

Objectives: To make the students aware about processing of unstructured data. To design the computer programs that considers and processes the textual data as human-like comprehension. To design the computer system that able to understand and draw inferences from textual data.

Unit I: Introduction: What is Natural Language Processing, Ambiguity and uncertainty in language. The Turing test, Models and Algorithms

Unit II: Regular Expressions, Automata and Finite-State Transducers: Basic Regular Expression Patterns, Disjunction, Grouping, and Precedence, Examples, Advanced Operators, Regular Expression Substitution, Using FSA to Recognize Sheeptalk, Formal Languages, Nondeterministic FSAs, Using an NFSA to accept strings, Recognition as Search, Relating Deterministic and Non-deterministic Automata, Regular Languages and FSAs, Survey of (Mostly) English Morphology, Inflectional Morphology, Derivational Morphology, The Lexicon and Morphotactics, Morphological Parsing with Finite-State Transducers, Orthographic Rules and Finite-State Transducers, Human Morphological Processing

Unit III: N-gram Language Models: Words, Corpora, Text Normalization, Minimum Edit Distance, N-Grams, Evaluating Language Models, Generalization and Zeros, Smoothing, Kneser-Ney Smoothing

Unit IV: Word Classes and Part-of-Speech Tagging: English Word Classes, Tagsets for English, Part of Speech Tagging, Rulebased Part-of-speech Tagging, Stochastic Part-of-speech Tagging, The Actual Algorithm for HMM tagging, Transformation-Based Tagging, How TBL rules are applied, How TBL Rules are Learned, Multiple tags and multiple words, Unknown words, Class-based N-grams

Unit V: Context-Free Grammars for English and Parsing: Constituency, Context-Free Rules and Trees, Sentence-Level Constructions, The Noun Phrase, The Verb Phrase and Subcategorization, Auxiliaries, Spoken Language Syntax, Grammar Equivalence & Normal Form, Finite State & Context-Free Grammars, Grammars & Human Processing, Parsing as Search, Top-Down and Bottom-Up Parsing with comparison, A Basic Topdown Parser, Adding Bottom-up Filtering, Problems with the Basic Topdown Parser, Left Recursion, Ambiguity

Unit VI: Semantics and Pragmatics: Syntax-Driven Semantic Analysis, Attachments for a Fragment of English, Idioms and Compositionality, Robust Semantic Analysis, WordNet: A Database of Lexical Relations, Reference Resolution, Text Coherence, Discourse Structure, Psycholinguistic Studies of Reference and Coherence,

Reference books

1. Speech and Language Processing by Daniel Jurafsky and James H. Martin Second edition, Prentice Hall
2. Natural Language Understanding by Allen James Second edition, Benjamin/Cumming
3. Statistical Language Learning by Charniack Eugene MIT Press

4. Foundations of Statistical Natural Language Processing by Manning Christopher and Heinrich Schutze MIT Press