CSE3201: 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.

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 and speech 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 (an indicative one)

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**		**6**
Word Counting and Simple N-Grams	**1**	
Training Sets, Test Sets and Evaluating N-Grams	**2**	
Smoothing Process	**3**	
UNIT-3 **Sequence Modelling**		**7**
English Word Classes, Tag Sets and POS-Tagging	**1**	
Rule-based approach to POS-Tagging, HMM POS-Tagging	**3**	
Markov Chains, Hidden Markov Model, Forward algorithm and Viterbi Algorithm	**3**	
UNIT-4 **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-5 **Semantic Analysis**		**6**
Lexical Semantics and Word Sense Disambiguation	**2**	
Compositional Semantics	**1**	
Semantic Role Labeling and Semantic Parsing	**3**	
UNIT-6 **Information Extraction**		**6**

Named Entity Recognition	**2**	
Relation Detection and Classification	**2**	
Temporal and Event Processing, Template Filling	**2**	
UNIT-7 **Additional Topics**		**7**
Question Answering (QA) – Information Retrieval	**2**	
Factoid QA Summarization – Single Documents and Multi-Documents	**2**	
Dialogue and Conversational Agents – Basic Dialogue System	**3**	

Textbooks and Reference books:

Textbook:

1. Daniel Jurafsky and James H. Martin, *"Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Pearson, 2nd edition, 2014.

Reference books:

1. Christopher D. Manning and Hinrich Schüze, "*Foundations of Statistical Natural Language Processing"*, The MIT Press. Cambridge. Massachusetts, London, England, 1999.

2. Daniel Jurafsky and James H. Martin, "*Speech and Language Processing - An Introduction to Natural Language Processing"* Computational Linguistics and Speech Recognition, Pearson, 3rd edition Draft, 2019.

   Web Link for the draft: *https://web.stanford.edu/~jurafsky/slp3/*

Evaluation Methods:

Item	**Weightage**
Active participation in Class and Piazza (if considerable participation is there)	5%
Mid Semester Exam	25%
Project Round – 1 and Round – 2: Report Sub mission	25%
End Semester Exam	45%

Note: If active participation in Piazza is not considerable enough, then the 5% weightage assigned to it will be added to the weightage of the End Semester Exams.

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.