****

**INSTRUCTION DIVISION**

**FIRST SEMESTER 2017-2018**

# Course Handout Part II

01-08-2017

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. Aruna Malapati (*[*arunam@hyderabad.bits-pilani.ac.in*](mailto:arunam@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 (Boolean, vector space, probabilistic, inference net, language modeling), collaborative filtering, Language translation and Multimedia information retrieval.

The student should be able to

* Design and implement Boolean and Vector space models for searching text documents.
* Analyze the effect of different scoring and ranking schemes for text search engines.
* Apply IBM models for language translation
* Implement recommender systems using Singular Value, CUR Decomposition and latent factor models
* Apply Google’s Page rank algorithm given a web graph.
* Compare the text retrieval techniques with Image, Video and Audio retrieval.

2. Pre requisites: Programming in Java or C, and knowledge of core data structures and algorithms.

##### 3.a. Text Book

* **T1**. C. D. Manning, P. Raghavan and H. Schutze. Introduction to Information Retrieval, Cambridge University Press, 2008.

**3.b. Reference Books**

* **R1:** Modern Information Retrieval, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley, 2000.
* **R2:** Search Engines: Information Retrieval in Practice by Bruce Croft, Donald Metzler, and Trevor Strohman, Addison-Wesley, 2009.

# R3: Cross-Language Information Retrieval by By Jian-Yun Nie Morgan & Claypool Publisher series 2010.

# R4: Multimedia Information Retrieval by Stefan M. Rüger Morgan & Claypool Publisher series 2010.

* **R5** Information Retrieval: Implementing and Evaluating Search Engines by S. Buttcher, C. Clarke and G. Cormack, MIT Press, 2010.
* **R6:** Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data by B. Liu, Springer, Second Edition, 2011.
* R7:Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P.B. (Eds.), Recommender Systems Handbook. 1st Edition., 2011, 845 p. 20 illus., Hardcover, ISBN: 978-0-387-85819-7

**4. Course Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| Lecture No | Learning Outcomes | Topics to be covered | Chapter in the Text Book |
| 1 | * List the course objectives and define the vocabulary used in IR | Introduction to the course | T1 Ch1 |
| 2,3 | The term vocabulary postings lists | T1 Ch 1 & 2,R1 Ch2 section 5 |
| 4 | * Define different query mechanisms | Introduction to ad-hoc search |
| 5 | * Evaluate and apply wild card queries and spelling correction | Wildcard queries  Spelling correction | T1 Ch 3 |
| 6 | * Apply edit distances and Phonetic based spelling correction | Edit distances  Phonetic correction |
| 7 | * Apply blocked sort indexing | Blocked sort-based indexing | T1 Ch 3 |
| 8 | * Evaluate the use of Single-pass in memory indexing | Single-pass in-memory indexing |
| 9 | * Evaluate the use of Distributed and Dynamic in memory indexing | Distributed indexing  Dynamic indexing |
| 10-11 | * Apply the TF-IDF for evaluating the query | Weighted zone scoring  Learning weights  Term frequency and weighting  Tf-idf weighting | T1 Ch 6 |
| 12 | * Evaluate the variant for TF-IDF evaluating the query | Dot products,  Queries as vectors,  Variant tf-idf functions, | T1 Ch 6 |
| 13 | Document and query weighting schemes |
| 14 | * List the challenges involved in IR with non-English queries | European Languages | R3 Ch 1 |
| East Asian Languages |
| Other Languages |
| 15 | * List and define the terms used to implement CLIR | Translation Approaches for CLIR | R3 Ch2 |
| 16 | * Apply the IBM models for language translation | Handling many Languages using IBM models |
| 17 | Using manually constructed Translation systems and resources for CLIR |
| 18 | * Define the terms used in multimedia queries | Basic Multimedia search technologies | R4 Ch2,3 |
| 19,20 | * Evaluate techniques used in Multimedia IR | Content based retrieval |
| 21,22 | Image and Audio data challenges |
| 23 | Multimedia IR Research |
| 24 | * Define the problem of recommender system | Introduction to recommendation system | R7  Ch1,2,3,4,5  Class notes |
| 25 | * Evaluate and apply different algorithms for recommender systems | Collaborative , Content based recommendation |
| 26 | Hybrid recommendation systems |
| 27,28  29  30  31 | * Evaluate the search engine architecture | Search Engine Architecture  Web characteristics | T1 Ch 19  R1 Ch13, R2 Ch2 |
| 33-38 | * Apply and evaluate the page rank and its variants | The Web as a graph, Google’s Pagerank | T1 Ch 21 and class notes |
| 38,40 | Hub and authorities (HITS), Web spam, SEO |

##### 5. Evaluation Scheme

**5.a Major Components**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Component** | **Duration** | **Weightage** | Date&Time | **Mode** |
| Project |  | **20%** | To be announced | Open Book /  Take Home |
| Quiz (2) |  | **15%** | To be announced | Closed Book |
| Mid-Term exam | 90 mins | **25%** | 13/10/2017  1.30 -- 3.00 PM | Closed Book |
| Comprehensive | 3 hours | **40%** | 12/12/2017 FN | Closed Book |

**5.b. Timeliness**

* Assignments are to be completed in time with no postponements.
* Submissions 24 hours from deadline will have a penalty of 2 Marks per day.

**6. Chamber Consultation:** To be announced.

**7. Notices:** All notices related to the course will be displayed on the **CSIS Notice Board**, and / or CMS.

**8. Make-up Policy:**

Make ups for Mid sem test shall be granted by the I/C on prior permission and only to genuine cases with the permission of the warden concerned.

No Makeups will be granted for Quizzes.

Make-up for comprehensive examination will be decided and scheduled by the Instruction Division.

**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**