**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**WORK INTEGRATED LEARNING PROGRAMMES**

**Digital Learning**

**Part A: Course Design**

|  |  |
| --- | --- |
| **Course Title** | Information Retrieval |
| **Course No(s)** | SS ZG537 |
| **Credit Units** | 4 |
| **Credit Model** |  |
| **Content Authors** | Dr.V.Maheswari |

**Course Objectives**

|  |  |
| --- | --- |
| **No** | **Course Objectives** |
| **CO1** | To understand structure and organization of various components of an IR system |
| **CO2** | To understand information representation models, term scoring mechanisms, etc. in the complete search system |
| **CO3** | To understand architecture of search engines, crawlers and the web search |
| **CO4** | To understand cross lingual retrieval and multimedia information retrieval |

**Text Book(s)**

|  |  |
| --- | --- |
| T1 | C. D. Manning, P. Raghavan and H. Schutze. Introduction to Information Retrieval, Cambridge University Press, 2008. [http://nlp.stanford.edu/IR-book/http://nlp.stanford.edu/IR-book/](http://nlp.stanford.edu/IR-book/) |

**Reference Book(s) & other resources**

|  |  |
| --- | --- |
| R1 | Modern Information Retrieval, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley, 2000. [http://people.ischool.berkeley.edu/~hearst/irbook/http://people.ischool.berkeley.edu/~hearst/irbook/](http://people.ischool.berkeley.edu/~hearst/irbook/) |
| R2 | 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 |
| 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 | Search Engines: Information Retrieval in Practice by Bruce Croft, Donald Metzler, and Trevor Strohman, Addison-Wesley, 2009. |
| R8 | Koehn P., “Statistical Machine Translation”, Cambridge University Press, 2010. |

**Learning Outcomes:**

|  |  |
| --- | --- |
| No. | Learning Outcomes |
| LO1 | Students will gain understanding about an information retrieval system as a whole and about its components. |
| LO2 | Students will have knowledge about the design issues and their solutions of different type of models including Boolean, vector space etc. |
| LO3 | Students will have detailed understanding about text indexing, mining, weighting schemes etc. |
| LO4 | Students will acquire knowledge about cross lingual and multimedia information retrieval. |
| LO5 | With the acquired knowledge students will be able to design and build different kind of information retrieval systems. |

**Content Structure**

**Modular Content Structure**

|  |  |
| --- | --- |
| **No** | **Title of the Module** |
| M1 | Introduction, Boolean Retrieval |
| M2 | Boolean Retrieval, Term vocabulary and postings list |
| M3 | Postings list, Phrase queries, Dictionaries |
| M4 | Tolerant Retrieval, Index Construction |
| M5 | Index Construction, Vector Space Model |
| M6 | Vector Space Model, Text Classification |
| M7 | Vector space classification, Evaluation of classification |
| M8 | Text Clustering |
| M9 | Web Search, Web Crawler, Link Analysis |
| M10 | Multimedia IR, Cross-Lingual IR |

**Glossary of Terms:**

1. Contact Hour (CH) stands for a hour long live session with students conducted either in a physical classroom or enabled through technology. In this model of instruction, instructor led sessions will be for 20 CH.
   1. Pre CH = Self Learning done prior to a given contact hour
   2. During CH = Content to be discussed during the contact hour by the course instructor
   3. Post CH = Self Learning done post the contact hour
2. RL stands for Recorded Lecture or Recorded Lesson. It is presented to the student through an online portal. A given RL unfolds as a sequences of video segments interleaved with exercises
3. SS stands for Self-Study to be done as a study of relevant sections from textbooks and reference books. It could also include study of external resources.
4. LE stands 7or Lab Exercises
5. HW stands for Home Work will consists could be a selection of problems from the text.

**Part B: Contact Session Plan**

|  |  |
| --- | --- |
| **Academic Term** | Second Semester 2018-2019 |
| **Course Title** | Information Retrieval |
| **Course No** | SS ZG537 |
| **Content Developer** | Dr. Lavika Goel |

**M1: Introduction, Boolean Retrieval**

|  |  |
| --- | --- |
| Type | Description |
| RL 1.1 | RL 1.1.1 - Information vs. Data Retrieval  RL 1.1.2 – IR task |
| RL 1.2 | RL 1.2.1 - Basic Concepts  RL 1.2.2 - Logical view of the documents  RL 1.2.3 – The Retrieval process |
| RL 1.3 | RL 1.3.1 -- Classical IR models  RL 1.3.2 - Boolean model  RL 1.3.3 - Vector space model  RL 1.3.4 – Probabilistic model |
| RL 1.4 | RL 1.4.1 – Example of an IR problem  RL 1.4.2 – Term document Matrix  RL 1.4.3 -Boolean Queries, Inverted Index |
| CS 1.1 | CS 1.1.1 - Difference between Information and Data Retrieval, Information need, IR task  CS 1.1.2 – Relevance , Motivation of IR  CS 1.1.3 – Examples of different IR models |
| CS 1.2 | CS 1.2.1 – Example of Term document Matrix, Building Inverted Index  CS 1.2.2 – Processing Boolean queries – related examples  CS 1.2.3 - Precision P and recall R |
| LE1.1 |  |
| SS1.1 | R1- Ch1, Ch2, T1- Ch2 |
| HW1.1 | Exercises at the end of T1 - Ch2 |
| QZ1.1 |  |

**M2: Boolean Retrieval, Term vocabulary and postings list**

|  |  |
| --- | --- |
| Type | Description |
| RL 2.1 | RL 2.1.1 – Query Optimization  RL 2.1.2 - Term Vocabulary-Preprocessing |
| RL 2.2 | RL 2.2.1 – Vocabulary of terms- Tokenization, Stop Words  RL 2.2.2 – Vocabulary of terms- Normalization, Stemming and Lemmatization |
| CS 2.1 | CS 2.1.1 – Intersection of postings list and query optimization  CS 2.1.2 – Preprocessing- document delineation |
| CS 2.2 | CS 2.2.1 – Tokenization  CS 2.2.2 – Stop words, Normalization  CS 2.2.3 – Stemming and Lemmatization : Demonstration using the Porter’s Stemming Tool.  CS 2.2.4 – Python’s NLTK |
| LE 2.1 |  |
| SS 2.1 | T1 - Ch2 |
| HW 2.1, | Exercises at the end of T1- Ch2 |
| QZ 2.1 |  |

**M3: Postings list, Phrase queries, Dictionaries**

|  |  |
| --- | --- |
| Type | Description |
| RL 3.1. | RL 3.1.1 – Faster postings list  RL 3.1.2 - Postings lists intersection with skip pointers |
| RL 3.2 | RL 3.2.1 – Positional postings and phrase queries  RL 3.2.2 – Handling phrase queries |
| RL 3.3 | RL 3.3.1 – Dictionary data structure  RL 3.3.2 – Tolerant retrieval – wildcard queries |
| CS 3.1 | CS 3.1.1 – Faster postings list via skip pointers  CS 3.1.2 – Skip pointers example, placing skip pointers |
| CS 3.2 | CS 3.2.1 – Phrase queries – examples  CS 3.2.2 – Biword indexes – example, extended biwords, issues  CS 3.2.3 – Positional indexes |
| CS 3.3 | CS 3.3.1 - Dictionary data structure – Hash table, Tree  CS 3.3.2 – Tolerant Retrieval- Wildcard Queries  CS 3.3.3 – Wildcard query processing – permuterm indexes  CS 3.3.4 - Wildcard query processing - k-gram indexes |
| LE3.1 |  |
| SS3.1 | T1 - Ch2, Ch3 |
| HW3.1 | Exercises at the end of T1- Ch2, Ch3 |
| QZ3.1 |  |

**M4: Tolerant Retrieval**

|  |  |
| --- | --- |
| Type | Description |
| RL 4.1 | RL 4.1.1 – Spelling correction- isolated word correction  RL 4.1.2 - Edit distance (Levenshtein distance – computation) |
| RL 4.2 | RL 4.2.1 – Using edit distance, k-gram index  RL 4.2.2 – Context sensitive spell correction  RL 4.3.1 – Phonetic correction, Soundex |
| CS 4.1 | CS 4.1.1 – Isolated word correction – standard lexicon, lexicon of indexed corpus  CS 4.1.2 – Edit distance- example  CS 4.1.3 – Computation of Levenshtein distance |
| CS 4.2 | CS 4.2.1- Context sensitive spell correction  CS 4.2.2 - Phonetic correction, Soundex algorithm, computation of Soundex codes |
| LE4.1 |  |
| SS4.1 | T1- Ch3 |
| HW4.1 | Exercises at the end of T1- Ch3 |
| QZ4.1 | Compute Edit distance between “side” and “slide” |

**M5: Index Construction, Vector Space Model**

|  |  |
| --- | --- |
| Type | Description |
| RL 5.1 | RL 5.1.1 - Blocked sort-based Indexing  RL 5.1.2 - Single‐Pass In‐Memory Indexing |
| RL 5.2 | RL 5.2.1 – Distributed Indexing  RL 5.2.1 – Dynamic Indexing |
| RL 5.3 | RL 5.3.1 - Term frequency and weighting  RL 5.3.2 - The vector space model for scoring |
| CS 5.1 | CS 5.1.1 - BSBI, BSBI index construction algorithm  CS 5.1.2 – SPIMI algorithm, Merits and Demerits of both algorithms |
| CS 5.2 | CS 5.2.1 - Distributed Indexing, Examples of search engines using distributed indexing  CS 5.2.2 – Mapreduce architecture for distributed indexing, Parsers, Inverters, Schema for index construction in Mapreduce  CS 5.2.3 – Dynamic Indexing, Issues with main and auxiliary indexing, Logarithmic merge, Dynamic Indexing used in search engines  CS 5.2.4 - Index Compression |
| CS 5.3 | CS 5.3.1 – Boolean retrieval issues, Ranked retrieval models, query – document matching scores  CS 5.3.2 – Bag of words model., Term frequency-tf, document frequency-idf, effect of idf on ranking, tf-idf weighting |
| LE5.1 |  |
| SS5.1 | T1 - Ch4,Ch5, Ch6 |
| HW5.1 | Exercises at the end of T1 - Ch4, Ch5, Ch6 |
| QZ5.1 |  |

**M6: Vector Space Model, Text Classification**

|  |  |
| --- | --- |
| Type | Description |
| RL 6.1 | RL 6.1.1 – Documents and Queries as vectors, Similarity measure using angle, cosine  RL 6.1.2 – Cosine similarity for length normalized vectors  RL 6.1.3 - Computing Cosine scores, tf-idf weighting variants |
| RL 6.2 | RL 6.2.1 – Text classification problem  RL 6.2.2 – Classification methods |
| RL 6.3 | RL 6.3.1 – Bayesian method of text classification  RL 6.3.2 – Bernoulli Model  RL 6.3.3 – Time complexity of NB |
| CS 6.1 | CS 6.1.1 – Difference between binary, count and weight Matrix, representation of documents and queries as vectors, Similarity measure using angle, its disadvantage.  CS 6.1.2 – cosine similarity measurement between query and document, advantage of using length normalized vectors, , example of cosine similarity computation  CS 6.1.3 – Algorithm for computing cosine scores, Weighting in queries vs. documents- Examples. |
| CS 6.2 | CS 6.2.1 – Introduction to Text classification problem, examples of search engines using text classification, Supervised learning  CS 6.2.2 – Classification methods – manual, rule-based, statistical |
| CS 6.3 | CS 6.3.1 – Relevance feedback, Bayes rule, Bayes rule for text classification  CS 6.3.2 – NB classifier model, Conditional Independence assumption, Parameter estimation, Positional Independence assumption  CS 6.3.3 – The problem with maximum likelihood estimates- Zeros, Add-one smoothing, Naive Bayes -Training, Testing, Examples  CS 6.3.4 – Time complexity |
| LE6.1 |  |
| SS6.1 | T1 – Ch6,Ch13 |
| HW6.1 | Exercises at the end of T1 – Ch6, Ch13 |
| QZ6.1 |  |

**M7: Vector space classification, Evaluation of classification**

|  |  |
| --- | --- |
| Type | Description |
| RL 7.1. | RL 7.1.1 – Vector space representation of classes  RL 7.1.2 – Rocchio classification |
| RL 7.2 | RL 7.2.1 – k Nearest Neighbour Classification  RL 7.2.2 - Evaluation of classification |
| CS 7.1 | CS 7.1.1 - Vector Space Representation, Contiguity hypothesis, Classes in the Vector Space, Vector Space Classification Methods  CS 7.1.2 - Illustration of Rocchio Text Categorization, Definition of centroid, Rocchio algorithm, Example |
| CS 7.2 | CS 7.2.1 – Explanation about k Nearest Neighbour Classification, kNN example  CS 7.2.2 – kNN algorithm, impact of parameter k, solved example. |
| LE 7.1 |  |
| SS7.1 | T1- Ch14 |
| HW7.1 | Exercises at the end of T1-Ch14 |
| QZ7.1 |  |

**M8: Text Clustering**

|  |  |
| --- | --- |
| Type | Description |
| RL 8.1 | RL 8.1.1 – Clustering concepts  RL 8.1.2 – Types of clustering Flat vs. Hierarchical clustering |
| RL 8.2 | RL 8.2.1 – Flat algorithms – K means |
| RL 8.3 | RL 8.3.1 - Hierarchical clustering (HAC)  RL 8.3.2 – Cluster similarity  RL 8.3.3 – Variants of HAC |
| CS 8.1 | CS 8.1.1 – Classification vs. Clustering, Cluster hypothesis, Applications of Clustering in IR, examples of search engines  CS 8.1.2 - Flat vs. Hierarchical clustering, Hard vs. Soft clustering  CS 8.1.3 – Basic idea of K-means, RSS, K-means algorithm, Example  CS 8.1.4 – seed choice, time complexity and optimality of K-means |
| CS 8.2 | CS 8.2.1 – HAC algorithm, Dendrogram  CS 8.2.2 – Single Link and Complete link HAC- examples  CS 8.2.3 – Centroid and Group Average – examples |
| LE8.1 |  |
| SS8.1 | T1-Ch16,Ch17 |
| HW8.1 | Exercises at the end of T1-Ch16,Ch17 |
| QZ8.1 |  |

**M9: Web Search, Web Crawler, Link Analysis**

|  |  |
| --- | --- |
| Type | Description |
| RL 9.1 | RL 9.1.1 - Web Search Basics  RL 9.1.2 – Web documents  RL 9.1.3 – Duplicate detection |
| RL 9.2 | RL 9.1.1 – Web crawler operation  RL 9.1.2 – Crawling process  RL 9.1.3 – Crawler Architecture |
| RL 9.3 | RL 9.3.1 – Link analysis  RL 9.3.2 – Page Rank  RL 9.3.3 – Hubs and authorities |
| CS 9.1 | CS 9.1.1 – Ads and Web search, Web IR: Differences from traditional IR  User Needs in Web search  CS 9.1.2 - Web documents, Dynamic content, Multilinguality, spam techniques  CS 9.1.3 – Near Duplicate detection example  CS 9.1.4 – Size of the web issues |
| CS 9.2 | CS 9.2.1 – Web Crawler Architecture  CS 9.2.2 – Distributed Indexes- Web  CS 9.2.3 – Page rank algorithm – illustration with an example  CS 9.2.4 - Hubs and authorities– illustration with an example |
| LE9.1 |  |
| SS9.1 | T1 – Ch19, Ch20, Ch21 |
| HW9.1 | Exercises at the end of T1 – Ch19, Ch20, Ch21 |
| QZ9.1 |  |

**M10: Multimedia IR, Cross-Lingual IR**

|  |  |
| --- | --- |
| Type | Description |
| RL 10.1 | RL 10.1.1 – Multimedia IR- Introduction, Feature Extraction  RL 10.1.2 – Search Techniques |
| RL 10.2 | RL 10.2.1 – Cross-Lingual IR  RL 10.2.2 – Query Translation methods |
| CS 10.1 | CS 10.1.1 – Examples of query translation methods  CS 10.1.2 – Examples of parallel corpora and comparable corpora  CS 10.1.3 - Statistical MT – IBM model1 word based Machine Translation |
| CS 10.2 | CS 10.2.1 – Phrase Based Translation Models  CS 10.2.2 – Examples of Phrase Based Translation Models |
| CS 10.3 | CS 10.3.1- Multimedia search examples  CS 10.3.2 – Examples of image, audio and video feature extraction |
| CS 10.4 | CS 10.4.1- Extracting color features: Average color, color histogram, matching  CS 10.4.2 – Examples of distance measures for comparing image histograms |
| CS 10.5 | CS 10.5.1 – Collaborative filtering in Recommender systems  CS 10.5.2 - Apache Prediction I/O |
| CS 10.6 | Review |
| LE10.1 |  |
| SS10.1 | R2, R3, R4, R8 |
| HW10.1 |  |
| QZ10.1 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Contact hour | Pre-contact hour prep | During Contact hour | Post-contact hour |
| 1 | None | CS 1.1 |  |
| 2 | RL 1.1, RL 1.2, RL 1.3, RL 1.4 | CS1.2 | HW 1.1 |
| 3 | RL2.1, RL 2.2 | CS2.1, CS 2.2 | HW 2.1 |
| 4 | RL 3.1, RL 3.2 | CS 3.1, CS 3.2 | HW 2.1, HW 3.1 |
| 5 | RL3.3 | CS3.3 | HW3.1 |
| 6 | RL4.1, RL4.2 | CS4.1, CS4.2 | HW4.1,QZ4.1, SS4.1 |
| 7 | RL 5.1 | CS 5.1 | HW 5.1 |
| 8 | RL 5.2 | CS 5.2 | HW 5.1 |
| 9 | RL5.3 | CS5.3 | HW 5.1 |
| 10 | RL 6.1 | CS6.1 | HW6.1 |
| 11 | RL6.2, RL6.3 | CS6.2,CS6.3, Review | HW 6.1 |
| 12 | RL7.1,RL7.2 | CS7.1, CS 7.2 | HW 7.1 |
| 13 | RL8.1, RL8.2 | CS8.1 | HW 8.1 |
| 14 | RL8.2,RL8.3 | CS8.2 | HW8.1 |
| 15 | RL9.1,9.2 | CS9.1,CS9.2 | HW9.1 |
| 16 | RL 9.3 | CS9.2 | HW9.1 |
| 17 | RL10.1 | CS10.1 |  |
| 18 | RL10.2 | CS10.2 |  |
| 19 | - | CS10.3 |  |
| 20 | - | CS 10.4 |  |
| 21 | - | CS 10.5 |  |
| 22 | - | Review |  |

**Demonstrative Components**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.No** | **Details** | **Tools** | **Module** |
| 1. | Illustration of Stemming Process | Porter Stemmer Tool | M2, CS2.2 |
| 2. | Stemming and Lemmatization Process | Python NLTK | M2, CS2.2 |
| 3. | Index Building | Lucene | M3, M4 |
| 4. | To build NB Classifier, Nearest Neighbour Classification etc.  K Means | DatumBox FrameWork  Weka | M7  M8 |
| 5. | Demo to build Recommender System using the Collaborative Filtering | Apache Prediction I/O | M10, CS10.3 |

**Evaluation Scheme**:

Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Evaluation Component** | **Name**  (Quiz, Lab, Project, Mid term exam, End semester exam, etc) | **Type**  (Open book, Closed book, Online, etc.) | **Weight** | **Duration** | **Day, Date, Session, Time** |
| **EC - 1** | Quiz | Online | 10% |  | February 14 to 28, 2019 |
| Assignment | Take home | 10% |  | March 14 to 28, 2019 |
| **EC - 2** | Mid-Semester Test | Closed Book | 30% | 2 hrs | 09/03/2019 (FN)  10 AM – 12 Noon |
| **EC - 3** | Comprehensive Exam | Open Book | 50% | 3 hrs | 04/05/2019 (FN)  9 AM – 12 Noon |

Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 to 6.

Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 11)

**Important links and information:**

Elearn portal: https://elearn.bits-pilani.ac.in

Students are expected to visit the Elearn portal on a regular basis and stay up to date with the latest announcements and deadlines.

Contact sessions: Students should attend the online lectures as per the schedule provided on the Elearn portal.

Evaluation Guidelines:

1. EC-1 consists of either two Assignments or three Quizzes. Students will attempt them through the course pages on the Elearn portal. Announcements will be made on the portal, in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: Use of books and any printed / written reference material (filed or bound) is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam which will be made available on the Elearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.