Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division

Digital Learning Handout

Part A: Content Design

Course Title	Natural Language Processing
Course No(s)	AIMLCZG530
Credit Units	4
Credit Model	3-1-0
Course Author	Prof. Chetana Gavankar
Version No:	2.0
Date:	5/02/2025

Course Description:

Natural Language Understanding and Generation, N-gram and Neural Language Models, Introduction to LLM, Introduction to prompt engineering, Word to Vectors / Word Embedding (Skip gram/CBOW, BERT), Part of Speech Tagging, Parsing, Word Sense Disambiguation, Semantic Web and Knowledge Graph, Introduction to Retrieval Augmented Generation (RAG)

Course Objectives

No	Course Objective
CO1	To identify and recall fundamental concepts and techniques in Natural Language Processing
CO2	To explain the computational properties of natural languages and articulate the algorithms commonly utilized for processing linguistic information.
CO3	To apply basic mathematical models and methods in NLP applications to solve practical problems
CO4	To examine research and development efforts in Natural Language Processing to discern trends, challenges, and opportunities.

Text Book(s):

T1	Jurafsky and Martin, SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, McGraw Hill, Jan 2025
T2	Manning and Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA

Reference Book(s) & other resources:

R1	Allen James, Natural Language Understanding
R2	Neural Machine Translation by Philipp Koehn

Learning Outcomes: Students will be able to

LO1	Should have a good understanding of the field of natural language processing.
LO2	Should have knowledge of important techniques like language modelling, parsing, used in natural language processing
LO3	Students will be able to analyse language modelling algorithms for real world problems
LO4	Should be able to apply NLP algorithms with deep learning algorithms for applications

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Modular Content Structure

1. **Natural Language Understanding and Generation**
   - Applications of Natural Language proessing
   - Evaluating Language Understanding Systems
   - The Different Levels of Language Analysis
   - The Organization of Natural Language Systems

2. Vector semantics and Embedding
   - Lexical and Vector semantics
   - Word and Vectors
   - TFIDF, Word2Vec, Skip gram and CBOW, Glove

3. N-gram Language Modelling
   - N-Grams, Smoothing
   - Evaluating Language Models

4 Neural Network and Neural Language Modelling
   - Feed-Forward Neural Networks
   - Training Neural Nets for language models
   - Neural Language Models

5. Introduction to LLM and Prompt Engineering
   - Introduction to LLM
   - Introduction to prompt engineering

6. Part-of-Speech Tagging
   - Part-of-Speech Tagging
   - Markov Chains , Hidden Markov Model
   - HMM Part-of-Speech Tagging

7. **Statistical, ML and Neural models of POS tagging**
   - Likelihood Computation: The Forward Algorithm
   - Decoding: The Viterbi Algorithm
   - Maximum Entropy Markov Model, Bidirectionality
   - Neural network models for POS

8. Parsing
   - A Bottom-Up Chart Parser
   - Probabilistic Context-Free Grammars
   - Probabilistic CKY Parsing of PCFGs

10. Dependency Parsing

- Dependency Relations
- Transition-Based Dependency Parsing
- Graph-Based Dependency Parsing
- Dependency parser using neural network

11. Encoder-Decoder Models, Attention and Contextual Embeddings

- Neural Language Models and Generation
- Encoder-Decoder Networks, Attention
- Applications of Encoder-Decoder Networks
- Self-Attention and Transformer Networks
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Contextual Word Representations

12. Word sense disambiguation

- Word Senses
- Relations between Senses
- WordNet: A Database of Lexical Relations
- Word Sense Disambiguation

13. Semantic web ontology and Knowledge Graph

- Introduction to semantic web
- Semantic web ontology, Semantic web languages
- Ontology Engineering ,Ontology Learning
- Knowledge graph –construction of graph

14. Retrieval Augmented Generation

- Large language model (LLM) enhancement
- Accessing external knowledge bases,
- Utilizing specific applications like Semantic search, chatbot or knowledge base enrichment

15. State of art applications

- Text Summarization
- Discussion on other state of art application

Part B: Learning Plan

Contact Session	List of Topic Title	Sub-Topics	Reference
1	**Natural Language Understanding and Generation**	• The Study of Language. • Applications of Natural Language Understanding.	T2 and other online references

		• Evaluating Language Understanding Systems. • The Different Levels of Language Analysis. • The Organization of Natural Language Understanding Systems.	
2	**Vector semantics and Embedding**	• Lexical semantics • Vector semantics • Word and Vectors • TFIDF • Word2Vec, Skip gram and CBOW • Glove • Visualizing Embedding's	T1 and lecture notes
3	**N-gram Language Modelling**	• N-Grams • Generalization and Zeros. • Smoothing • The Web and Stupid Back off • Evaluating Language Models	T1, Chapter 3 and other online references
4	**Neural network and neural language modelling**	• Feed-Forward Neural Networks • Training Neural Nets for language models • Neural Language Models	R2, Chapter 4
5	**Introduction to LLM and Prompt Engineering**	• Introduction to LLM • Introduction to prompt engineering • n-shot prompting • chain-of-thought (CoT) prompting • Generated knowledge prompting	Online research papers and references
6	**Part-of-Speech Tagging**	• (Mostly) English Word Classes • The Penn Treebank Part-of-Speech Tag set • Part-of-Speech Tagging • Markov Chains • The Hidden Markov Model • HMM Part-of-Speech Tagging	T1 Chapter 8 and other online references
7	**Statistical, ML and Neural models of**	• Likelihood Computation: The Forward Algorithm	T1 Appendix A, research

	POS tagging	• Decoding: The Viterbi Algorithm • Maximum Entropy Markov Model • Bidirectionality • Neural network models for POS	papers and other online references
8	**Review**	Review of contact session 1 to 7	
9	**Parsing**	• Grammars and Sentence Structure. • What Makes a Good Grammar • A Bottom-Up Chart Parser • Probabilistic Context-Free Grammars • Probabilistic CKY Parsing of PCFGs • Ways to Learn PCFG Rule Probabilities • Problems with PCFGs • Improving PCFGs by Splitting Non-Terminals, Lexicalized CFGs	T2 Chapter3. T1 Chapter 14
10	**Dependency Parsing**	• Dependency Relations • Dependency Formalisms • Dependency Treebanks • Transition-Based Dependency Parsing • Graph-Based Dependency Parsing • Dependency parsers using neural network	T1 Chapter 19
11	**Encoder-Decoder Models, Attention and Contextual Embeddings**	• Neural Language Models and Generation • Encoder-Decoder Networks, Attention • Applications of Encoder-Decoder Networks • Self-Attention and Transformer Networks • BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding • Contextual Word Representations	T1 Chapter10
12	**Word Sense Disambiguation**	• Word Senses • Relations between Senses • WordNet: A Database of Lexical Relations • Word Sense Disambiguation • Alternate WSD algorithms and Tasks	T1 Chapter 15

		• Word Sense Induction	
13	**Semantic Web Ontology and Knowledge Graphs**	• Introduction • Ontology and Ontologies • Ontology Engineering • Ontology Learning • Generation of knowledge graph	R1 chapter 24, research papers and online resources
14	**Retrieval Augmented Generation**	• large language model (LLM) enhancement • Accessing external knowledge bases, • Semantic search • Utilizing specific applications like chatbot or knowledge base enrichment	Online research papers and references
15	**State of art applications**	• Text Summarization • Discussion on other state of art application	Research papers
16	**Review**	Review of contact session 9 to 15	

Experiential Learning Components:

Describe objective, outcome of Experiential Learning Component and the lab infrastructure needed (virtual, remote, open source etc...) number of lab exercises needed, etc.

1. Lab work: 10
2. Project work: 0
3. Case Study: 4 Webinars
4. Simulation: 0
5. Work Integrated Learning Assignment- 2 Assignments
6. Design work/ Field work: 0

Objective of Experiential Learning Component:
Hands on sessions on implementation of fundamental NLP algorithms using state of art tools
Scope of Experiential Learning Component:
Programming language - Python
Tools and libraries: NLTK, SPACY, Jupyter, ScikitLearn, OpenAI, neo4j, Huggingface, Tensorflo etc
Lab Infrastructure:
Online/ Open source/Google Colab
List of Experiments:

Lab No	Lab Objective	Session Reference
1	Introduction to NLTK, Spacy and other tools	1
2	Word embedding - Skip gram, CBOW model	2
3	N-gram Language Modeling	3

4	Neural language Modeling	4
5	Part of speech tagging	6,7
6	Parsing	9,10
7	BERT, Large language models	5,11
8	WordNet, Ontology and Knowledge Graph	12,13
9	Retrieval Augment Generation	14
10	Text Summarization	15

Evaluation Scheme:

Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

Evaluation Component	Name (Quiz, Lab, Project, Mid-term exam, End semester exam, etc.)	Type (Open book, Closed book, Online, etc.)	Weigh t	Duration	Day, Date, Session, Time
EC – 1*	Quiz	Online	10%	1 week	To be announced
	Assignment/Lab Assignment / Lab Exams	Online	20 %	10 days	To be announced
EC - 2	Mid-Semester Test	Closed Book	30%	2 hours	To be announced
EC - 3	Comprehensive Exam	Open Book	40%	2 ½ Hours	To be announced

EC1* (20% - 30%): Quiz (optional): 5-10 %, Lab Assignment/Assignment: 20% - 30%
Syllabus for Mid-Semester Test (Closed Book): Topics in Contact session: 1 to 8
Syllabus for Comprehensive Exam (Open Book): All topics

Important Links and Information:

eLearn Portal: https://elearn.bits-pilani.ac.in

Students must visit the eLearn portal regularly and stay updated with the latest announcements and deadlines.

Contact Sessions: Students should attend the online lectures as per the schedule provided on the eLearn portal.

Evaluation Guidelines:

1. EC-1 consists of either two Assignments or three Quizzes. Students will attempt them through the course pages on the eLearn portal. Announcements will be made on the portal in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: "open book" means text/ reference books (publisher copy only) and does not include any other learning material. No other learning material will be permitted during the open book examinations. For Detailed Guidelines refer to the attached document.
   EC3 Guidelines
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam, which will be made available on the eLearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self-study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignments/Quizzes, Mid-Semester Tests and Comprehensive Exams according to the evaluation scheme provided in the handout.

****