NATURAL LANGUAGE PROCESSING

(Effective from the Academic Year 2022 - 2023)

VI SEMESTER

Course Code	**AM622I2A**	CIA Marks	50
Number of Contact Hours/Week (L: T: P: S)	3:0:2:0	SEE Marks	50
Total Hours of Pedagogy	40L + 20P	Exam Hours	03

CREDITS – 4

COURSE PREREQUISITES:

- Fundamentals of Automata Theory and Basic knowledge of English Grammar.

COURSE OBJECTIVES:

- Define the natural language and analyze the importance of natural language.
- Analyze spelling error detection and correction methods and parsing techniques in NLP.
- Understand the Applications of natural language processing.
- Illustrate the information retrieval models in natural language processing.

TEACHING - LEARNING STRATEGY:

Following are some sample strategies that can be incorporate for the Course Delivery

- Chalk and Talk Method/Blended Mode Method
- Power Point Presentation
- Expert Talk/Webinar/Seminar
- Video Streaming/Self-Study/Simulations
- Peer-to-Peer Activities
- Activity/Problem Based Learning
- Case Studies
- MOOC/NPTEL Courses
- Any other innovative initiatives with respect to the Course contents

COURSE CONTENTS

MODULE - I

Overview and language modeling: Overview: Origins and challenges of NLP-Language and Grammar- Processing Indian Languages- NLP Applications, **Language Modeling**: Statistical Language Model- N-gram model- (**unigram, bigram**), Paninion Framework, Karaka theory, Smoothing Technique.	**8 Hours**

MODULE - II

Word Level Analysis: Regular Expressions, Finite State Automata, Morphological Parsing, Spelling Error Detection and Correction, Words and Word Classes-Part-of Speech Tagging. **Syntactic Analysis:** Context-free Grammar, Constituency, top-down and bottom-up Parsing, CYK parsing.	**8 Hours**

MODULE - III	
Naive Bayes and Sentiment Classification: Naive Bayes Classifiers, Training the Naive Bayes Classifier, worked example, Optimizing for Sentiment Analysis, Naive Bayes for other text classification tasks, Naive Bayes as a Language Model.	**8 Hours**
MODULE - IV	
Information Retrieval and Lexical Resources: Information Retrieval: Design features of Information Retrieval Systems-Classical, Non-classical, Alternative Models of Information Retrieval- Custer model, Fuzzy model, LSTM model**, Major Issues in Information Retrieval.** **Lexical Resources**: World Net, Frame Net, Stemmers, POS Tagger- Research Corpora.	**8 Hours**
MODULE - V	
Machine Translation: Recurrent Neural Networks, The LSTM, Machine Translation using Encoder-Decoder, Details of the Encoder-Decoder Model, Translating in low-resource situations, MT Evaluation, Bias, and Ethical Issues.	**8 Hours**

COURSE OUTCOMES

Upon completion of this course, the students will be able to:

CO No.	Course Outcome Description	Bloom's Taxonomy Level
CO1	Discuss the concepts of NLP and demonstrate the statistical-based language models and smoothing techniques	CL3
CO2	Demonstrate morphological analysis and parsing using finite-state transducers	CL3
CO3	Apply the Naïve Bayes classifier and sentiment analysis for Natural language problems and text classifications.	CL3
CO4	Illustrate Information Retrieval in the context of NLP and discuss the lexical dictionaries.	CL3
CO5	Develop the Machine Translation applications using the Encoder and Decoder model.	CL3

LABORATORY COMPONENTS

Exp. No.	Experiment Description	CO No.	Bloom's Taxonomy Level
1	Consider the following Corpus of three sentences a) There is a big garden. b) Children play in a garden c) They play inside a beautiful garden Calculate P for the sentence "They play in a big Garden" assuming a bi-gram language model.	CO1	CL3

2	Find the bigram count for the given corpus. Apply Laplace smoothing and find the bigram probabilities after add-one smoothing (up to 4 decimal places)	CO1	CL3
3	Implement rule-based tagger and stochastic tagger for the give corpus of sentences.	CO2	CL3
4	Implement top-down and bottom-up parsing using python NLTK.	CO2	CL3
5	Given the following short movie reviews, each labeled with a genre, either comedy or action: a) fun, couple, love, love : **comedy** b) fast, furious, shoot : **action** c) couple, fly, fast, fun, fun :**comedy** d) furious, shoot, shoot, fun :**action** e) fly, fast, shoot, love :**action** and a new document D: ***fast, couple, shoot, fly*** Compute the most likely class for D. Assume a naive Bayes classifier and use add-1 smoothing for the likelihoods.	CO3	CL3
6	The dataset contains following 5 documents. a) D1: "Shipment of gold damaged in a fire" b) D2: "Delivery of silver arrived in a silver truck" c) D3: "Shipment of gold arrived in a truck" d) D4: "Purchased silver and gold arrived in a wooden truck" e) D5: "The arrival of gold and silver shipment is delayed." Find the top two relevant documents for the query document with the content "**gold silver truck** " using the vector space model and latent semantic space model. Use the following similarity measure and analyze the result. a) Euclidean distance b) Manhattan distance c) Cosine similarity d) Jaccard similarity e) Dice Similarity Coefficient	CO3	CL3
7	Extract Synonyms and Antonyms for a given word using WordNet.	CO4	CL3
8	Implement a machine translator for 10 sentences using an encoder-decoder model for any two languages.	CO5	CL3

CO-PO-PSO MAPPING

CO No.	Programme Outcomes (PO)												Programme Specific Outcome (PSO)	
	1	2	3	4	5	6	7	8	9	10	11	12	1	2
CO1	2	2	2	1	1	1						2		2
CO2	2	2	2	1								2		2
CO3	2	2	2	1	1	1						2		2
CO4	2	2	2	1	1	1						2		2
CO5	2	2	2	1	1	1						2		2

3: Substantial (High) **2: Moderate (Medium)** **1: Poor (Low)**

ASSESSMENT STRATEGY

Assessment will be both CIA and SEE. Students learning will be assessed using Direct and Indirect methods:

Sl. No.	Assessment Description	Weightage (%)	Max. Marks
1	**Continuous Internal Assessment (CIA)**	**100 %**	**50**
	Continuous Internal Evaluation (CIE)	60 %	30
	Practical Session (Laboratory Component)	40 %	20
2	**Semester End Examination (SEE)**	**100 %**	**50**

<table>
<tr><td colspan="4">ASSESSMENT DETAILS</td><td></td></tr>
<tr><td colspan="3">Continuous Internal Assessment (CIA) (50%)</td><td>Practical Sessions (40%)</td><td rowspan="3">Semester End Exam (SEE) (50%)</td></tr>
<tr><td colspan="3">Continuous Internal Evaluation (CIE) (60%)</td><td rowspan="2"></td></tr>
<tr><td>I</td><td>II</td><td>III</td></tr>
<tr><td colspan="3">Syllabus Coverage</td><td>Syllabus Coverage</td><td>Syllabus Coverage</td></tr>
<tr><td>40%</td><td>30%</td><td>30%</td><td>100%</td><td>100%</td></tr>
<tr><td>MI</td><td></td><td></td><td>MI</td><td>MI</td></tr>
<tr><td>MII</td><td>MII</td><td></td><td>MII</td><td>MII</td></tr>
<tr><td></td><td>MIII</td><td></td><td>MIII</td><td>MIII</td></tr>
<tr><td></td><td></td><td>MIV</td><td>MIV</td><td>MIV</td></tr>
<tr><td></td><td></td><td>MV</td><td>MV</td><td>MV</td></tr>
</table>

NOTE:

- Assessment will be both CIA and SEE.
- The practical sessions of the IPCC shall be for CIE only.
- The Theory component of the IPCC shall be for both CIA and SEE respectively.
- The questions from the practical sessions shall be included in Theory SEE.

Note: For Examinations (both CIE and SEE), the question papers shall contain the questions mapped to the appropriate Bloom's Level. Any COs mapped with higher cognitive Bloom's Level may also be assessed through the assignments.

SEE QUESTION PAPER PATTERN:

1. The question paper will have **TEN** full questions from **FIVE** Modules
2. There will be 2 full questions from each module. Every question will carry a maximum of 20 marks.
3. Each full question may have a maximum of four sub-questions covering all the topics under a module.
4. The students will have to answer FIVE full questions, selecting one full question from each module.

TEXT BOOKS:

1. Tanveer Siddiqui, U.S. Tiwary, "Natural Language Processing and Information Retrieval", Oxford University Press, 2008.
2. D. Jurafsky, J. H. Martin, "Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3e)", Pearson Education, 2023.

REFERENCE BOOKS:

1. Akshay Kulkarni, Adarsha Shivananda, "Natural Language Processing Recipes - Unlocking Text Data with Machine Learning and Deep Learning using Python", Apress, 2019
2. James Allen, "Natural Language Understanding", 2nd edition, Benjamin/Cummings publishing company, 1995.
3. Gerald J. Kowalski and Mark.T. Maybury, "Information Storage and Retrieval systems", Kluwer Academic Publishers, 2000