# INFORMATION RETRIEVAL SYSTEMS
(ELECTIVE II)

**Course Code:** 15IT2115

| L | P | C |
|---|---|---|
| 3 | 0 | 3 |

**Course outcomes**: At the end of the course, a student will be able to

**CO1:** Identify pre-processing methods for effective information retrieval.
**CO2:** Apply tolerant information retrieval.
**CO3:** Describe the index compression process.
**CO4:** Transform textual information into vectors.
**CO5:** Analyse ranked and unranked search results.

**UNIT –I** (10-Lectures)

**Boolean Retrieval**: An example information retrieval problem, A first take at building an inverted index, Processing Boolean queries, The extended Boolean model versus ranked retrieval.

**The Term vocabulary and postings lists** : Document delineation and character sequence decoding, Obtaining the character sequence in a document, Choosing a document unit, Determining the vocabulary of terms ,Tokenization, Dropping common terms: stop words,
Normalization (equivalence classing of terms) stemming and lemmatization, Faster postings list intersection via skip pointers, Positional postings and phrase querie, Biword indexes , Positional indexes , Combination schemes

**UNIT –II** (10-Lectures)

**Dictionaries and tolerant retrieval** : Search structures for dictionaries ,Wildcard queries, General wildcard queries, k-gram indexes for wildcard queries, Spelling correction , Implementing spelling correction, Forms of spelling correction , Edit distance , k-gram indexes for spelling correction, Context sensitive spelling correction , Phonetic correction.

**Index construction**: Hardware basics, Blocked sort-based indexing, Single-pass in-memory indexing, Distributed indexing, Dynamic indexing, Other types of indexes

**UNIT –III** (10-Lectures)

**Index compression**: Statistical properties of terms in information retrieval, Heaps' law: Estimating the number of terms, Zipf's law: Modeling the distribution of terms, Dictionary compression, Dictionary as a string, Blocked storage, Postings file compression, Variable byte codes, ã codes.

**Scoring, term weighting** : Parametric and zone indexes, Weighted zone scoring, Learning weights, The optimal weight g, Term frequency and weighting , Inverse document frequency, Tf-idf weighting.

**UNIT –IV** (10-Lectures)

**The vector space model**: The vector space model for scoring, Dot products, Queries as vectors, Computing vector scores, Variant tf-idf functions, Sublinear tf scaling, Maximum tf normalization, Document and query weighting schemes, Pivoted normalized document length.

**UNIT –V** (10-Lectures)

**Evaluation in information retrieval**: Information retrieval system evaluation, Standard test collections, Evaluation of unranked retrieval sets, Evaluation of ranked retrieval results, Assessing relevance, Critiques and justifications of the concept of Relevance, A broader perspective: System quality and user utility, System issues , User utility, Refining a deployed system, Results snippets.

**TEXT BOOKS:**

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *"An Introduction to Information Retrieval"*, 1st Edition, Cambridge University Press, 2008.

## REFERENCES:

1. G.G. Chowdhury, "*Introduction to Modern Information Retrieval",* 3rd Edition, neal-schuman publishers, 2010.
2. Gerald J.Kowalski, Mark T.Maybury*, "Information storage and Retrieval systems: theory and implementation",* 2nd Edition, kluwer academic publishers, 2009.

## WEB REFERENCES:

1. http:// nlp.stanford.edu/IR-book/
2. ftp://mail.im.tku.edu.tw.seke.slide/baezaats/chap10_user_interfaces_ and_visualization_ir.pdf