Course Proposal: Special Topics in Natural Language Processing
Indian Institute of Technology, Kanpur
Department of Computer Science and Engineering

Course Number: CS698O

Course Title: Special Topics in Natural Language Processing

Units: 3-0-0-0-9

Proposed by: Ashutosh Modi

Pre-requisites : Instructor's consent and
Must: Introduction to Machine Learning (CS771) or equivalent course, Proficiency in Linear Algebra, Probability and Statistics, Proficiency in Python Programming
Desirable: Probabilistic Machine Learning (CS772), Topics in Probabilistic Modeling and Inference (CS775), Deep Learning for Computer Vision (CS776)

Level of the course: Senior UG and PG (6xx level)

Course Description:
Natural language (NL) refers to the language spoken/written by humans. NL is the primary mode of communication for humans. With the growth of the world wide web, data in form on textual natural language has grown exponentially. This calls for development of algorithms and techniques for processing natural language for the purposes of automation and for the development of intelligent machines. This course will primarily focus on understanding and developing techniques/learning algorithms/models for processing text. We will have a statistical approach to Natural Language Processing (NLP), wherein we will learn how one could develop natural language understanding models from regularities in large corpora of natural language texts.

Tentative Topics (total 40 lectures of 50 minutes each):

1. Introduction to Natural Language (NL) [1 lecture]: why is it hard to process NL, linguistics fundamentals, etc.

2. Language Models [4 lectures]: n-grams, smoothing, class-based, brown clustering

3. Sequence Labeling [5 lectures]: HMM, MaxEnt, CRFs, related applications of these models e.g. Part of Speech tagging, etc.

4. Parsing [4 lectures]: CFG, Lexicalized CFG, PCFGs, Dependency parsing

5. Applications [3 lecture]: Named Entity Recognition, Coreference Resolution, text classification, toolkits e.g. Spacy, etc.

6. Distributional Semantics [2 lecture]: distributional hypothesis, vector space models, etc.
7. Distributed Representations [3 lecture]: Neural Networks (NN), Backpropogation, Softmax, Hierarchical Softmax
8. Word Vectors [4 lecture]: Feedforward NN, Word2Vec, GloVE, Contextualization (ELMo etc.), Subword information (FastText, etc.)
9. Deep Models [5 lectures]: RNNs, LSTMs, Attention, CNNs, applications in language, etc.
10. Sequence to Sequence models [5 lectures]: machine translation and other applications
11. Transformers [4 lectures]: BERT, transfer learning and applications

References: There are no specific references, this course gleans information from a variety of sources like books, research papers, other courses, etc. Relevant references would be suggested in the lectures. Some of frequent references are as follows:

1. Speech and Language Processing, Daniel Jurafsky, James H.Martin,
2. Foundations of Statistical Natural Language Processing, CH Manning, H Schtze
3. Introduction to Natural Language Processing, Jacob Eisenstein
4. Natural Language Understanding, James Allen