Name of the Programme : **M.Sc. in Data Science**
Course Code : **CSD-601**
Title of Course : **Natural Language Processing**
Number of Credits : **4(3L+ 1T)**
Contact hours : **60 hours (45L-15T)**
Effective from AY : **2023-24**

Pre-requisites for the Course	Python Programming and Machine Learning	
Course Objectives	This course will provide a foundational understanding of NLP methods and strategies, evaluate strengths and weaknesses of various NLP technologies and frameworks, and gain practical experience in NLP toolkits.	
Content	Introduction, Machine Learning and NLP, ArgMax Computation, Word Sense Disambiguation: WordNet, Wordnet; Application in Query Expansion, Measures of WordNet Similarity. Resnick's work on WordNet Similarity, Parsing Algorithms, Evidence for Deeper Structure; Top-Down Parsing Algorithms, Noun Structure; Top-Down Parsing Algorithms, Non-noun Structure and Parsing Algorithms.	**15 hours**
	Probabilistic parsing; Sequence labelling, PCFG, Probabilistic parsing: Training issues, Arguments and Adjuncts, Probabilistic parsing; inside-outside probabilities. Speech: Phonetics, Hidden Markov Model, Morphology, Graphical Models for Sequence Labelling in NLP, Consonants (place and manner of articulation) and Vowels.	**15 hours**
	Forward Backward probability; Viterbi Algorithm, Phonology, Sentiment Analysis and Opinions on the Web, Machine Translation and MT Tools - GIZA++ and Moses, Text Alignment, POS Tagging. Phonology; ASR, Speech Synthesis, Hidden Markov Model and Viterbi, Precision, Recall, F-score, Map, Semantic Relations; UNL; Towards Dependency Parsing. Universal Networking Language, Semantic Role Extraction, Baum Welch Algorithm; HMM training.	**15 hours**

	Tutorial assignments: 1. Import nltk and download the 'stopwords' and 'punkt' packages and Import spacy and load the language model 2. Program to tokenize a given text, to get the sentences of a text document 3. program to tokenize a text using th`transformers` package, tokenize text with stopwords as delimiters, remove, stop words in a text, add custom stop words in spaCy remove punctuations, and perform stemming. 4. Program to lemmatize a given text, extract usernames from emails, find the most common words in the text excluding stopwords 5. Program to do spell correction in a given text, tokenize tweets, extract all the nouns in a text, extract all the pronouns in a text,find similarity between two words, find similarity between two documents, find the cosine similarity of two documents.	**3x5=15 hours**
Pedagogy	Lectures/ Tutorials/Hands-on assignments/Self-study/Flipped classroom	
References/ Readings	1. Allen, J. (1995). Natural language understanding. Benjamin-Cummings Publishing Co., Inc.. 2. Charniak, E. (1996). Statistical language learning. MIT press. 3. Jurafsky, D. (2008). Martin, and H. James. Speech and Language Processing (2nd Edition)(Prentice Hall Series in Artificial Intelligence). 4. Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT press.	
Course Outcomes	After completion of this course, students will be able to: 1. apply various NLP methods and strategies for tasks such as text representation, tokenization, part-of-speech tagging, and syntactic analysis. 2. Analyze sentence structure using syntactic analysis and parsing techniques, including constituency and dependency parsing. 3. Explore real-world applications of NLP	

(Back to Index)