20CS020 NATURAL LANGUAGE PROCESSING

Hours Per Week :

L	T	P	C
3	-	-	3

Total Hours :

L	T	P
45	-	-

WA/RA	SSH/HSH	CS	SA	S	BS
15	30	-	5	5	-

Course Description and Objectives:

The main objectives of this course are to acquire basic understanding of linguistic concepts and natural language complexity, variability, to acquire basic understanding of machine learning techniques as applied to language to processes the language tasks.

Course Outcomes:

Upon the Completion of the course, students will be able to:

- ✓ Understand the basic issues related with processing natural languages.
- ✓ Apply computational methods to processes Natural Languages.
- ✓ Analyze various case studies on several Natural Language Tasks.
- ✓ Design feasible solutions for natural language tasks.

UNIT I

Introduction and Overview: What is Natural Language Processing, hands-on demonstrations. Ambiguity and uncertainty in language. The Turing test.

Regular Expressions: Chomsky hierarchy, regular languages, and their limitations. Finite- state automata. Practical regular expressions for finding and counting language phenomena. A little morphology. Exploring a large corpus with regex tools.

UNIT II

Context Free Grammars: Constituency, CFG definition, use and limitations. Chomsky Normal Form. Top-down parsing, bottom-up parsing, and the problems with each. The desirability of combining evidence from both directions

Non-probabilistic Parsing: Efficient CFG parsing with CYK, another dynamic programming algorithms. Early parser. Designing a little grammar, and parsing with it on some test data.

UNIT III

Language modeling and Naive Bayes: Probabilistic language modeling and its applications. Markov models. N-grams. Estimating the probability of a word, and smoothing. Generative models of language. Part of Speech Tagging and Hidden Markov Models, Viterbi Algorithm for Finding Most Likely HMM Path Dynamic programming with Hidden Markov Models, and its use for part-of-speech tagging.

UNIT IV

Probabilistic Context Free Grammars: Weighted context free grammars. Weighted CYK, Pruning and beam search.

Parsing with PCFGs: A tree bank and what it takes to create one. The probabilistic version of CYK. Also: How do humans parse? Experiments with eye-tracking. Modern parsers. Maximum Entropy Classifiers: The maximum entropy principle and its relation to maximum likelihood. Maximum entropy classifiers and their application to document classification, sentence segmentation, and other language tasks.

UNIT V

Maximum Entropy Markov Models & Conditional Random Fields: Part-of-speech tagging, noun-phrase segmentation and information extraction models that combine maximum entropy and finite-state machines. State-of-the-art models for NLP.

Lexical Semantics: Mathematics of Multinomial and Dirichlet distributions, Dirichlet as a smoothing for multinomial's.

TEXT BOOKS:

1. Jurafsky and Martin, "Speech and Language Processing" Prentice Hall
2. Manning and Schutze "Statistical Natural Language Processing", MIT Press
3. James Allen "Natural Language Understanding". The Benajmins/ Cummings, Publishing Company

REFERENCE BOOKS:

1. Cover, T. M. and J. A. Thomas: Elements of Information Theory. Wiley.
2. Charniak, E. "Statistical Language Learning". The MIT Press.
3. Jelinek, F."Statistical Methods for Speech Recognition". The MIT Press.
4. Lutz and Ascher - "Learning Python", O'Reilly