



Government of Maharashtra
GOVERNMENT POLYTECHNIC, NAGPUR
 (An Autonomous Institute of Govt. of Maharashtra)
 Near Mangalwari Bazar, Sadar, Nagpur-440001



COURSE CURRICULUM

Program	:	Diploma in AI/IT
Course Category	:	DSE
Course Code	:	AI303H
Course Title	:	Natural Language Processing

I Rationale:

The subject provides a comprehensive introduction to core NLP techniques. Language Modelling are introduced by explaining models like n-grams approaches for understanding word sequences. Next, Naïve Bayes and Text Classification covers essential text classification techniques, highlighting its use in applications. Vector Semantics focuses on retrieving relevant information from large datasets. Annotating Linguistic Structure teaches Context Free Grammar. The course concludes with Applications of NLP. This structure equips students with theoretical knowledge and practical skills, preparing them to build NLP solutions for various industries.

II Industry Identified Competency:

The Student will be able to do in Industry at entry level:

Work with NLP techniques for text and language understanding.

III Course Outcomes (COs):

After completing this course students will be able to:

CO1: Construct Regular Expressions based on the given data.

CO2: Develop programs for language models, tokenization and smoothing.

CO3: Implement the Naïve Bayes algorithm to classify data and evaluate its performance.

CO4: Develop programs for computing text similarity using given measures.

CO5: Develop programs for deriving sentences using Context Free Grammar.

CO6: Design any applications using Chatbots.

IV Learning Scheme:

Classroom Learning (CL)	Tutorial Learning (TL)	Laboratory Learning (LL)	Self-Study Learning (SL)	Notional Learning Hours (NLH)	Credits
3	-	2	1	6	3

V Assessment Scheme:

Classroom Learning				Tutorial/Laboratory Learning				Self-Study Learning	
FA	SA	Total		FA		SA		SA	
Max	Max	Max	Min	Max	Min	Max	Min	Max	Min
30	70	100	40	25	10	25+	10	25	10

VI Classroom Learning Content:

Unit No	Specific Learning Outcomes (SLO) (In Cognitive Domain)		Topics and Sub-topics		Aligned Cos
UNIT I: Introduction	1a	Construct regular expression for the given statement.	1.1	Introduction to NLP, Regular Expressions, Basic Regular Expression Patterns, Disjunction, Grouping, and Precedence, A Simple Example, More Operators, A More Complex Example	CO1
	1b	Compute number of types using Herdan's Law.	1.2	Words, utterance, disfluency, lemma, wordform, word type, word token, Herdan's Law, Corpora, code switching, datasheet	
	1c	Perform Lemmatization using porter stemmer based on the given data.	1.3	Text Normalization, Word Tokenization, Byte-Pair Encoding for Tokenization, Word Normalization, Lemmatization and Stemming, The Porter Stemmer, Sentence Segmentation	
	1d	Calculate Minimum Edit Distance for given data.	1.4	Minimum Edit Distance, The Minimum Edit Distance Algorithm	

UNIT II: N-gram Language Models	2a	Identify the given N-gram based on the given data.	2.1	N-gram Language Models, N-grams, bigram, Markov, maximum likelihood estimation	CO2
	2b	Differentiate extrinsic evaluation and intrinsic evaluation based on the given points.	2.2	Evaluating Language Models, extrinsic evaluation, intrinsic evaluation, training set, test set, development test set, Perplexity	
	2c	Compare generalization and zeroes based on the given points.	2.3	Sampling sentences from a language model, Generalization and Zeros	
	2d	Calculate the given term using given smoothing technique.	2.4	Smoothing, Laplace Smoothing, discounting, Add-k smoothing, Backoff and Interpolation, Advanced: Kneser-Ney Smoothing	
UNIT III: Naive Bayes and Text Classification	3a	Justify the Naive Bayes Classifier equation with given terms.	3.1	Naive Bayes Classifiers	CO3
	3b	Predict the given class using Naive Bayes Classifier	3.2	Training the Naive Bayes Classifier , Worked example	
	3c	Find the probability of given text using Naive Bayes Language Model	3.3	Naive Bayes for other text classification tasks, Naive Bayes as a Language Model	
	3d	Compute the given term based on given data.	3.4	Evaluation: Confusion Matrix, Accuracy, Precision, Recall, F-measure, Evaluating with more than two classes, Test sets and Cross-validation	
	3e	Justify Statistical Significance Testing with given terms.	3.5	Statistical Significance Testing, Avoiding Harms in Classification	
ics and	4a	Compare given vectors based on given points	4.1	Lexical Semantics, Vector Semantics, Words and Vectors	

UNIT IV: Vector Semant Embeddings	4b	Compute the Cosine for measuring similarity using given technique for given data.	4.2	Cosine for measuring similarity, TF-IDF: Weighing terms in the vector, Pointwise Mutual Information (PMI)	CO4
	4c	Justify the given terms in Wrod2Vec with respet to given parameters.	4.3	Applications of the tf-idf or PPMI vector models, Word2vec, Visualizing Embeddings, Semantic properties of embeddings	
UNIT V: Annotating Linguistic Structure	5a	Identify the given grammar using given relations.	5.1	Context-Free Grammars and Constituency Parsing: Constituency, Context-Free Grammars.	CO5
	5b	Compare given normal forms based on given parameters.	5.2	Treebanks, Grammar Equivalence and Normal Form, Ambiguity	
	5c	Convert given grammar into given Normal Form.	5.3	CKY Parsing: A Dynamic Programming Approach, Conversion to Chomsky Normal Form	
	5d	Classify the given phrase into given term.	5.4	Dependency Parsing: Dependency Relations, Dependency Formalisms, Projectivity, Dependency Treebanks	
	5e	Justify function for given Dependency Parser.	5.5	Transition-Based Dependency Parsing	
UNIT VI: Applications of NLP	6a	Justify the given terms.	6.1	Question Answering and Information Retrieval: Information Retrieval, IR-based Factoid Question Answering	CO6
	6b	Distinguish given Question Answering based on given points.	6.2	Knowledge-based Question Answering, Evaluation of Factoid Answers	
	6c	Identify different properties of Human Conversation	6.3	Chatbots & Dialogue Systems: Properties of Human Conversation, Chatbots	
	6d	Draw architecture of given system.	6.4	GUS: Simple Frame-based Dialogue Systems	
	6e	Classify the given dialogue system based on the following points.	6.5	The Dialogue-State Architecture, Evaluating Dialogue Systems	

VII Laboratory Learning Content:

Unit No	Specific Learning Outcomes (SLO) (In Psychomotor Domain)		Hours	Aligned COs
I	1	Develop and execute Regular Expressions based on the given data using following URL https://regex101.com or any other regex processing tool.	2#	CO1
I	2	Open account in Google Colab or Install Python and explore GUI for executing programs.	2#	CO1
I	3	Import NLTK in python editor, download the 'stopwords' and 'punkt' packages and execute different commands based on the given packages.	2#	CO1
II	4	Import spacy, load the language model and execute different commands based on the given packages.	2#	CO2
II	5	Implement a program to tokenize a given text and to get sentences from text document.	2#	CO2
II	6	Implement a program for listing the words using N-grams.	2#	CO2
II	7	Implement a program for any one smoothing technique.	2	CO2
III	8	Implement a program for various evaluation metrics such as Accuracy, precision, recall, etc.	2#	CO3
III	9	Implement a program for Naive Bayes Classifier with NLTK.	2#	CO3
IV	10	Implement a program for text similarity using Cosine and TF-IDF algorithm with NLTK.	2#	CO4
IV	11	Develop a program for demonstrating word2vec or PMI with NLTK.	2	CO4
V	12	Develop a program to define Context Free Grammar and show parsing on simple sentence using NLTK.	2#	CO5
V	13	Develop a program to define own Context Free Grammar and show parse Tree using NLTK.	2#	CO5
VI	14	Develop a program for IR-based Question Answering.	2#	CO6
VI	15	Design a program for building simple Chatboat using NLTK library.	2	CO6

Note: # Compulsory

VIII Self-Study Learning (SLO in Cognitive/Psychomotor/Affective Domain)

1. Prepare microproject on Regular Expressions with examples.
2. Prepare microproject on creating Words and Corpora for any real life problem.
3. Prepare microproject on Normalization and Minimum Edit Distance based on real life problem.
4. Prepare microproject on N-gram language model.
5. Prepare microproject on different smoothing techniques.
6. Prepare microproject on Naive Bayes Classifiers with suitable dataset.
7. Prepare microproject on standard evaluating measures like accuracy, precision, recall, f-measure.
8. Prepare microproject on Vector semantics, words and vectors using any dataset.
9. Prepare microproject on Cosine, TF-IDF, PMI and PPMI using four real life problems.
10. Prepare microproject on Context-Free Grammars with five different examples.
11. Prepare microproject on Grammar Equivalence and Normal Forms with examples.
12. Prepare microproject on Grammar ambiguity and conversion of Normal Forms with examples.
13. Prepare microproject on IR-based Factoid Question Answering.
14. Prepare microproject on Knowledge-based Question Answering.
15. Prepare microproject based on any topic relevant with this curriculum.

IX Specification Table for Classroom Learning Assessment:

Unit No.	Units	Classr oom Learn ing Hours	C/O	Levels from Cognition Process Dimension			Total Marks
				R	U	A	
1	Introduction	6	C	02	08	00	10
			O	00	00	04	4
2	N-gram Language Models	8	C	02	04	04	10
			O	00	00	06	6
3	Naive Bayes and Text Classification	8	C	02	00	10	12
			O	00	04	06	10
4	Vector Semantics and Embeddings	8	C	04	00	10	14
			O	02	04	00	6
5	Annotating Linguistic Structure	8	C	02	00	12	14
			O	04	00	04	8
6	Applications of NLP	7	C	02	08	00	10
			O	02	04	00	6
Total		45	C	14	20	36	70
			O	8	12	20	40

*C - Compusory

O - Optional

*R – Remember

U – Understand

A – Analyze / Apply

X Question Paper Format for Summative Assessment (SA):

Q. No.	Bit 1	Bit 2	Bit 3	Bit 4	Bit 5	Bit 6	Bit 7	Options	
	TLM	TLM	TLM	TLM	TLM	TLM	TLM	C	O
1	1R2	2R2	3R2	5R2	6R2	4R2	6R2	5	7
2	1U4	2U4	4R4	5R4	3U4			3	5
3	3A4	4A4	6U4	1A4	5A4			3	5
4	2A4	1U4	6U4	6U4	4U4			3	5
5	5A6	4A6	3A6					2	3
6	5A6	3A6	2A6					2	3

*T- Unit/Topic Number

L- Level of Question

M- Marks

*R – Remember

U – Understand

A – Analyze / Apply

*1R2 means Unit Number No- 1, Level of Question –Remember, Marks – 2 Marks

XI Scheme of Laboratory Formative Assessment (FA):

S.N.	Criteria	Max. Marks
1	Write algorithm or Draw flow chart	5
2	Performance	10
3	Execution and result	5
4	Viva Voce	5
TOTAL		25

XII Scheme of Self-Learning Summative Assessment (SA):

S.N.	Criteria	Max. Marks
1	Content Management and Formatting	10
2	Algorithm / Flowchart / Steps of solution	5
3	Execution and results	5
4	Viva Voce	5
TOTAL		25

XIII COs-POs/PSOs Mapping Matrix:

Course Outcomes	Program Outcomes							Program Specific	
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PSO1	PSO2
C01	3	-	-	-	-	-	-	-	-
C02	3	2	2	3	-	-	-	-	3
C03	3	2	2	3	-	-	-	-	3
C04	3	-	-	3	-	-	-	-	3
C05	3	2	2	3	-	-	-	-	3
C06	3	2	2	3	-	-	-	-	3

XIV Textbooks, BIS Codes References:

S.N.	Title	Author, Publisher, Edition and Year of Publication	ISBN Number
1	Speech and Language Processing	Dan Jurafsky and James Martin, 3rd Edition , Pearson Publication, 2023	9781593279288
2	Natural Language Processing with Python: Analysing Text with the Natural Language Toolkit	Steven Bird, Ewan Klein, Edward Loper, O'Reilly Publication	9780596158064
3	Natural Language Understanding	Allen James, Lebanon, Second Edition , Indian Publication	9781491919538
4	Foundations of Statistical Natural Language Processing	Manning, Christopher and Heinrich, Schutze, MIT Press, 1999	9781735467221

XV e-References:

1. <https://web.stanford.edu/~jurafsky/slp3/> , accessed on 18/09/2024
2. <https://www.simplilearn.com/tutorials/machine-learning-tutorial/naive-bayes-classifier> , accessed on 18/09/2024
3. <https://pythonprogramming.net/part-of-speech-tagging-nltk-tutorial/> , accessed on 20/09/2024
4. <https://pythonprogramming.net/naive-bayes-classifier-nltk-tutorial/> , accessed on 20/09/2024
5. <https://stackoverflow.com/questions/42454072/qa-query-system-on-corpus> , accessed on 21/09/2024
6. <https://medium.com/@batuhansenoglu/introduction-to-question-answering-23117be9f8c8> , accessed on 21/09/2024
7. https://spotintelligence.com/2022/12/19/text-similarity-python/#1_Text_similarity_with_NLTK , accessed on 21/09/2024
8. <https://stackoverflow.com/questions/8897593/how-to-compute-the-similarity-between-two-text-documents> , accessed on 21/09/2024
9. https://jofrhwd.github.io/teaching/courses/2022_lin517/lectures/ngram/02_smoothering.html , , accessed on 22/09/2024
10. <https://regex101.com/r/ry8aTb/1> , accessed on 22/09/2024

XVI List of Major Equipment/Machineries with Specification:

1. Computer System
2. COLAB compiler (Open Source) or Python

XVII List of Industry Experts and Faculties who contributed for this curriculum:

S.N.	Name	Designation	Institute / Industry
1	Dr. A. R. Mahajan	HoD Information Technology	Government Polytechnic, Nagpur
2	Dr. Paresh R. Kamble	Machine Learning specialist	OZ SportZ, Nagpur, HQ-Reykjavik, Iceland
3	Dr. Vijaya Balpande	Associate Professor	Priyadarshini College of Engg., Nagpur
4	Mrs. Shifa Azharuddin	Lecturer in IT	Government Polytechnic, Nagpur
5	Dr. S. S. Gharde	Lecturer in IT	Government Polytechnic, Nagpur
6	Mr. A. G. Barsagade	Lecturer in IT	Government Polytechnic, Nagpur



(Dr. A. R. Mahajan)
HOD & Chairman PBOS, AI



(Dr. G. V. Gotmare)
Member Secretary PBOS, AI