CEN - 807: Natural Language Processing and Information Extraction

L	T	P		
4	0	0	External:	60 Marks
Credits : 4			**Total:**	**100 Marks**
Duration of Exam : 3 Hours				

Course Outcomes:

---

1. Students will be able to implement text processing tasks and develop probabilistic language models.
2. Students will be able to implement text classification and sequence modelling on various problems.
3. Students will be able to implement lexical semantics tasks: word similarity and word sense disambiguation.
4. Students will be able to understand distributional semantics, word embeddings and neural language models.
5. Students will be able to implement information extraction tasks: named entity recognition and relation extraction.

Unit 1. Text Processing Tasks and Probabilistic Language Models new topic: Byte pair Encoding

Introduction to Text, Speech and Language Technologies, Basic Text Processing Tasks, Normalization, Max Match Algorithm, Lemmatization, Porter Stemmer, Minimum Edit Distance, Probabilistic Language Models: N Grams, Bigram Probabilities, Perplexity, Smoothing Techniques: La Place, ~~Good Turing, Kneser Ney, Interpolation.~~

Unit 2. Text Classification and Sequence Modelling

Text Classification: Bag of words, Conditional Independence, Multinomial Naïve Bayes Classifier, Maximum Likelihood Estimation, Evaluation of Text Classification Model. Sentiment Analysis: Entity based and aspect Based Feature Extraction, Baseline Algorithm, Sentiment Lexicons, Polarity Analysis. Building Sentiment Lexicons: Semi supervised Algorithm, Turney Algorithm. Sequence Modelling: Markov Models, HMM, ~~Beam, Greedy and Viterbi inference, HMM, CRF, LSTM based POS tagging.~~

Unit 3. Lexical Semantics

Word Senses and Word Relations, Wordnet. Computing Word Similarities: Path Based, Information Content, Word Sense Disambiguation, Thesaurus based WSD using Wordnet, Lesk Algorithm, Typical Features of WSD, Supervised WSD, ~~Semi supervised WSD.~~

Unit 4. Distributional Semantics

Vector Semantics: Distributed Representations, Word Context Matrix Generation, Weighting Methods, Dimensionality Reduction, Similarity Measures. Word Embeddings, ~~Learning of Neural Embeddings.~~

Unit 5. Information Extraction

Named Entity Recognition: Hand Written Regular Expressions, Typical Features for NER, Classification models, Sequence Models. ~~Relation Extraction: Binary Relation Association, Relation Extraction from Wikipedia, Supervised Relation Extraction, Semi-supervised Relation Extraction, Distant Supervision~~.

Books:

- Daniel Jurafsky and James H. Martin, "Speech and Language Processing", 2nd Edition, Pearson Education, 2013.
- Yoav Goldberg, "Neural Network Methods in Natural Language Processing", Morgan & Claypool Publishers, 2017.
- Steven Bird, Ewan Klein, Edward Loper "Natural Language Processing with Python", O'Reilly, 2009.
- Manning and Schuetze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999.