09-04-2025 02:35:25 PM

NATURAL LANGUAGE PROCESSING **Course Code : 315329**

Programme Name/s : **Artificial Intelligence/ Artificial Intelligence and Machine Learning/ Data Sciences**

Programme Code : **AI/ AN/ DS**

Semester : **Fifth**

Course Title : **NATURAL LANGUAGE PROCESSING**

Course Code : **315329**

I. RATIONALE

This course emphasizes foundational knowledge and practical skills in language processing. It equips students with the ability to develop text-based applications, handle text data processing and preparing them for role of NLP Engineer in the software industry.

II. INDUSTRY / EMPLOYER EXPECTED OUTCOME

Design and develop NLP-based applications and use NLP toolkits.

III. COURSE LEVEL LEARNING OUTCOMES (COS)

Students will be able to achieve & demonstrate the following COs on completion of course based learning

- CO1 - Explain key concepts linguistics and NLP.
- CO2 - Implement Text Normalization and Text Preprocessing techniques to the text.
- CO3 - Apply Part Of Speech ,Parsing ,Named Entity Recognition techniques to the text.
- CO4 - Generate text embedding in NLP.
- CO5 - Use Transformer in NLP applications.

IV. TEACHING-LEARNING & ASSESSMENT SCHEME

Course Code	Course Title	Abbr	Course Category/s	Learning Scheme					Credits	Assessment Scheme											
				Actual Contact Hrs./Week			SLH	NLH		Paper Duration	Theory			Based on LL & TL				Based on SL		Total Marks
														Practical						
				CL	TL	LL					FA-TH	SA-TH	Total	FA-PR		SA-PR		SLA		
											Max	Max	Max	Min	Max	Min	Max	Min	Max	Min	
315329	NATURAL LANGUAGE PROCESSING	NLP	DSE	4	-	2	-	6	2	3	30	70	100	40	25	10	25#	10	-	-	150

Total IKS Hrs for Sem. : 1 Hrs

Abbreviations: CL- ClassRoom Learning , TL- Tutorial Learning, LL-Laboratory Learning, SLH-Self Learning Hours, NLH-Notional Learning Hours, FA - Formative Assessment, SA -Summative assessment, IKS - Indian Knowledge System, SLA - Self Learning Assessment

Legends: @ Internal Assessment, # External Assessment, *# On Line Examination , @$ Internal Online Examination

Note :

1. FA-TH represents average of two class tests of 30 marks each conducted during the semester.
2. If candidate is not securing minimum passing marks in FA-PR of any course then the candidate shall be declared as "Detained" in that semester.
3. If candidate is not securing minimum passing marks in SLA of any course then the candidate shall be declared as fail and will have to repeat and resubmit SLA work.
4. Notional Learning hours for the semester are (CL+LL+TL+SL)hrs.* 10 Weeks
5. 1 credit is equivalent to 30 Notional hrs.
6. * Self learning hours shall not be reflected in the Time Table.
7. * Self learning includes micro project / assignment / other activities.

V. THEORY LEARNING OUTCOMES AND ALIGNED COURSE CONTENT

NATURAL LANGUAGE PROCESSING **Course Code : 315329**

Sr.No	Theory Learning Outcomes (TLO's)aligned to CO's.	Learning content mapped with Theory Learning Outcomes (TLO's) and CO's.	Suggested Learning Pedagogies.
1	TLO 1.1 Explain the significance of Language Syntax, Structure and Semantics. TLO 1.2 Use different Text Corpora. TLO 1.3 Describe various applications of NLP.	**Unit - I Natural Language Basics** 1.1 Overview of NLP ,The need of NLP, Areas of study under linguistics 1.2 Language Syntax and Structure: Words, Phrases, Clauses, Grammar, Word Order Typology, Word Order–Based Language Classification 1.3 Language Semantics: Lexical Semantic Relations,Semantic Networks and Models 1.4 Text Corpora:Corpora Annotation and Utilities, Popular Corpora 1.5 Applications of NLP.	Lecture Using Chalk-Board Presentations
2	TLO 2.1 Apply regex patterns to match and manipulate text. TLO 2.2 Apply Text normalization techniques. TLO 2.3 Apply word tokenization , lemmatization, stemming techniques. TLO 2.4 Generate N-grams.	**Unit - II Text Normalization and Preprocessing** 2.1 Regular Expressions :Basic Regular Expression Patterns ,Disjunction, Grouping, and Precedence, sets of characters, operators for counting 2.2 Text Preprocessing: Removing Special Characters ,Expanding Contractions , Case Conversions , Removing Stopwords 2.3 Word Tokenization: rule-based tokenization, Byte-Pair Encoding 2.4 Lemmatization and Stemming ,porter stemmer, snowball stemmer 2.5 N-grams Vectors :unigram,bigram, trigram,n-gram	Lecture Using Chalk-Board Demonstration Hands-on
3	TLO 3.1 Assign grammatical categories to individual words in a text. TLO 3.2 Perform NER for chunks of text. TLO 3.3 Perform dependency parsing to construct dependency tree. TLO 3.4 Perform constituency-based parsing to generate parse trees.	**Unit - III Text Syntax and Structure** 3.1 Part-of-speech tagging : English Word Classes ,Part-of-Speech Tagging 3.2 Named entity recognition :Named Entities and Named Entity Tagging ,IOB/ BIO tagging 3.3 Parsing Techniques :Partial parsing/chunking ,Dependency parsing	Lecture Using Chalk-Board Demonstration Hands-on
4	TLO 4.1 Explain the process of Vector Space Models . TLO 4.2 Apply Cosine Similarity to measure and compare the similarity between word vectors. TLO 4.3 Use TF-IDF vector embedding. TLO 4.4 Use Word2Vec embedding. TLO 4.5 Use contextual embedding.	**Unit - IV Text Feature Extraction** 4.1 Vector Space Models :Words and Vectors , Cosine for measuring similarity 4.2 One hot encoding ,Bag-of-words ,TF-IDF 4.3 Word2vec:continuous bag of words, skip gram 4.4 Contextual Embeddings : Contextual Embeddings ,Word Sense	Lecture Using Chalk-Board Demonstration Hands-on

NATURAL LANGUAGE PROCESSING **Course Code : 315329**

Sr.No	Theory Learning Outcomes (TLO's)aligned to CO's.	Learning content mapped with Theory Learning Outcomes (TLO's) and CO's.	Suggested Learning Pedagogies.
5	TLO 5.1 Explain different phases of implementing Sentiment analysis. TLO 5.2 Explain Need of Transformer. TLO 5.3 Explain Transfer Learning. TLO 5.4 Use Hugging face Transformer. TLO 5.5 Enlist challenges in Transformer.	**Unit - V NLP Application and Transformers** 5.1 NLP application for text classification :NLP application pipeline ,Evaluating Classification Model, Develop a text classification application 5.2 Transformers :Encoder-Decoder Framework ,Attention Mechanisms ,Transformer architecture ,Challenges with Transformers 5.3 Transfer Learning in NLP 5.4 Hugging Face Transformers : Transformer Applications ,The Hugging Face Hub, Hugging Face Tokenizers ,Hugging Face Datasets ,Hugging Face Accelerate	Lecture Using Chalk-Board Demonstration Hands-on

VI. LABORATORY LEARNING OUTCOME AND ALIGNED PRACTICAL / TUTORIAL EXPERIENCES.

Practical / Tutorial / Laboratory Learning Outcome (LLO)	Sr No	Laboratory Experiment / Practical Titles / Tutorial Titles	Number of hrs.	Relevant COs
LLO 1.1 Use different type of Text corpus.	1	Implement a program to use text corpus. i)Brown corpus, ii)Penn Treebank Corpus.	2	CO1
LLO 2.1 Use Regular Expression LLO 2.2 Word Segmentation using Re and nltk.	2	i)Write program for Sentence Segmentation Techniques. ii)Write program Word Segmentation using Re and nltk.	2	CO2
LLO 3.1 Apply tokenization on text .	3	Implement Penn Treebank tokenization, word_tokenize,wordpunct_tokenize,sent_tokenize,WhitespaceTokenizer.	2	CO2
LLO 4.1 Use various Stemmer and Lemmatizer for text processing .	4	*Apply various Lemmatization Techniques and Stemming Techniques such as porter stemmer,Lancaster Stemmer,Snowball Stemmer on text.	2	CO2
LLO 5.1 Apply text normalization techniques LLO 5.2 Generate unigram ,bigram, trigram.	5	*Write program on Text Normalization using nltk: i)Tokenizing text ii)Removing special charactersiii)Expanding contractions iv)case conversion. ii)*Generate unigram, bigram, trigram for given text.	2	CO2
LLO 6.1 Perform POS using ntlk or spacy.	6	*Write program for POS tagging on the given text.	2	CO3

NATURAL LANGUAGE PROCESSING **Course Code : 315329**

Practical / Tutorial / Laboratory Learning Outcome (LLO)	Sr No	Laboratory Experiment / Practical Titles / Tutorial Titles	Number of hrs.	Relevant COs
LLO 7.1 Implement Named Entity Recognizer.	7	*Write program to find Named Entity Recognition(NER) for the given text .	2	CO3
LLO 8.1 Generate Dependency parse tree LLO 8.2 Perform chunking on text.	8	i)Implement a program for dependency parse tree on the sentence using nltk or spacy. ii)Write program for performing chunking on the given text .Extract Noun Phrases, Verb Phrases, Adjective Phrases.	2	CO3 CO4
LLO 9.1 Use Word Embedding.	9	*Write program to generate word embedding using word2Vec and BERT embedding(use Hugging Face).	2	CO4
LLO 10.1 Sentiment analysis. LLO 10.2 Detect fake content.	10	Perform the prediction task using NLP and ML classifiers: a)Sentiment analysis b) Fake news detection.	2	CO5
LLO 11.1 Implement text classification using Hugging Face.	11	* Implement program to fine-tune a pre-trained model from Hugging Face's for text classification.	2	CO5

Note : Out of above suggestive LLOs -

- '*' Marked Practicals (LLOs) Are mandatory.
- Minimum 80% of above list of lab experiment are to be performed.
- Judicial mix of LLOs are to be performed to achieve desired outcomes.

VII. SUGGESTED MICRO PROJECT / ASSIGNMENT/ ACTIVITIES FOR SPECIFIC LEARNING / SKILLS DEVELOPMENT (SELF LEARNING)

Micro project

- 1)Sentiment Analysis – Develop a model to classify text as positive, negative, or neutral using NLP techniques. 2)Fake News Detection – Train a classifier to differentiate between real and fake news articles based on linguistic patterns. 3)Keyword Extraction – Extract the most relevant keywords from a document using NLP algorithms like TF-IDF or RAKE.

NATURAL LANGUAGE PROCESSING **Course Code : 315329**

Note :

- Above is just a suggestive list of microprojects and assignments; faculty must prepare their own bank of microprojects, assignments, and activities in a similar way.
- The faculty must allocate judicial mix of tasks, considering the weaknesses and / strengths of the student in acquiring the desired skills.
- If a microproject is assigned, it is expected to be completed as a group activity.
- SLA marks shall be awarded as per the continuous assessment record.
- For courses with no SLA component the list of suggestive microprojects / assignments/ activities are optional, faculty may encourage students to perform these tasks for enhanced learning experiences.
- If the course does not have associated SLA component, above suggestive listings is applicable to Tutorials and maybe considered for FA-PR evaluations.

VIII. LABORATORY EQUIPMENT / INSTRUMENTS / TOOLS / SOFTWARE REQUIRED

Sr.No	Equipment Name with Broad Specifications	Relevant LLO Number
1	Computer system - (Computer System which is available in lab with 4GB RAM)	All
2	Python 3.7 onwards	All
3	Colab	All

IX. SUGGESTED WEIGHTAGE TO LEARNING EFFORTS & ASSESSMENT PURPOSE (Specification Table)

Sr.No	Unit	Unit Title	Aligned COs	Learning Hours	R-Level	U-Level	A-Level	Total Marks
1	I	Natural Language Basics	CO1	4	4	4	4	12
2	II	Text Normalization and Preprocessing	CO2	10	2	8	6	16
3	III	Text Syntax and Structure	CO3	8	2	2	12	16
4	IV	Text Feature Extraction	CO4	10	2	4	8	14
5	V	NLP Application and Transformers	CO5	8	2	4	6	12
Grand Total				**40**	**12**	**22**	**36**	**70**

X. ASSESSMENT METHODOLOGIES/TOOLS

Formative assessment (Assessment for Learning)

- Two unit tests of 30 marks each conducted during the semester.
- Continuous assessment based on process and product related performance indicators. Each practical will be assessed considering 60% weightage to process, 40% weightage to product. A continuous assessment based term work.

Summative Assessment (Assessment of Learning)

- End semester examination, Lab performance, Viva voce

XI. SUGGESTED COS - POS MATRIX FORM

09-04-2025 02:35:25 PM

NATURAL LANGUAGE PROCESSING **Course Code : 315329**

Course Outcomes (COs)	Programme Outcomes (POs)							Programme Specific Outcomes* (PSOs)		
	PO-1 Basic and Discipline Specific Knowledge	PO-2 Problem Analysis	PO-3 Design/ Development of Solutions	PO-4 Engineering Tools	PO-5 Engineering Practices for Society, Sustainability and Environment	PO-6 Project Management	PO-7 Life Long Learning	PSO-1	PSO-2	PSO-3
CO1	3	-	-	1	1	-	-			
CO2	2	3	2	3	1	-	1			
CO3	2	3	2	3	1	-	1			
CO4	2	3	2	3	1	-	1			
CO5	2	3	3	3	2	2	1			

Legends :- High:03, Medium:02,Low:01, No Mapping: -
*PSOs are to be formulated at institute level

XII. SUGGESTED LEARNING MATERIALS / BOOKS

Sr.No	Author	Title	Publisher with ISBN Number
1	Daniel Jurafsky	Speech and Language Processing -ch2 2.1,2.3,2.4 ch3,ch4	Pearson Publication ISBN :978-0131873216
2	Dipanjan Sarkar	Text Analytics with Python ch1 ch5 5.1	Apress ISBN-13 (pbk): 978-1-4842-2387-1
3	Steven Bird, Ewan Klein, and Edward Loper	Natural Language Processing with python ch5 5.2 5.3 5.4	Oreally ISBN:978-0-596-51649-9
4	Akshay Kulkarni Adarsha Shivananda	Natural Language Processing Recipes_ Unlocking Text Data with Machine Learning and Deep Learning using Python . for lab 1 to 11	Apress ISBN-13 (pbk): 978-1-4842-4266-7
5	Pushpak Bhattacharyya and Aditya Joshi	Natural Language Processing ch2-2.5 ch3-3.3	Willey ISBN:978-93-5746-238-9

XIII . LEARNING WEBSITES & PORTALS

Sr.No	Link / Portal	Description
1	https://web.stanford.edu/~jurafsky/slp3/	NLP e-book and PPT
2	https://github.com/Donges-Niklas/Intro-to-NLP-with-NLTK/blob /master/nltk.ipynb	Text Segmentation, Stop Words & Word Segmentation, Stemming ,Parsing (Speech Tagging & Chunking),programs
3	https://github.com/samiramunir/Simple-Sentiment-Analysis-usi ng-NLTK/blob/master/live_classifier.py	sentiment Analysis Program
4	https://www.youtube.com/watch?v=yLDRHyNJSXA&list=PLPIwNooIb9 vimsumdWeKF3BRzs9tJ-_gy&index=38	Sentiment Analysis theory content
5	https://www.youtube.com/watch?v=fM4qTMfCoak&list=PLZoTAELRMX VMdJ5sqbCK2LiM0HhQVWNzm	NLP concept playlist

Note :

- Teachers are requested to check the creative common license status/financial implications of the suggested online educational resources before use by the students

09-04-2025 02:35:25 PM

NATURAL LANGUAGE PROCESSING **Course Code : 315329**

MSBTE Approval Dt. 24/02/2025 **Semester - 5, K Scheme**