Information Retrieval

Course Title: Information Retrieval
Course No: CSC413
Nature of the Course: Theory + Lab
Semester: VII

Full Marks: 60 + 20 + 20
Pass Marks: 24 + 8 + 8
Credit Hrs: 3

Course Description:
This course familiarizes students with different concepts of information retrieval techniques mainly focused on clustering, classification, search engine, ranking and query operations techniques.

Course Objective:
The main objective of this course is to provide knowledge of different information retrieval techniques so that the students will be able to develop information retrieval engine.

Course Contents:

Unit 1: Introduction to IR and Web Search (2 Hrs.)
Introduction, Data vs Information Retrieval, Logical view of the documents, Architecture of IR System, Web search system, History of IR, Related areas

Unit 2: Text properties, operations and preprocessing (5 Hrs.)
Tokenization, Text Normalization, Stop-word removal, Morphological Analysis, Word Stemming (Porter Algorithm), Case folding, Lemmatization, Word statistics (Zipf's law, Heaps' Law), Index term selection, Inverted indices, Positional Inverted index, Natural Language Processing in Information Retrieval, Basic NLP tasks – POS tagging; shallow parsing

Unit 3: Basic IR Models (5 Hrs.)
Classes of Retrieval Model, Boolean model, Term weighting mechanism – TF, IDF, TF-IDF weighting, Cosine Similarity, Vector space model , Probabilistic models (the binary independence model ,Language models; · KL-divergence; · Smoothing), Non-Overlapping Lists, Proximal Nodes Mode

Unit 4: Evaluation of IR (2 Hrs.)
Precision, Recall, F-Measure, MAP (Mean Average Precision), (DCG) Discounted Cumulative Gain, Known-item Search Evaluation

Unit 5: Query Operations and Languages (4 Hrs.)
Relevance feedback and pseudo relevance feedback, Query expansion (with a thesaurus or WordNet and correlation matrix), Spelling correction (Edit distance, K – Gram indexes, Context sensitive spelling correction), Query languages (Single-Word Queries, Context Queries, Boolean Queries, Structural Query, Natural Language)

Unit 6: Web Search (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 (HITS, Page ranking), Query log analysis, Handling "invisible" Web – Snippet generation, CLIR (Cross Language Information Retrieval)

Unit 7: Text Categorization (4 Hrs.)
Categorization, Learning for Categorization, General learning issues, Learning algorithms: Bayesian (naïve), Decision tree, KNN, Rocchio)

Unit 8: Text Clustering (4 Hrs.)
Clustering, Clustering algorithms (Hierarchical clustering, k-means, k-medoid, Expectation maximization (EM), Text shingling)

Unit 9: Recommender System (3 Hrs.)
Personalization, Collaborative filtering recommendation, Content-based recommendation

Unit 10: Question Answering (5 Hrs.)
Information bottleneck, Information Extraction, Ambiguities in IE, Architecture of QA system, Question processing, Paragraph retrieval, Answer processing

Unit 11: Advanced IR Models (5 Hrs.)
Latent Semantic Indexing (LSI), Singular value decomposition, Latent Dirichlet Allocation, Efficient string searching, Knuth – Morris – Pratt, Boyer – Moore Family, Pattern matching

Laboratory Works:
The laboratory should contain all the features mentioned in a course. The Laboratory work should contain at least following tasks

1. Program to demonstrate the Boolean Retrieval Model and Vector Space Model
2. Tokenize the words of large documents according to type and token
3. Program to find the similarity between documents
4. Implement Porter stemmer
5. Build a spider that tracks only the link of nepali documents
6. Group the online news onto different categorize like sports, entertainment, politics
7. Build a recommender system for online music store

Recommended Books:

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