance in a course will be given grade ‘F’ (Fail) and will not be allowed to take end term exams. International students who will be leaving for visa during semester should not use any days off except for visa trip. Otherwise they could reach short attendance.

* Withdraw Policy

Students may withdraw from a course till the end of the 12th week of the semester. Consequently, grade W will be awarded to the student which shall have no impact on the calculation of the GPA of the student. A Student withdrawing after the 12th week shall be automatically awarded “F” grade which shall count in the GPA.

* Moodle

UMT –LMS (Moodle) is an Open Source Course Management System (CMS), also known as a learning Management System (LMS). Participants should regularly visit the course website on MOODLE Course Management system, and fully benefit from its capabilities. If you are facing any problem using moodle, visit <http://oit.umt.edu.pk/moodle>. For further query send your queries to moodle@umt.edu.pk

* Harassment Policy

Sexual or any other harassment is prohibited and is constituted as punishable offence. Sexual or any other harassment of any participant will not be tolerated. All actions categorized as sexual or any other harassment when done physically or verbally would also be considered as sexual harassment when done using electronic media such as computers, mobiles, internet, emails etc.

* Use of Unfair Means/Honesty Policy

Any participant found using unfair means or assisting another participant during a class test/quiz, assignments or examination would be liable to disciplinary action.

* Plagiarism Policy

All students are required to attach a “Turnitin” report on every assignment, big or small. Any student who attempts to bypass “Turnitin” will receive “F” grade which will count towards the CGPA. The participants submit the plagiarism report to the resource person with every assignment, report, project, thesis etc. If student attempts to cheat “Turnitin”, he/she will receive a second “F” that will count towards the CGPA. There are special rules on plagiarism for final reports etc. all outlined in your handbook.
* Communication of Results

The results of quizzes, midterms and assignments are communicated to the participants during the semester and answer books are returned to them. It is the responsibility of the course instructor to keep the participants informed about his/her progress during the semester. The course instructor will inform a participant at least one week before the final examination related to his or her performance in the course.

**Course Outline**

|  |  |
| --- | --- |
| Program | MS Data Sciences |
| Credit Hours | 3 |
| Duration | 15 Weeks |
| Prerequisites (If any) | N/A |
| Resource PersonName and Email |
| Counseling Timing |  |
| Contact no. |  |
| Web Links:-(Face book, Linked In, Google Groups, Other platforms) |  |

**Chairman/Director Program signature………………………………….Date……………………..**

**Dean’s signature………………………… ………………….Date………………………………………….**

**Grade Evaluation Criteria**

Following is the criteria for the distribution of marks to evaluate final grade in a semester.

**Marks Evaluation Marks in percentage**

Quizzes 15%

Assignments 15%

Mid Term 25%

Attendance & Class Participation 5%

Term Project and Presentation 40%

Total 100%

**Reference Books:**

* Text Book: Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press, 2008.
* MG = Managing Gigabytes, by I. Witten, A. Moffat, and T. Bell. IRAH = Information Retrieval: Algorithms and Heuristics by D. Grossman and O. Frieder. IR = Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto. FOA = Finding Out About, by R. Belew. MTW = Mining the Web, by S. Chakrabarti. FSNLP = Foundations of Statistical Natural Language Processing, by C. Manning and H. Schütze.
* These books all have useful information on topics that we cover and are recommended as references. MG is particularly good as a detailed reference for technical IR in the first half of the course. MTW covers many of the topics from the latter part of the course.

**Course: -** **Information Retrieval Course code: Book:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **Topics to be****covered in the course** | **Learning Objective****of this topic** | **Expected Outcomes from Students** | **Teaching Method** | **Assessment Criteria** | **Deadlines and Homework** |
| 1 | Introduction to Information Retrieval. | Basic Concept of Information Retrieval and IR Models | Students will be able to understand the basic terminology Information Retrieval | Lecture | Class Participation | Within a Week |
| 2 | Inverted indices and Boolean queries | Understanding the concept of Inverted indices and Boolean queries and extended Boolean queries | Students will be able to understand Boolean and extended Boolean Query  | Case Discussion | Class activity | Within a Week |
| 3 | Optimization | Understanding the nature of unstructured and semi-structured text. | They will be in position to understand the key nature of unstructured and semi-structured text. | Lecture | Assignment | Within a Week |
| 4 | The term vocabulary and postings lists and Text encoding. | Learn the basics term vocabulary and postings lists and Text encoding: tokenization, stemming, lemmatization, stop words, phrases | After this lecture, students will be able to determine vocabulary and Text encoding. | Lecture | Quiz | Within a Week |
| 5 | Optimizing indices with skip lists. Proximity and phrase queries. Positional indices. | Working on indices using proximity and phrase queries. | They will be able to work on Optimizing indices. | Case Discussion + Lecture | Class Activity | Within a Week |
| 6 | Index construction.  | Understand the Postings size estimation, sort-based indexing, dynamic indexing, positional indexes, n-gram indexes, distributed indexing, real-world issues. | Students will understand Index construction, it’s types and real world issues  | Lecture  | Quiz | Within two Weeks |
| 7 | Index compression | Understand the Index compression that includes lexicon compression and postings lists compression. Gap encoding, gamma codes, Zipf's Law, variable-byte encoding. Blocking. Extreme compression. | Students will learn Index compression and its different methods | Lecture  | Assignment  | Within two Weeks |
| 8 | Mid Term | Mid Term | Mid Term | Mid Term | Mid Term | Mid Term |
| 9 | Dictionaries and tolerant retrieval. Wild-card queries, permuterm indices, n-gram indices. | Understanding the Dictionaries and tolerant retrieval. Dictionary data structures, Wild-card queries, permuterm indices, n-gram indices.  | After this lecture, students have understood Dictionaries and tolerant retrieval. Dictionary data structures, Wild-card queries, permuterm indices, n-gram indices. | Case Discussion | Assignment | Within a Week |
| 10 | Spelling correction and synonyms:  | How to perform Spelling correction and synonyms: edit distance, soundex, language detection. | Now they will be able to perform Spelling correction and synonyms | Lecture + Case Study | Class Activity | Within a Week |
| 11 | Scoring, term weighting, and the vector space model. Parametric or fielded search. Document zones. | Understanding Scoring, term weighting, and the vector space model. Parametric or fielded search. Document zones. | After this lecture, they are able to do Scoring, term weighting, and the vector space model. Parametric or fielded search. Document zones. | Lecture | Quiz | Within a Week |
| 12 | The vector space retrieval model. | Understanding The vector space retrieval model. tf.idf weighting. The cosine measure. Scoring documents. | Their understanding with the vector space retrieval model. tf.idf weighting. The cosine measure. Scoring documents | Lecture + Case Discussion | Assignment | Within two Weeks |
| 13 | Computing scores in a complete search system | How to perform various tasks on Computing scores in a complete search system: Components of an IR system. Efficient vector space scoring. Nearest neighbor techniques, reduced dimensionality approximations, random projection. | They will be able to understand Computing scores in a complete search system. | Lecture | Quiz | Within two Weeks |
| 14 | K Nearest Neighbors, Decision boundaries, Vector space classification using centroids. Comparative results | How to use K Nearest Neighbors, Decision boundaries, Vector space classification using centroids. Comparative results. | Now they will be in position to understand K Nearest Neighbors, Decision boundaries, Vector space classification using centroids. Comparative results | Lecture | Class Activity | Within a Week |
| 15 | Latent semantic indexing (LSI). Applications to clustering and to information retrieval. | Working on how to use Latent semantic indexing (LSI). Applications to clustering and to information retrieval. | Students will learn how to use Latent semantic indexing (LSI). Applications to clustering and to information retrieval. | Lecture + Case Discussion | Class Participation | Within a Week |