CSE 3201: Natural Language Processing

Programme: B.Tech (CSE)	**Year:** Third	**Semester: Fifth**
Course: Program Elective	**Credits:** 3	**Hours:** 40

Course Context and Overview:

The main objective of this course is to make students understand and apply various automated textual processing methods for processing and analyzing textual data (English). This course will help students to gain knowledge on various existing models and algorithms to process and analyze the textual data. This course will also equip students with the skills to use the state-of-the-art tools and applications for analyzing textual data. This course requires the knowledge of Python to complete the projects.

Prerequisites Courses: Design and Analysis of Algorithms, Computer Programming, Theory of Computation

Course outcomes (COs):

On completion of this course, the students will have the ability to:
CO1: Demonstrate the knowledge of fundamental concepts in natural language processing
CO2: Demonstrate the understanding of various algorithms to process textual data
CO3: Show a working knowledge of various levels of textual data processing in order to process the linguistic data
CO4: Implement the algorithms studied, in various situations, to process and analyze textual data

Course Topics with hours for each section

Contents	Lecture Hours	
UNIT – 1 **Introduction**		**2**
History, Ambiguity, Knowledge in speech and NLP	**1**	
The State of the Art, Models and Algorithms	**1**	

UNIT –2 **N Grams and Sequence Modelling**		**6**
Word Counting, Simple N-Grams and Evaluating N-Grams	**2**	
Smoothing Process	**1**	
English Word Classes, Tag Sets and POS-Tagging	**1**	
HMM POS-Tagging, Markov Chains, Hidden Markov Model, Forward algorithm and Viterbi Algorithm	**2**	
UNIT-3 **Synctactic Parsing**		**6**
Top-Down Parsing, Bottom-up Parsing, CKY Parsing and The Earley Parsing	**2**	
Probabilistic Context-Free Grammar (PCFG), PCFG for Disambiguation and Language Modeling, Probabilistic CKY Parsing of PCFG and Learning PCFG rule Probabilities	**2**	
The Collins Parser	**2**	
UNIT-4 **Semantic Analysis**		**8**
Lexical Semantics and Compositional Semantics Word Sense Disambiguation - Naive Bayes Classifier, Dictionary and Thesaurus Methods – Lesk and Yarrowsky's Algorithm Word Similarity – Thesaurus Method	**4**	
Vector Semantics and Embeddings – Words and Vectors, Cosine Similarity, TF-IDF, Pointwise Mutual Information, Word2vec, Skip-gram embeddings	**4**	
UNIT-5 **Information Extraction**		**6**
Named Entity Recognition	**2**	
Relation Detection and Classification	**2**	
Temporal and Event Processing, Template Filling	**2**	

UNIT-6 **Introduction to Deep Learning for Natural Language Processing**		**8**
Neural Networks, Feed-Forward Neural Networks, Neural Language Models	**3**	
Deep Learning Architectures for Sequence Processing – Recurrent Neural Networks as Language Models, LSTMs, GRUs and Transformers as Autoregressive Language Models	**5**	
UNIT-7 **Additional Topics**		**4**
Question Answering (QA) – Information Retrieval	**2**	
Factoid QA Summarization – Single Documents and Multi-Documents	**2**	

Textbooks and Reference books:

Textbooks:

1. Speech and Language Processing - *An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Daniel Jurafsky and James H. Martin, Pearson, 2nd edition, 2014.
2. *Speech and Language Processing - An Introduction to Natural Language Processing,* Computational Linguistics and Speech Recognition, Daniel Jurafsky and James H. Martin, Pearson, 3rd edition Draft, Dec 2021. (Book not yet published)
   Web Link for the draft: *https://web.stanford.edu/~jurafsky/slp3/*

Evaluation Methods:

Item	**Weightage**
Quiz 1	10%
Quiz 2	10%
Mid Semester Exam	20%

Project Round – 1 and Report submission just after Mid Semester Exam	10%
Project Round – 2 and Report submission just after End Semester term	15%
End Semester Exam	35%

Prepared By: Sakthi Balan Muthiah in April 2019.
Updated By: Sakthi Balan Muthiah in June 2019.
Updated By: Sakthi Balan Muthiah in April 2020.
Updated By: Sakthi Balan Muthiah in May 2020.
Updated By: Sakthi Balan Muthiah in Aug 2020 (Evaluation updated).
Updated By: Sakthi Balan Muthiah in Aug 2021.
Updated By: Sakthi Balan Muthiah in Aug 2022 (Evaluation updated)