

Course Code:	Course Title	Credit
CSDO7011	Natural Language Processing	3

Prerequisite: Artificial Intelligence and Machine Learning, Basic knowledge of Python	
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 various language models and POS tagging techniques
4	To design and learn NLP applications such as Information Extraction, Question answering
5	To design and implement applications based on natural language processing
Course Outcomes:	
1	To have a broad understanding of the field of natural language processing
2	To design language model for word level analysis for text processing
3	To design various POS tagging techniques
4	To design, implement and test algorithms for semantic analysis
5	To develop basic understanding of Pragmatics and to formulate the discourse segmentation and anaphora resolution
6	To apply NLP techniques to design real world NLP applications

Module		Content	Hrs
1		Introduction	4
	1.1	Origin & History of NLP, The need of NLP, Generic NLP System, Levels of NLP, Knowledge in Language Processing, Ambiguity in Natural Language, Challenges of NLP, Applications of NLP.	
2		Word Level Analysis	8
	2.1	Tokenization, Stemming, Segmentation, Lemmatization, Edit Distance, Collocations, Finite Automata, Finite State Transducers (FST), Porter	

		Stemmer, Morphological Analysis, Derivational and Reflectional Morphology, Regular expression with types.	
	2.2	N –Grams, Unigrams/Bigrams Language Models, Corpora, Computing the Probability of Word Sequence, Training and Testing.	
3		Syntax analysis	8
	3.1	Part-Of-Speech Tagging (POS) - Open and Closed Words. Tag Set for English (Penn Treebank), Rule Based POS Tagging, Transformation Based Tagging, Stochastic POS Tagging and Issues –Multiple Tags & Words, Unknown Words.	
	3.2	Introduction to CFG, Hidden Markov Model (HMM), Maximum Entropy, And Conditional Random Field (CRF).	
4		Semantic Analysis	8
	4.1	Introduction, meaning representation; Lexical Semantics; Corpus study; Study of Various language dictionaries like WordNet, Babelnet; Relations among lexemes & their senses –Homonymy, Polysemy, Synonymy, Hyponymy; Semantic Ambiguity	
	4.2	Word Sense Disambiguation (WSD); Knowledge based approach (Lesk's Algorithm), Supervised (Naïve Bayes, Decision List), Introduction to Semi-supervised method (Yarowsky), Unsupervised (Hyperlex)	
5		Pragmatic & Discourse Processing	6
	5.1	Discourse: Reference Resolution, Reference Phenomena, Syntactic & Semantic constraint on coherence; Anaphora Resolution using Hobbs and Cantering Algorithm	
6		Applications (preferably for Indian regional languages)	5
	6.1	Machine Translation, Information Retrieval, Question Answers System, Categorization, Summarization, Sentiment Analysis, Named Entity Recognition.	
	6.2	Linguistic Modeling – Neurolinguistics Models- Psycholinguistic Models – Functional Models of Language – Research Linguistic Models- Common Features of Modern Models of Language.	

Textbooks:	
1	Daniel Jurafsky, James H. and Martin, Speech and Language Processing, Second Edition, Prentice Hall, 2008.
2	Christopher D.Manning and HinrichSchutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
References:	
1	Siddiqui and Tiwary U.S., Natural Language Processing and Information Retrieval, Oxford University Press, 2008.
2	Daniel M Bikel and ImedZitouni — Multilingual natural language processing applications: from theory to practice, IBM Press, 2013.
3	Nitin Indurkha and Fred J. Damerau, —Handbook of Natural Language Processing, Second Edition, Chapman and Hall/CRC Press, 2010.

<u>Assessment:</u>	
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 total six questions.
2	All question carries equal marks
3	Questions will be mixed in nature (for example supposed Q.2 has part (a) from module 3 then part (b) will be from any module other than module 3)
4	Only Four question need to be solved
5	In question paper weightage of each module will be proportional to number of respective lecture hours as mention in the syllabus

Useful Links
