Course Code	Course Name	Credits
DLO8012	**Natural Language Processing**	**4**

Course objectives:

1. To understand natural language processing and to learn how to apply basic algorithms in this field.
2. To get acquainted with the basic concepts and algorithmic description of the main language levels: morphology, syntax, semantics, and pragmatics.
3. To design and implement applications based on natural language processing
4. To implement various language Models.
5. To design systems that uses NLP techniques

Course outcomes: On successful completion of course learner should:

1. Have a broad understanding of the field of natural language processing.
2. Have a sense of the capabilities and limitations of current natural language technologies,
3. Be able to model linguistic phenomena with formal grammars.
4. Be able to Design, implement and test algorithms for NLP problems
5. Understand the mathematical and linguistic foundations underlying approaches to the various areas in NLP
6. Be able to apply NLP techniques to design real world NLP applications such as machine translation, text categorization, text summarization, information extraction...etc.

Prerequisite: Data structure & Algorithms, Theory of computer science, Probability Theory.

Module No.	Unit No.	Topics	Hrs.
1	**Introduction**	History of NLP, Generic NLP system, levels of NLP , Knowledge in language processing , Ambiguity in Natural language , stages in NLP, challenges of NLP ,Applications of NLP	4
2	**Word Level Analysis**	Morphology analysis –survey of English Morphology, Inflectional morphology & Derivational morphology, Lemmatization, Regular expression, finite automata, finite state transducers (FST) ,Morphological parsing with FST , Lexicon free FST Porter stemmer. N –Grams- N-gram language model, N-gram for spelling correction.	10
3	**Syntax analysis**	Part-Of-Speech tagging(POS)- Tag set for English (Penn Treebank) , Rule based POS tagging, Stochastic POS tagging, Issues –Multiple tags & words, Unknown words. Introduction to CFG, Sequence labeling: Hidden Markov Model (HMM), Maximum Entropy, and Conditional Random Field (CRF).	10
4	**Semantic Analysis**	Lexical Semantics, Attachment for fragment of English- sentences, noun phrases, Verb phrases, prepositional phrases, Relations among lexemes & their senses –Homonymy, Polysemy, Synonymy, Hyponymy, WordNet, Robust Word Sense Disambiguation (WSD) ,Dictionary based approach	10

5	**Pragmatics**	Discourse –reference resolution, reference phenomenon , syntactic & semantic constraints on co reference	8
6	**Applications (preferably for Indian regional languages)**	Machine translation, Information retrieval, Question answers system, categorization, summarization, sentiment analysis, Named Entity Recognition.	10

Text Books:

1. Daniel Jurafsky, James H. Martin "Speech and Language Processing" Second Edition, Prentice Hall, 2008.
2. Christopher D.Manning and Hinrich Schutze, " Foundations of Statistical Natural Language Processing ", MIT Press, 1999.

Reference Books:

1. Siddiqui and Tiwary U.S., Natural Language Processing and Information Retrieval, Oxford University Press (2008).
2. Daniel M Bikel and Imed Zitouni " Multilingual natural language processing applications" Pearson, 2013
3. Alexander Clark (Editor), Chris Fox (Editor), Shalom Lappin (Editor) " The Handbook of Computational Linguistics and Natural Language Processing " ISBN: 978-1-118-
4. Steven Bird, Ewan Klein, Natural Language Processing with Python, O'Reilly
5. Brian Neil Levine, An Introduction to R Programming
6. Niel J le Roux, Sugnet Lubbe, A step by step tutorial : An introduction into R application and programming

Assessment:

Internal Assessment:
Assessment consists of two class tests of 20 marks each. The first class test is to be conducted when approx. 40% syllabus is completed and second class test when additional 40% syllabus is completed. Duration of each test shall be one hour.

End Semester Theory Examination:
1 Question paper will comprise of 6 questions, each carrying 20 marks.
2. The students need to solve total 4 questions.
3. Question No.1 will be compulsory and based on entire syllabus.
4. Remaining question (Q.2 to Q.6) will be selected from all the modules.

Laboratory Work/Case study/Experiments:

Description: The Laboratory Work (Experiments) for this course is required to be performed and to be evaluated in CSL803: Computational Lab-II

The objective of Natural Language Processing lab is to introduce the students with the basics of NLP which will empower them for developing advanced NLP tools and solving practical problems in this field.

Reference for Experiments: **http://cse24-iiith.virtual-labs.ac.in/#**

Reference for NPTEL: http://www.cse.iitb.ac.in/~cs626-449

Sample Experiments: possible tools / language: R tool/ Python programming Language

Note: Although it is not mandatory, the experiments can be conducted with reference to any Indian regional language.

1. Preprocessing of text (Tokenization, Filtration, Script Validation, Stop Word Removal, Stemming)
2. Morphological Analysis
3. N-gram model
4. POS tagging
5. Chunking
6. Named Entity Recognition
7. Case Study/ Mini Project based on Application mentioned in Module 6.