BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

WORK INTEGRATED LEARNING PROGRAMMES

COURSE HANDOUT

Part A: Content Design

Course Title	Natural Language Processing
Course No(s)	
Credit Units	4 units
Course Author	Dr. Chetana Gavankar
Version No	1.0
Date	September 2022

Course Objectives

No	**Course Objective**
CO1	To learn the fundamental concepts and techniques of natural language processing (NLP) including Language Models, Word Embedding, Part pf speech Tagging, Parsing
CO2	To learn computational properties of natural languages and the commonly used algorithms for processing linguistic information
CO3	To introduce basic mathematical models and methods used in NLP applications to formulate computational solutions.
CO4	To introduce students research and development work in Natural language Processing

Text Book(s)

T1	Jurafsky and Martin, SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, McGraw Hill
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
R3	Semantic Web Primer (Information Systems) By Antoniou, Grigoris; Van Harmelen, Frank

Modular Content Structure

1. **Natural Language Understanding and Generation**

- The Study of Language.
- Applications of Natural Language Understanding.
- Evaluating Language Understanding Systems.
- The Different Levels of Language Analysis.
- The Organization of Natural Language Understanding Systems.

2. N-gram Language Modelling

- N-Grams
- Generalization and Zeros.
- Smoothing
- The Web and Stupid Backoff
- Evaluating Language Models
- Smoothing
- The Web and Stupid Backoff

3 Neural networks and Neural language Models

- Units
- The XOR problem
- Feed-Forward Neural Networks
- Training Neural Nets
- Neural Language Models -expand spend more time

4. 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
- Part-of-Speech Tagging for Morphological Rich Languages

5. Hidden Markov Models and MEMM

- The Hidden Markov Model
- Likelihood Computation: The Forward Algorithm
- Decoding: The Viterbi Algorithm
- HMM Training: The Forward-Backward Algorithm
- Maximum Entropy Markov Models
- Bidirectionality

6. Topic Modelling

- Mathematical foundations for LDA : Multinomial and Dirichlet distributions
- Intuition behind LDA
- LDA Generative model
- Latent Dirichlet Allocation Algorithm and Implementation
- Gibbs Sampling

7. Vector semantics and Embedding

- Lexical semantics
- Vector semantics
- Word and Vectors
- TFIDF
- Word2Vec, Skip gram and CBOW
- Glove
- Visualizing Embedding's

8. Grammars and Parsing.

- Grammars and Sentence Structure.
- What Makes a Good Grammar
- A Top-Down Parser.
- Bottom-Up Chart Parser.
- Top-Down Chart Parsing.
- Finite State Models and Morphological Processing.
- Grammars and Logic Programming.

9. Statistical Constituency Parsing

- Probabilistic Context-Free Grammars
- Probabilistic CKY Parsing of PCFGs
- Ways to Learn PCFG Rule Probabilities
- Problems with PCFGs
- Improving PCFGs by Splitting Non-Terminals
- Probabilistic Lexicalized CFGs

10. Dependency Parsing

- Dependency Relations
- Dependency Formalisms
- Dependency Treebanks
- 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: A Contextual Introduction
- The Illustrated BERT, ELMo, and co.
- XLM

12. Word sense disambiguation

- Word Senses
- Relations between Senses
- WordNet: A Database of Lexical Relations
- Word Sense Disambiguation
- Alternate WSD algorithms and Tasks
- Using Thesauruses to Improve Embedding's
- Word Sense Induction

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. Introduction to NLP Applications

- Brief introduction of state of art applications
- Text Summarization
- Machine Translation

Part B: Contact Session Plan

Academic Term	
Course Title	
Course No	
Lead Instructor	

Course Contents

Contact session	List of Topic Title (from content structure in Part A)	Topic # (from content structure in Part A)	Text / Ref Book / External resource
1	**Natural Language Understanding and Generation** 1.1 The Study of Language. 1.2 Applications of Natural Language Understanding. 1.3 Evaluating Language Understanding Systems. 1.4 The Different Levels of Language Analysis. 1.5 The Organization of Natural Language Understanding Systems.	Chapter1	T2
2	**N-gram Language Modelling** • N-Grams • Generalization and Zeros. • Smoothing • The Web and Stupid Backoff • Evaluating Language Models • Smoothing • The Web and Stupid Backoff	Chapter 3	T1
3	**Neural Network and Neural Language Modelling** • Units • The XOR problem • Feed-Forward Neural Networks • Training Neural Nets	Chapter 4	R2

	• Neural Language Models		
4	**Vector semantics and Embedding** • Lexical semantics • Vector semantics • Word and Vectors • TFIDF • Word2Vec, Skip gram and CBOW • Glove • Visualizing Embedding's	Chapter 6	T1 and lecture notes https://www.youtube.com/watch?v=hQwFeIupNP0
5	**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 • Part-of-Speech Tagging for Morphological Rich Languages	Chapter8	T1 and class notes
6	**Hidden Markov Model Algorithms** • Likelihood Computation: The Forward Algorithm • Decoding: The Viterbi Algorithm • HMM Training: The Forward-Backward Algorithm • Maximum Entropy Markov Model • Bidirectionality	Appendix chapter A	T1 and class notes
7	**Topic modelling** • Mathematical foundations for LDA • Multinomial and Dirichlet distributions • Intuition behind LDA • LDA Generative model • Latent Dirichlet Allocation Algorithm and Implementation • Gibbs Sampling		Class Notes
	Review of M1 to M7		
9	**Grammars and Parsing** • Grammars and Sentence Structure. • What Makes a Good Grammar • A Top-Down Parser. • A Bottom-Up Chart Parser. • Top-Down Chart Parsing. • Finite State Models and Morphological Processing. • Grammars and Logic Programming. • Parsing	Chapter3	T2

10	**Statistical Constituency Parsing** • Probabilistic Context-Free Grammars • Probabilistic CKY Parsing of PCFGs • Ways to Learn PCFG Rule Probabilities • Problems with PCFGs • Improving PCFGs by Splitting Non-Terminals • Probabilistic Lexicalized CFGs	Chapter 14	T1
11	**Dependency Parsing** • Dependency Relations • Dependency Formalisms • Dependency Treebanks • Transition-Based Dependency Parsing • Graph-Based Dependency Parsing • Dependency parsers using neural network	Chapter 19	T1 and class notes
12	**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: A Contextual Introduction • The Illustrated BERT, ELMo, and co. • XLM	Chapter10	T1 https://colab.research.google.com/drive/1iqs9Y5_zLI6R6mAwlnapcxcUbKjpv2CC?usp=sharing
13	**Word sense and word net** • Word Senses • Relations between Senses • WordNet: A Database of Lexical Relations • Word Sense Disambiguation • Alternate WSD algorithms and Tasks • Using Thesauruses to Improve Embedding • Word Sense Induction	Chapter15	T1
14	**Semantic web ontology and Knowledge Graphs** • Introduction • Ontology and Ontologies • Ontology Engineering • Ontology Learning	Chapter 24	R1 and class notes
15	**State of art applications**		Class Notes and web references

16	**Review of session 9 to session 15**		

Detailed Plan for Lab work

Lab No.	Lab Objective	Lab Sheet Access URL	Session Reference
1	Introduction to NLTK, Spacy and other open source tools		1
2	Language Modelling- Neural		2,3
3	Part of speech tagging		4,5
4	Topic Modeling		7
5	Parsing-Dependency-neural		9,10,11
6	Wordnet, Ontology and Knowledge Graph		12,13,14

Evaluation Scheme

Evaluation Component	**Name** (Quiz, Lab, Project, Midterm exam, End semester exam, etc)	**Type** (Open book, Closed book, Online, etc.)	**Weight**	**Duration**	**Day, Date, Session, Time**
EC – 1	Quiz		10%		To be announced
EC – 2	Assignment		20%		To be announced
EC – 3	Mid-term Exam	Open book	30%		To be announced
EC – 4	End Semester Exam	Open book	40%		To be announced

Important Information

Syllabus for Mid-Semester Test (Closed Book): Topics in Weeks 1-8 (1-18 Hours)
Syllabus for Comprehensive Exam (Open Book): All topics given in plan of study

Notes

- Quiz and Assignments timelines will be announced on the canvas portal.
- **Deadlines for evaluation components will NOT be extended** and the student is requested not to wait for the deadline to start working on Quiz/Assignment
- Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 to 8
- Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 16)
- **Strictly NO MAKEUPS for Quiz and Assignments** and all submissions after the announced deadlines will not be considered for evaluation.
- **All assignments will be subjected to plagiarism check, and if violated will be subject to disciplinary action apart from nullifying all the marks/grades assigned.**

Important links and information:

Canvas: Students are expected to visit the Canvas portal on a regular basis and stay up to date with the latest announcements and deadlines.

Contact sessions: Students should attend the online lectures as per the schedule provided.
Evaluation Guidelines:

1. EC-1 consists of Assignments and Quizzes. Announcements regarding the same will be made in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
3. For Open Book exams: Use of prescribed and reference text books, in original (not photocopies) is permitted. Class notes/slides as reference material in filed or bound form is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
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. The genuineness of the reason for absence in the Regular Exam shall be assessed prior to giving permission to appear for the Make-up Exam. Make-Up Test/Exam will be conducted only at selected exam centres.

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 lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.

Learning Outcomes:

No	Learning Outcomes
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	Should be able to apply NLP algorithms along with deep learning algorithms for state of art areas like word embedding