# INFORMATION RETRIEVAL SYSTEMS (ELECTIVE-II)

**Course Code:** 13IT2116 **L P C**
**4 0 3**

**Course Educational Objectives:**.

This course is to introduce fundamental concepts of information retrieval and the procedures for evaluation

1. Understand the existing problems and potentials of current IR systems.
2. Learn and use different retrieval algorithms and systems
3. Use k-gram indexes for spelling correction
4. Implement Dictionary compression
5. Use Parametric and zone indexes

**Course outcomes:**

At the end of the course the student will be able to

1. Understand the information pre-processing concept.
2. Understand the Term Vocabulary And Postings Lists
3. Understand the Index Construction
4. Understand the Index Compression
5. Understand the Vector Space Model

**UNIT -I**

**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 queries , Biword indexes , Positional indexes , Combination schemes

**UNIT -II**

**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**

**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**

**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**

**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/baeza-ates/chap10_user_interfaces_and_visualization-modern_ir.pdf