Birla Institute of Technology & Science, Pilani
Hyderabad Campus

FIRST SEMESTER 2020-2021
Course Handout Part II

17-08-2020

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)

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 Google's Page rank algorithm given a web graph.
- Apply IBM models for language translation
- Implement recommender systems using Singular Value, CUR Decomposition and latent factor models
- Compare the text retrieval techniques with Image, Video and Audio retrieval.

2. Pre requisites: Programming in Java or C however programming in python will be an advantage, 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. The entire book is available at http://nlp.stanford.edu/IR-book/

3.b. Reference Books

- **R1:** Modern Information Retrieval, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley, 2000. http://people.ischool.berkeley.edu/~hearst/irbook/
- **R2:** Statistical Machine Translation, Philipp Koehn, CAMBRIDGE UNIVERSITY PRESS,2010
- **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**: Mining of Massive Datasets,Jure Leskovec,Anand Rajaraman,Jeffrey D. Ullman, Cambridge University Press

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
2-4		Inverted Index constructions and merge algorithm, IR Pipeline, Skip Lists, Phrase queries	T1 Ch 1 & 2,R1 Ch2 section 5
5	• Evaluate and apply wild card queries and spelling correction	Dictionary data structures, Wildcard queries	T1 Ch 3
6	• Evaluate and apply different spelling correction techniques	Edit distances, Soundex algorithm, N-gram overlap, Context-sensitive correction	T1 Ch 3
7-9	• Apply tf-idf and cosine score to score documents against a query	Jaccard score, TF-IDF and its variants for ranked retrieval	T1 Ch 6
10-12	• Formulate Google's Page Ranks algorithm	Page Rank, Teleportation, Topic Specific Page rank, Spam, Hub and authorities (HITS), Web spam, web farms	T1 Ch 21
13-14	• Formulate the search as near duplicate detection	Latent Semantic Analysis	T1 Ch 18 Topic 18.4
		Locality sensitive hashing	
15	• Compare different metrics for evaluating search engines	Precision, Accuracy, Recall, Mean Average Precision, Precision and Recall in ranked retrieval	T1 Ch 8
16-24	• Compare and evaluate models for recommender systems	Recommender systems problem formulation and its solution using collaborative filtering, content based filtering, Singular Value Decomposition, CUR Decomposition and Latent Factor modeling	R6 Ch 9
25-29	• Formulate IR problem using Probabilistic approach and Near duplicates approach	Probabilistic model for IR	T1 Ch 11
30-38	Identify challenges in cross language IR and devise solutions using statistical machine translation	Cross Language IR, Statistical Machine Translations using word and Phrase based models	R2 Ch 4,5
39-42	• Define the terms used in multimedia queries • Compare the techniques for implementing multimedia IR	Basic Multimedia search technologies, Content based retrieval, Image and Audio data challenges	R4 Ch2,3

5. Evaluation Scheme

5.a Major Components

Component	Duration	Weightage	Date&Time	Mode
Test 1	30 mins	**12%**	September 10 – September 20 (during scheduled class Hour)	Open
Test 2	30 mins	**12%**	October 9-October 20(during scheduled class hour)	Open
Test 3	30 mins	**12%**	November 10-November 20 during scheduled class hour)	Open
Assignments (Min 3)	-	**30%**		Open
Comprehensive	2 hours	**34%**		Open

6. Chamber Consultation: Mon 4-5 PM via Google meet.

7. Notices: All notices related to the course will be displayed on the **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.

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

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