GUJARAT TECHNOLOGICAL UNIVERSITY
MASTER OF COMPUTER APPLICATIONS (MCA)
SEMESTER: V

Subject Name: **Language Processing (LP) (Elective-III)**
Subject Code: **650014**

Learning Objectives:

- Words: Fundamental building block in a language
- Computational models of spelling and correction of morphology of words
- Regular expressions, Finite State Automata (FSA), Finite State Transducers (FST)
- N-Gram models of word sequences
- Computational models for Part-of-Speech (POS), Phrases, Words' dependence
- POS Tagging, Modeling English as CFG, Parsing
- Ways to represent the meaning of utterances thru' the First Order Predicate Calculus
- Algorithms for Reference Resolution

Pre-requisites
- Statistics

Contents

1. **Introduction to Language Processing, Regular Expressions & Automata**   **[5 Sessions]**
   Introduction: Morphology; Syntax; Semantics; Pragmatics; Discourse Convention; Ambiguity; Disambiguation; Models & Algorithms; Regular Expressions; Regular Expressions Substitutions; Memory and ELIZA; Finite State Automata (FSA); Formal Languages; Non-Deterministic FSA (NFSA); Using an NFSA to Accept Strings; Recognition as Search; Relating DFSA & NFSA.

2. **Morphology and Finite State Transducers (FST)**   **[5 Sessions]**
   Introduction; Survey of (English) Morphology; Inflectional and Derivational Morphology; Finite State Morphological Parsing: Introduction; Lexicon & Morpho-tactics; Morphological Parsing with FST; Orthographic Rules & FST; Combining FST Lexicons & Rules; Lexicon-Free FSTs: Porter Stemmer; Human Morphological Processing

3. **Probabilistic Models of Spelling, N-Grams**   **[5 Sessions]**
   Introduction; Dealing with Spelling Errors; Spelling Error Patterns; Determining Non-Word Errors; Probabilistic Models; Applying the Bayesian Model to Spelling; Minimum Edit Distance; Introduction to N-Grams; Counting Words in Corpora; Simple (Unsmoothed) N-Grams; N-Grams for Spelling; Entropy

4. **Word Class & Part-of-Speech (POS) Tagging**   **[3 Sessions]**
   Introduction; English Word Classes; Tag Sets for English; POS Tagging; Rule-Based and Stochastic POS Tagging

5. Context-Free Grammar (CFG) for English, Parsing with CFG   [6 Sessions]
Introduction; Constituency; Context-Free Rules & Trees; Sentence-Level Constructions; The Noun Phrase; Coordination; Agreement; The Verb Phrase & Sub-Categorization; Auxiliaries; Introduction to Parsing with CFG; Parsing as Search; Top-Down & Bottom-Up Parsing; A Basic Top-Down Parser; Finite State Parsing Methods

6. Features & Unification   [3 Sessions]
Introduction; Feature Structures; Unification of Feature Structures, Feature Structures in Grammar

7. Representing Meaning   [5 Sessions]
Introduction; Computational Representation; Meaning Structure of Language; First Order Predicate Calculus; Linguistically Relevant Concepts such as Categories, Events, Representing Time, Aspect, Representing Beliefs, Pitfalls; Related Representational Approaches; Alternate Approaches to Meaning

8. Semantic Analysis   [3 Sessions]
Introduction; Syntax-Driven Semantic Analysis; Attachments for a Fragment of English

9. Lexical Semantics   [4 Sessions]
Introduction; Relation among Lexemes and their Senses; WordNet: A Database of Lexical Relations; The Internal Structure of Words

10. Discourse   [5 Sessions]
Introduction; reference Resolution; Text Coherence

Text Book:

1. Daniel Jurafsky & James H. Martin, "Speech and Language Processing", Pearson, 5th Impression (2011) ISBN 378-81-317-1672-4

Reference Books:

1. John C. Martin, "Introduction to Languages and the Theory of Computation", Tata McGraw-Hill, (2003), 3rd Edition, ISBN: 007049939X
2. Stuart Russell & Peter Norvig, "Artificial Intelligence: A Modern Approach (Specifically Chapters 22, 23)", PHI (2005) Rs. 395/-, ISBN-81-203-2382-3
3. Rob Callan, "Artificial Intelligence (Specifically Chapters 18, 19)", Palgrave Macmillan (2006), Rs. 525/-, ISBN-0-333-80136-9
4. Dan W. Patterson, "Introduction to Artificial Intelligence and Expert Systems (Specifically Chapters 12)", PHI (2010) Rs. 275/-, ISBN-978-81-203-0777-3
5. Ben Coppin, "Artificial Intelligence Illuminated (Specifically Chapters 20)", Narosa (2005) Rs. 295/-, ISBN-81-7319-671-0

Course Coverage (From Text Book):

Unit-1: Chapter-1 (1.1 to 1.5), Chapter-2 (2.1, 2.2)
Unit-2: Chapter-3
Unit-3: Chapter-5 (5.1 to 5.6), Chapter-6 (6.1, 6.2, 6.6, 6.7)
Unit-4: Chapter-8 (8.1 to 8.5)
Unit-5: Chapter-9 (9.1 to 9.8), Chapter-10 (10.1, 10.2, 10.5)
Unit-6: Chapter-11 (11.1 to 11.3)
Unit-7: Chapter-14
Unit-8: Chapter-15 (15.1 to 15.2)
Unit-9: Chapter-16 (16.1 to 16.3)
Unit-10: Chapter-18 (18.1, 18.2)

Accomplishment of Students after Completing the Course

Students shall learn lexical, syntactic, semantic, and pragmatic analysis of English language text. In particular, they will develop the ability to apply:
- FSA, FST and N-Gram models for morphological parsing, stemming, and spelling correction.
- Computational models for POS tagging, and parsing with CFG
- First Order Predicate Calculus and computational processes to represent meaning.
- The algorithms for reference (pronoun) resolution, and application of text coherence for reference resolution.