SARVAJANIK UNIVERSITY

Sarvajanik College of Engineering and Technology

Master of Computer Applications

MCA Semester III

Subject Name: Natural Language Processing **Subject Code:** MTCA14309

Type of course: Professional Elective Course

Prerequisite: Data Structure, Theory of Computation

List of Courses where this course will be prerequisite:
Machine Translation, Speech Processing

Rationale: To understand the algorithms available for the processing of linguistic information and computational properties of natural languages. To conceive basic knowledge on various morphological, syntactic and semantic NLP tasks. To develop systems for various NLP problems and familiarize with various publicly available NLP software libraries and data sets to be used for development. This course will boost students' knowledge to a research level where they can work on real world computational linguistic problems.

Teaching and Examination Scheme:

TEACHING SCHEME				Theory Marks			Practical Marks		Total
L	T	P	C	TEE	CA1	CA2	TEP	CA3	
3	0	4	5	60	25	15	60	40	200

CA1: Continuous Assessment (assignments/projects/open book tests/closed book tests **CA2:** Sincerity in attending classes/class tests/ timely submissions of assignments/self-learning attitude/solving advanced problems **TEE:** Term End Examination **TEP:** Term End Practical Exam (Performance and viva on practical skills learned in course) **CA3:** Regular submission of Lab work/Quality of work submitted/Active participation in lab sessions/viva on practical skills learned in course

Content:

Sr. No.	Content	Teaching Hrs.	Module Weightage
1	**Introduction:** NLP – introduction and applications, NLP phases, Difficulty of NLP including ambiguity; Spelling error and Noisy Channel Model; Concepts of Parts-of-speech and Formal Grammar of English. Language Modeling with N-gram, Simple N-gram models, Smoothing (basic techniques), Evaluating language models;	**12**	**28%**
2	**Linguistic essentials:** Lexical syntax- Morphology and Finite State Transducers - Part of speech Tagging - Rule-Based Part of Speech Tagging - Markov Models - Hidden Markov Models	**11**	**24%**
3	**Syntax Analysis:** Syntax Parsing - Grammar formalisms and tree banks - Parsing with Context Free Grammars - Features and Unification -Statistical parsing and probabilistic CFGs (PCFGs)-Lexicalized PCFGs.	**11**	**24%**
4	**Semantics - Representing Meaning** – Semantic Analysis – Vector Semantics; Words and Vector; Measuring Similarity; Semantics with dense vectors; SVD and Latent Semantic Analysis; Embedding from prediction: Skip-gram and CBOW; Concept of Word Sense; Introduction to WordNet	**11**	**24%**

Suggested Specification table with Marks (Theory):

Distribution of Theory Marks					
R Level	U Level	A Level	N Level	E Level	C Level
20	20	15	15	15	15

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom's Taxonomy)

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

Reference Books:

Sr. no.	Title of book /article	Author(s)	Publisher and details like ISBN	Year of publication	Publication Edition
1	Speech and Language Processing	Daniel Jurafsky and James H. Martin	Prentice Hall	2008	2nd Edition
2	Foundations of Statistical Natural Language Processing	Christopher D. Manning and HinrichSchuetze	MIT Press	1999	
3	Natural Language Processing with Python	Steven Bird, Ewan Klein and Edward Loper	O'Reilly Media	2009	1 edition
4	Foundations of Computational Linguistics: Human- Computer Communication in Natural Language, Paperback	Roland R. Hausser	MIT Press	2011	

Course Outcomes:

Sr. No.	CO Statement After learning this subject, students will be able to	Marks % weightage
CO-1	Describe the concepts of morphology, syntax &semantics	20
CO-2	Demonstrate understanding of the relationship between NLP and	20

	statistics & machine learning	
CO-3	Discover various linguistic and statistical features relevant to the basic NLP task, namely, spelling correction, Preprocessing	25
CO-4	Understand and apply morphological analysisparts-of-speech tagging, parsing and semantic analysis.	25
CO-5	Develop systems for various NLP problems with moderate complexity and evaluate.	10

Mapping with POs:

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PS O1	PS O2	PS O3
CO-1	**3**	**3**	**3**	**3**	**2**	**0**	**1**	**1**	**0**	**0**	**1**	**1**			
CO-2	**3**	**3**	**3**	**3**	**2**	**0**	**1**	**1**	**0**	**0**	**1**	**1**			
CO-3	**3**	**3**	**3**	**3**	**2**	**0**	**1**	**1**	**0**	**0**	**1**	**1**			
CO-4	**3**	**3**	**3**	**3**	**2**	**0**	**1**	**1**	**0**	**0**	**1**	**1**			
CO-5	**3**	**3**	**3**	**3**	**2**	**0**	**1**	**1**	**0**	**0**	**1**	**1**			
Rationale*															

Rationale*: Explaining why it is matching this particular program outcome

List of Open learning website:

- https://blog.algorithmia.com/introduction-natural-language-processing-nlp/
- https://www.udacity.com/course/natural-language-processing-nanodegree--nd892

- https://www.coursera.org/learn/language-processing https://towardsdatascience.com/a-practitioners-guide-to-natural-languageprocessing-part-i-processing-understanding-text-9f4abfd13e72
- https://www.edx.org/course/natural-language-processing

List of Open Source Software:

FOR LAB SESSIONS:

List of Experiments:

Students have to develop projects, which should center on the analysis of a specific data source and include at least some of the methods covered in the course. Final projects that employ new and/or unique datasets and reach innovative conclusions will receive the highest scores. A full written explanation of the scope and utility of the project, at least 2 pages in length, will be required by the due date of the final project. All project coding and use of data sources will be closely reviewed, and the potential impact of the project will play a major role in its assessment.

Some examples of Projects

Text Classification, Summarization, Topic Modeling, Document retrieval, WSD Applications, and Sentiment Analysis