COURSE TITLE	**INFORMATION RETRIEVAL AND NATURAL LANGUAGE PROCESSING**
COURSE CODE	**01CT0723**
COURSE CREDITS	**4**

Objective:

1 This course provides with an understanding of common and emerging methods of organizing, summarizing, and analyzing large collections of unstructured and lightly structured text. Students will gain an in-depth knowledge of the commonly used algorithms for processing the natural languages. The course examines NLP models and algorithms using both traditional symbolic recent statistical approaches.

2 This course provides an understanding of common and emerging methods of organizing, summarizing, and analyzing large collections of unstructured and lightly structured text. Students will gain an in-depth knowledge of the commonly used algorithms for processing the natural languages. The course examines NLP models and algorithms using both traditional symbolic recent statistical approaches.

Course Outcomes: After completion of this course, student will be able to:

1 Comprehend types of text analysis, Information retrieval, IR system architecture, query processing models and probabilistic models.

2 To understand natural language processing and importance of word representation.

3 Manage information retrieval systems by performing network management, search engine optimization, records compliance and risk management.

4 Perform indexing, compression, information categorization and sentiment analysis

5 Apply deep learning to solve natural language problems such as language modelling, machine translation, POS tagging, Seq2Seq generation

6 Solve NLP problem of real time scenario

Pre-requisite of course: Programming, Algorithms, Data Structures, Machine Learning, Artificial Intelligence

Teaching and Examination Scheme

Theory Hours	Tutorial Hours	Practical Hours	ESE	IA	CSE	Viva	Term Work
3	0	2	50	30	20	25	25

Contents : Unit	Topics	Contact Hours
1	**Introduction to Information Retrieval** Boolean retrieval, Vector Space Model, Feature Vectors, Document/Passage Retrieval, Search Engines, Relevance Feedback & Query Expansion, Document Filtering and Categorization, flat and hierarchical clustering, Latent Semantic Analysis, Web Crawling and the Google algorithm	4

MR. CHANDRASINH DHIRUBHA PARMAR
Digitally signed by (Name of HOD)

DR. RAJENDRASINH BAHADURSINH JADEJA
Digitally signed by (Name of Dean/ Principal)

Contents : Unit	Topics	Contact Hours
2	**Language models and IR systems** Unigram, Bigram language models,, generating queries from documents, Language models and smoothing, , ranking with language models, Kullback Leibler divergence, , Divergence from randomness, Passage retrieval and ranking, Management of Information Retrieval Systems: Knowledge management, Information management, Digital asset management, Network management, Search engine optimization, Records compliance and risk management, Version control, Data and data quality, Information system failure. , Types of information retrieval systems: Web retrieval and mining, Semantic web, XML information retrieval, Recommender systems and expert locators, Knowledge management systems, Decision support systems, , Geographic information system (GIS). Indexing: Inverted indices, Index components and Index life cycle, Interleaving Dictionary and Postings lists, Index construction	12
3	**Statistical Natural Language Processing** Sequence Labeling: POS-tagging, Named Entity Recognition and Normalization, Syntactic Parsing: Dependency Syntactic Parsing, Ambiguity in language, Shallow Semantic Parsing: Predicate Argument Structures, Relation Extraction (supervised and semi-supervised), Discourse Parsing: Coreference Resolution and discourse connective classification	8
4	**Neural NLP Models** Word Window Classification, Neural Networks for text, N-gram Language Models, Perplexity, Hidden Markov Models, Recurrent Neural network, Vanishing Gradients and exploding gradient, 1D-CNN for NLP, Contextual Representations, Encoder-Decoder, Transformers	9
5	**Joint NLP and IR applications** Deep Linguistic Analysis for Question Answering: QA tasks (open, restricted, factoid, non-factoid), NLP Representation, Question Answering Workflow, QA Pipeline, Question Classification, Fine-Grained Opinion Mining: automatic review classification, automatic product extraction and review	9
Total Hours		**42**

Suggested List of Experiments:

Contents : Unit	Topics	Contact Hours
1	**Experiment-1** Implementation of the NLP Boolean model	2
2	**Experiment-2** Implement the Vector Space model	2
3	**Experiment-3** To implement the lemmatization and Stemming Algorithms	2

MR. CHANDRASINH DHIRUBHA PARMAR
Digitally signed by (Name of HOD)

DR. RAJENDRASINH BAHADURSINH JADEJA
Digitally signed by (Name of Dean/ Principal)

Suggested List of Experiments:

Contents : Unit	Topics	Contact Hours
4	**Experiment-4** Index creation for IR system: Inverted Files	2
5	**Experiment-5** Index creation for IR system: Signature Files	2
6	**Experiment-6** To construct the thesaurus from the word corpus	2
7	**Experiment-7** Text document pre-processing and Classification	2
8	**Experiment-8** To construct the Ontology relation among the words and sentences	2
9	**Experiment-9** Text processing practice: sentence segmentation, word tokenization, stemming and lemmatization, preparation of dictionary, etc	2
10	**Experiment-10** Implement a n-gram model	2
11	**Experiment-11** Programming exercises on employing HMM for PoS tagging	2
12	**Experiment-12** Programming exercises to implement log-linear model for PoS tagging problem	2
13	**Experiment-13** Programming exercises for employing CRF on NER tasks	2
14	**Experiment-14** Programming exercises for using existing NLP tools (CoreNLP/NLTK) and obtaining the syntactic parsing of the text	2
15	**Experiment-15** Programming exercises for employing a multi-layer feedforward network on PoS tagging and NER tasks	2
16	**Experiment-16** Programming exercises for using LSTM in text classification task	2
17	**Experiment-17** Programming exercises for employing CNN on the text classification task	2
18	**Experiment-18** Programming exercises for using a sequence to sequence model on machine translation tasks	2
Total Hours		**36**

Textbook :

1 Speech and Language processing an introduction to Natural Language Processing, Computational Linguis , Daniel Jurafsky and James H. Martin, Prentice Hall, 2008

2 Natural Language Processing and Information Retrieval , Siddiqui and Tiwari, Oxford University Press, 2008

References:

1 Natural Language Processing with Python , Natural Language Processing with Python , Steven Bird, Ewan Klein and Edward Lopper, O'Reilly, 2009

2 Information Retrieval , Information Retrieval , Butcher S., Clarke C.L.A. and Cormack G., MIT Press, 2010

3 Understanding Information Retrieval Systems , Understanding Information Retrieval Systems , Bates M.J., CRC press, 2011

Suggested Theory Distribution:

The suggested theory distribution as per Bloom's taxonomy is as follows. This distribution serves as guidelines for teachers and students to achieve effective teaching-learning process

Distribution of Theory for course delivery					
Remember / Knowledge	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Higher order Thinking / Creative**
15.00	20.00	30.00	20.00	10.00	5.00

Instructional Method:

1 The course delivery method will depend upon the requirement of content and need of students. The teacher in addition to conventional teaching method by black board, may also use any of tools such as demonstration, role play, Quiz, brainstorming, MOOCs etc.

2 The internal evaluation will be done on the basis of continuous evaluation of students in the laboratory and class-room.

3 Practical examination will be conducted at the end of semester for evaluation of performance of students in laboratory.

4 Students will use supplementary resources such as online videos, NPTEL videos, e-courses, Virtual Laboratory.

Supplementary Resources:

1 https://cloud.google.com/natural-language

2 https://github.com/delip/PyTorchNLPBook

3 https://www.nltk.org/book/