CSC-405 Information Retrieval
Tribhuvan University
Institute of Science and Technology
Soch College of Information Technology
Bachelor of Science in Computer Science and Information Technology

Course Title: Information Retrieval
Course no: CSC-405 ---------- **Full Marks:** 60+20+20
Credit hours: 3 ---------- **Pass Marks:**24+8+8
Nature of course: Theory (3 Hrs.) + Lab (3 Hrs.)
Course Synopsis: Advanced aspects of Information Retrieval and Search Engine
Goal: To study advance aspects of information retrieval and working principle of search engine, encompassing the principles, research results and commercial application of the current technologies.

Course Contents:

Unit 1 Introduction: 2 Hrs.
Introduction, History of Information Retrieval, The retrieval process, Block diagram and architecture of IR System, Web search and IR, Areas and role of AI for IR

Unit 2 Basic IR Models: ---------- 4 Hrs.
Introduction, Taxonomy of information retrieval models, Document retrieval and ranking, A formal characterization of IR models, Boolean retrieval model, Vector-space retrieval model, probabilistic model, Text-similarity metrics: TF-IDF (term frequency/inverse document frequency) weighting and cosine similarity.

Unit 3 Basic Tokenizing, Indexing, and Implementation of Vector-Space Retrieval: ---------- 4 Hrs.
Simple tokenizing, Word tokenization, Text Normalization, Stop-word removal, Word Stemming (Porter Algorithm), Case folding, Lemmatization, Inverted indices (Indexing architecture), Efficient processing with sparse vectors, Sentence segmentation and Decision Trees

Unit 4 Experimental Evaluation of IR: ----------4 Hrs.
Relevance and Retrieval, performance metrics, Basic Measures of text retrieval (Recall, Precision and F-measure)

Unit 5 Query Operations and Languages: 3 Hrs.
Relevance feedback and pseudo relevance feedback, Query expansion/reformulation (with a thesaurus or WordNet, Spelling correction like techniques), Query languages (Single-Word Queries, Context Queries, Boolean Queries, Natural Language)

Unit 6 Text Representation: ---------- 3 Hrs.
Word statistics (Zipf's law), Morphological analysis, Index term selection, Using thesauri, Metadata, Text representation using markup languages (SGML, HTML, XML)

Unit 7 Search Engine: ---------- 6 Hrs.
Search engines (working principle), Spidering (Structure of a spider, Simple spidering algorithm, multithreaded spidering, Bot), Directed spidering(Topic directed, Link directed) ,Crawlers (Basic crawler architecture), Link analysis (e.g. hubs and authorities, Page ranking, Google Page Rank) , Shopping agents

Unit 8 Text Categorization and Clustering: ---------- 6 Hrs.
Categorization algorithms (Rocchio; naive Bayes; decision trees; and nearest neighbor), Clustering algorithms (agglomerative clustering; k-means; expectation maximization (EM)), Applications to information filtering; organization

Unit 9 Recommender Systems: ---------- 3 Hrs.
Personalization, Collaborative filtering recommendation, Content-based recommendation

Unit 10 Information Extraction and Integration:----------3 Hrs.
Information extraction and applications, Extracting data from text, Evaluating IE Accuracy, XML and Information Extraction, Semantic web (purpose, Relation to hypertext page), Collecting and integrating specialized information on the web

Unit 11 Advanced IR Models with indexing and searching text:---------- 4 Hrs.
Probabilistic models, Generalized Vector Space Model, Latent Semantic Indexing (LSI), efficient string searching, Pattern matching

Unit 12 Multimedia IR: ---------- 3 Hrs.
Introduction, multimedia data support in commercial DBMSs, Query languages, Trends and research issues

Laboratory Works: The laboratory should contain all the features mentioned in a course

Samples
Program to demonstrate the Boolean Retrieval Model and Vector Space Model
Program to find the similarity between documents
Tokenize the words of large documents according to type and token.
Segment the documents according to sentences
Implement Porter stemmer
Try to build a stemmer for Nepali language
Build a spider that tracks only the link of nepali documents

Group the online news onto different categorize like sports, entertainment, politics
Build a recommender system for online music store

Reference Books:

Modern Information Retrieval, Ricardo Baeza-Yates, Berthier Ribeiro-Neto.
Information Retrieval; Data Structures & Algorithms: Bill Frakes