Syllabus: Foundations of Natural Language Processing – CS 3101

Credits: 3 – 0 – 0

Module-1: Introduction (2-3 L)

What is Language? What is Natural Language Processing?; Why NLP? - Applications/ Motivation; Understanding Language- The Turing test, The Chinese Room Argument; Ambiguity and other Challenges in language. - Non-Standard English, Slangs, Neologism, Bias/Stereotypes, Exclusion and others.

Module-2: Probabilistic Language Model (3 L)

Zipf's law, Collocations, Language Models: n-grams, Probability Estimation, Unknown words, smoothing, Evaluation

Module-3: Morphology & Syntax (10-12 L)

Word Morphology - tokenisation, boundary detection, Stemming vs Lemmatization, Porter Stemmer, Text Normalization, Spelling Correction (String Edit Distance) -LCS; Word Class: Part of Speech Tagger- HMM & Viterbi; Parsing – Constituent (CKY parsing) vs Dependency Parsing, PCFG, Scope Ambiguity and Attachment Ambiguity Resolution; Named Entity Recognition, Evaluation (Exact vs Relaxed match).

Module-4: Text Classification (1-2 L)

Text Classification (sentiment, abstract vs concrete (psycholinguistics), Hate vs not hate), Policies (hopeful vs critical)

Module-5: Semantics & Meaning (5-6 L)

Meaning Representation; Lexical Semantics: Word Sense, Semantic Relations, Relation Extraction, Hearst Patterns, Distant Supervision; Word similarity vs relatedness, Sense disambiguation, Selectional Preference. Figurative texts (Contextual Incongruity); Pragmatics?

Module-6: Word Embeddings (2-4 L)

Bag of words, TF-IDF, Term-context Matrix, PPMI Word Embeddings, word2vec, Evaluating Word embeddings- *Analogy, Similarity, Emotion*

Module-7: Applications (3-4 L)

Applications such as Text Summarization, Machine Translation, Question-Answering.

Books/References:

(A) Speech and Language Processing by Daniel Jurafsky and James H. Martin

(B) Natural Language Processing with Python. (updated edition based on Python 3 and NLTK 3) Steven Bird et al. O'Reilly Media