Course Code	UDS21303J	Course Name	INTRODUCTION TO NATURAL LANGUAGE PROCESSING	Course Category	C	Professional Core Course	L	T	P	C
							4	0	2	5

Pre-requisite Courses	Nil	Co-requisite Courses	Nil	Progressive Courses	Nil
Course Offering Department	Computer Applications	Data Book / Codes/Standards	Nil		

Course Learning Rationale (CLR):	The purpose of learning this course is to,	Learning

		Learning			Program Learning Outcomes (PLO)														
		1	2	3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
CLR-1 :	To make the participants comfortable with the fundamentals of Natural Language Processing, their working principles and their functions in a business scenario.																		
CLR-2 :	To teach the participants to build intelligent and automated real-world natural language processing applications and use cases spanning healthcare, retail, energy verticals by intelligently analyzing different datasets collected from diverse data sources.																		
CLR-3 :	To teach the participants the various layers of Natural Language processing architecture, detailed steps are involved in transforming raw data into training datasets for critical decision making.																		
CLR-4 :	To teach the students the about the overall process involved in text processing and building an enterprise grade natural language processing solutions																		
CLR-5 :	To learn how to apply Natural language processing models to business problems, build data for efficient data collection, preparation, provisioning, model pipelines for model engineering and validation tasks.																		
CLR-6 :	To bring the learners to an alignment, apply their learning to a real-world business problem, and then performs research, design, development, and delivers an end-to-end Natural language processing solution for a given industry problem. The students will be working either in a group or individually.																		

Course Learning Outcomes (CLO):	At the end of this course, learners will be able to:	Level of Thinking (Bloom)	Expected Proficiency (%)	Expected Attainment (%)	Fundamental Knowledge	Application of Concepts	Link with Related Disciplines	Procedural Knowledge	Skills in Specialization	Ability to Utilize Knowledge	Skills in Modeling	Analyze, Interpret Data	Investigative Skills	Problem Solving Skills	Communication Skills	Analytical Skills	ICT Skills	Professional Behavior	Life Long Learning
CLO-1 :	Have a strong control over the fundamental concepts of Natural Language Processing including the ability to clearly define Natural Language Processing from both academic and industry perspective.	2	85	80	H	H	H	H	H	H	H	H	H	M	M	H	H	H	H
CLO-2 :	Gain hands-on solid skills, knowledge and expertise of real-world situations the applicability of tools and techniques in extracting valuable insights from the data of different formats on time.	3	85	80	H	H	H	H	H	H	H	H	H	M	M	H	H	H	H
CLO-3 :	Have solid hands-on skills, knowledge and expertise in Data gathering, Data collection, Model training, and model evaluation with domain-specific components.	3	85	80	H	H	H	H	H	H	H	H	H	M	M	H	H	H	H

CLO-4 :	Have a good Hands-on skills and knowledge to apply all the required processes on texts	3	85	80	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*M*	*M*	*H*	*H*	*H*	*H*
CLO-5 :	Have solid hands-on skills, knowledge and expertise in setting up a data platform for building enterprise-grade natural language processing solutions.	3	85	80	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*M*	*M*	*H*	*H*	*H*	*H*
CLO-6 :	Design and develop natural language processing solution artifacts and ultimately demonstrate an "end-to-end" machine learning solution for a given problem statement either in a group or individually.	3	85	80	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*M*	*M*	*H*	*H*	*H*	*H*

Note: All our curriculum, study materials, assignments, quizzes, lab works, and learning resources are personalized and dynamically generated using machine learning models based on the learner's learning ability. Users can review our learning curriculum only through our intelligent learning management platform (iLMSP), and our learning resources and lab infrastructures are available only in the digital form on our cloud infrastructures.

Duration (hour)		**18**	**18**	**18**	**18**	**18**
S-1	SLO-1	**Unit 1: Natural Language Processing Defined - Academic and Industry Perspective**	Pattern Mining	Topic Modelling	DeBERTa	Adding Packages
	SLO-2	What is Natural Language Processing?	Evaluation and Deployment	Text Classification	**Unit 10: Natural Language Processing Data Requirements**	**Unit 12: Natural Language Processing Data Requirements**
S-2	SLO-1	Natural Language Processing defined from Academic and Industry perspective	**Unit 5: Natural Language Processing Architecture**	Keyword Classification	How much data is needed	Patient Readmittance with discharge summaries
	SLO-2	Functions of a Natural Language Processing system	Components of machine learning solution	Lemmatization	Is your data good enough?	Who is going to get readmitted?
S-3	SLO-1	What does a Natural Language Processing system do?	Data Generation	Stemming	Data Structure	When will they get readmitted
	SLO-2	How a business uses Natural Language Processing	Data Collection	Part of speech tagging	Data Format	Why will they get readmitted
S-4	SLO-1	How Natural Language Processing works?	Feature Engg pipeline	Coreference resolution	Data Type	Problem statement
	SLO-2	**Unit 2: Demystifying Artificial Intelligence and Natural Language Processing**	Training	**Unit 8: What Problem Natural Language Processing Solves**	Source System	Problem type
S-5 & S-6	SLO-1	**Lab 1 : Import the nltk package in python and download**	**Lab 4 : Create a monolingual corpus of 200,000 words. Segment it**	**Lab 7: Choose a corpus of at least 20,000 words of online text,**	**Lab 10 : Estimate how much storage space is necessary for the**	**Lab 13: Extract the the topics from the any texts of your choice**
	SLO-2					

		'stopwords', 'punkt' packages, tokenize the string using the `transformers` package	**into words, and compute the frequency of each word. How many distinct words are there? count frequencies of bigrams (two consecutive words) and trigrams (three consecutive words).**	**and verify Zipf's law experimentally. Define an error measure and find the value of α where Zipf's law best matches your experimental data**	**index to a 100 billion-page corpus of Web pages. Show the assumptions you made**	**with Latent dirichlet algorithm**
S-7	SLO-1	What are Natural Language Processing promises and challenges?	Evaluation	Machine Translation	Target system	Data engineering
	SLO-2	Natural Language Processing Architecture, Libraries, Technologies and Framework	Task Orchestration	Named Entity Recognition	Training Data	Data pipeline
S-8	SLO-1	Why is Natural Language Processing so important?	Prediction	Text/Classification	Validation Data	Model selection
	SLO-2	Components of Natural Language Processing ✓ Natural language Understanding ✓ Natural language Generation	Infrastructure	Text Summarization	Test Data	Model engineering
S-9	SLO-1	Phases of Natural Language Processing ✓ Lexical Analysis ✓ Syntactic Analysis ✓ Semantic Analysis ✓ Disclosure Integration ✓ Pragmatic Analysis	Authentication	Topic Modelling	**Unit 11: Natural Language Processing Data Requirements**	Model Outcome
	SLO-2	**Unit 3: Natural Language Processing in Real World Applications**	Interaction	Keyword Extraction	Building a NLP Hardware system	Model Analysis
S-10	SLO-1	NLP in healthcare	Monitoring	Information Retrieval	Benefits	Model Optiization
	SLO-2	NLP in Retail	Building your NLP Architecture	Automatic Image annotation	Challenges	Model pipeline
S-11 &	SLO-1 SLO-2	**Lab 2 :** **With your knowledge of the English language, split 10**	**Lab 5 :**	**Lab 8:** **Create a corpus of spam email and one of non-spam**	**Lab 11:** **Write a regular expression or a short program to extract**	**Lab 14:** **Extract the the topics from the any texts of your choice**

S-12		**sentences of your choice into words and punctuation: Find out the words words that don't usually appear in a standard lexicon? The separators are: whitespaces, quote ('), full-stop/period (.), parenthesis, are kept as tokens, tokenize the earlier sentence.**	**Write a program to do segmentation of words without spaces. Given a string, such as the URL "thelongestlistofthelongeststuffatthelongestdomainnameatlonglast.com," return a list of component words: ["the," "longest," "list," …]. This task is useful for parsing URLs, for spelling correction when words runtogether, and for languages such as Chinese that do not have spaces between words**	**mail. Examine each corpus and decide what features appear to be useful for classification: unigram words? bigrams? message length, sender, time of arrival?**	**company names. Test it on a corpus of business news articles. Report your recall and precision.**	**using Non-negative Matrix Factorization**
S-13	SLO-1	NLP in Energy	**Unit 6: Natural Language Processing Implementation Framework**	**Unit 9: Natural Language Processing Models**	High level decisions	Data visualization
	SLO-2	NLP in Oil & Gas	What is a NLP framework?	BERT	Choosing the hardware components (GPU, TPU)	User interface
S-14	SLO-1	NLP in Automobile	Features of a good NLP framework	GPT2	Building a NLP Software system	
	SLO-2	**Unit 4: Natural Language Processing Workflow**	Popular NLP frameworks ✓ NLTK ✓ Gensim ✓ SpaCy ✓ CoreNLP	XLNet	Benefits	
S-15	SLO-1	Text pre-processing ✓ Contraction Mapping ✓ Tokenization ✓ Noise Cleaning ✓ Spell Checking ✓ Stop words Removal ✓ Stemming ✓ Lemmatization	**Unit 7: Natural Language Processing - Techniques an Overview**	Electra	Challenges	
	SLO-2	Exploratory Data Analysis	Pattern Recognition	Text to Text Transfer Transformer	High level decisions	
S-16	SLO-1	Text pre-processing	Named Entity Recognition	RoBERTa	Choosing the software components	
	SLO-2	Text Representation & Feature Engineering	Text Summarization	ALBERTA	Choosing the OS	

S-17 & S-18	SLO-1 SLO-2	**Lab 3:** **Design a NLP application which measures the edit distance between words using the chartbased algorithm.** **Provide the filled data structure resulting from the application of the algorithm to the pair "easy" and "tease". Briefly justify your answer.**	**Lab 6:** **Perform word segmentation implementation on a bigger example corpus. E.g., try the first N words in the Brown corpus.**	**Lab 9:** **Create a test set of ten queries, and pose them to three major Web search engines. Evaluate each one for precision at 1, 3, and 10 documents. Can you explain the differences between engines?**	**Lab 12:** **Implement Soft Cosine Similarity in python**	**Lab 15:** **Utilize Word2Vec model for representing words and plot the word embedding from the output of the word2Vec model**

Learning Resources	1. The textbook for the course will be the second edition of Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, by Daniel Jurafsky and James H. Martin 2. James A.. Natural language Understanding 2e, Pearson Education, 1994 3. Bharati A., Sangal R., Chaitanya V.. Natural language processing: a Paninian perspective, PHI, 2000	4. https://www.nltk.org/book/ 5. Siddiqui T., Tiwary U. S.. Natural language processing and Information retrieval, OUP,2008

Learning Assessment

	Bloom's Level of Thinking	Continuous Learning Assessment (50% weightage)								Final Examination (50% weightage)	
		CLA – 1 (10%)		CLA – 2 (10%)		CLA – 3 (20%)		CLA – 4 (10%) #			
		Theory	Practice	Theory	Practice	Theory	Practice	Theory	Practice	Theory	Practice
Level 1	Remember	20%	15%	20%	15%	20%	15%	20%	15%	20%	15%
	Understand										
Level 2	Apply	20%	20%	20%	20%	20%	20%	20%	20%	20%	20%
	Analyze										
Level 3	Evaluate	10%	15%	10%	15%	10%	15%	10%	15%	10%	15%
	Create										
	Total	100 %		100 %		100 %		100 %		100 %	

CLA – 4 can be from any combination of these: Assignments, Seminars, Tech Talks, Mini-Projects, Case-Studies, Self-Study, MOOCs, Certifications, Conf. Paper etc.,

Course Designers

Experts from Industry	Experts from Higher Technical Institutions	Internal Experts
Mr.Jothi, Periyasamy , Chief AI Architect DeepSphere.AI, CA, USA	Dr.S.Gopinathan, Associate Professor, University of Madras, Chennai	Dr.Pandiyan, SRMIST
		Dr.S.Sivakumar, SRMIST