|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Course Code** | **Course name** | | **L** | **T** | **P** | **C** |
| **CSEG4040P** | **Computational Linguistics and Natural Language Processing** | | **4** | **0** | **1** | **5** |
| **Total Units to be Covered: 05** | | **Total Contact Hours: 90** | | | | |
| **Prerequisite(s):** | **Applied Machine Learning - CSAI2017P** | | **Syllabus version: 1.0** | | | |

**Course Objectives**

1. The objectives of this course are:
2. To introduce the concept of Natural Language Understanding & Natural Language Generation.
3. To develop the concept of statistical and probabilistic approach of language modelling.
4. To extend the knowledge of Large Language Model.
5. To enrich the knowledge with different corpuses and different tools being used for machine translation.
6. To provide programming skills necessary for processing natural language.

**Course Outcomes**

On completion of this course, the students will be able to

CO1: Understand the techniques in NLP.

CO2: To understand the Large Language Model

CO3: To comprehend the natural language generation.

CO4: To understand the machine translation & information retrieval techniques.

**CO-PO Mapping**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Program**  **Outcomes**  **Course Outcomes** | **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PO12** | **PSO1** | **PSO2** | **PSO3** |
| **CO 1** | 1 | **-** | **-** | 1 | 1 | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | 3 |
| **CO 2** | 1 | **-** | **-** | 1 | 1 | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | 3 |
| **CO 3** | 1 | 2 | **-** | 2 | 1 | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | 3 |
| **CO4** | 1 | **-** | **-** | 3 | 1 | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | 3 |
| **Average** | 1 | .5 | **-** | 1.75 | 1 | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | **-** | 3 |

1 – Weakly Mapped (Low) 2 – Moderately Mapped (Medium)

3 – Strongly Mapped (High) “\_” means there is no correlation

**Syllabus**

**Unit I: Classical Approaches of NLP 12 Lecture Hours**

Introduction, Classical approaches to natural language processing , Understanding linguistics, Level 1: Morphology, Level 2: Syntax, Level 3: Semantics, Level 4: Pragmatics, Understanding linguistics, Traditional approach, Example: Automatic summarization using NLP, Drawbacks, Text processing, Ambiguities and computational challenges in processing various natural languages, Introduction to Real life applications of NLP such as spell and grammar checkers, information extraction, question answering, and machine translation.

**Unit II: Empirical & Statistical Approaches 12 Lecture Hours**

Corpus creation, Corpus linguistics, Types of corpora, Lexicographical implementations in corpora, Timeline of corpus linguistics, Usage areas of corpora, Traits of a good text corpus, Annotations in text corpus, NLP task-specific training corpora, Treebank annotation, Usage of annotations and corpora, Kinds of annotations, Tree banks and its construction, Need for tree bank, Types of tree bank corpus, Ambiguity in language, Segmentation, Stemming, Tokenization, Representation of word, Sentence, Word embedding, Word Senses, Linguistic Structure: Dependency Parsing. Fundamental statistical techniques, Problems of the traditional approach, how statistics helps, Problems of the traditional approach and how statistics helps, Hidden Markov model, Maximum entropy Markov model, Conditional random field model, Support vector machine, N-GRAM, Perplexity, POS Tagging, Word sense disambiguation, POS tag and Hidden Markov model, POS tagging using HMM, Viterbi algorithm, Recurrent Neural network, Vanishing Gradients and exploding gradients. Parsing, Statistical parsing, Approaches to parsing, Statistical approach, Lexicalized statistical parsing, Top-down parsing, Bottom-up parsing, Left corner parsing method, Statistical parsing: Probabilistic parser,

**Unit III: Language Modelling 12 Lecture Hours**

Word similarity and text similarity, Text similarity methods, Jaccard similarity, K-means, Cosine similarity, Word Mover’s distance, Word sense disambiguation, Complications in WSD, Methods in WSD, Evaluation of WSD, the role of language models, estimating parameters and smoothing. Evaluating language models, LSTM (Long short-term memory), GRU (Gated recurrent Unit), Part of speech tagging, BERT, XLnet, 1D-CNN for NLP, Sub-word Models, Contextual Representations, Transformers, Self-Attention for Generative Models.

**Unit IV: Machine Translation 12 Lecture Hours**

Machine translation, Rule-based machine translation, Statically Machine Translation, Neural Machine Translation, Seq2Seq Modelling, Attention, Question Answering Bot, Natural Language Generation, Neural Machine Translation, Case studies on Amazon Alexa, Google Assistant, Microsoft Cortona etc.

**Unit V: Applications of Natural Language Processing 12 Lecture Hours**

Text Summarization, Document Summarization, Sentiment Analysis, Question Answering system, Sarcasm Detection, Hostile detection Information retrieval in NLP, Image caption generation, Intelligent Tutoring System

**Total lecture Hours 60**

# References\*

|  |  |
| --- | --- |
| **Textbooks** | 1. Christopher D. Manning, and Hinrich Schutze, “Foundations of Natural Language Processing”, 6th Edition, The MIT Press Cambridge, Massachusetts London, England, 2003. 2. Daniel Jurafsky and James H. Martin “Speech and Language Processing”, 3rd Edition, Prentice Hall, 2009. |
| **Reference books** | 1. Nitin Indurkhya, and Fred J. Damerau “Handbook of Natural Language Processing”, 2nd Edition, CRC Press, 2010. 2. James Allen, “Natural Language Understanding”, 8th Edition, Pearson Publication, 2012. 3. Christopher D. Manning and Hinrich Schütze, “Foundations of Statistical Natural Language Processing”, 2nd edition, MIT Press Cambridge, MA, 2003. 4. Hobson Lane, Cole Howard, and Hannes Hapke, “Natural language processing in action”, Manning Publications, 2019. 5. Alexander Clark, Chris Fox, and Shalom Lappin, “The Handbook of Computational Linguistics and Natural Language Processing”, Wiley-Blackwell, 2012. |
| **Web Resources** |  |
| **Journals** |  |
| **MOOCs, online courses** |  |

**Modes of Evaluation: Quiz/Assignment/ presentation/ extempore/ Written Examination etc.**

**Examination Scheme**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Components** | **IA** | **MID SEM** | **End Sem** | **Total** |
| Weightage (%) | 50 | 20 | 30 | 100 |

**Computational Linguistics and Natural Language Processing Lab**

**List of Experiments**

**EXPERIMENT-1:**

**Title: Installing various packages required for analytics in python and write code in Python for following program**

1. Write code to load CSV file containing information about employee of a company in python and draw graph showing average salary department wise.

**EXPERIMENT-2:**

**Title: Text Retrieval**

1. Connect to Twitter account and Extract first 100 tweets from it in a file.
2. Study and Implementation of Processing text (Word and Sentence Tokenization)

**EXPERIMENT-3:**

**Title: Processing Data**

1. Python code to read a text document and perform basic pre-processing techniques on the text like tokenization, stop-word-removal, lemmatization etc.
2. Study and Implementation of Morphological analysis.

**EXPERIMENT-4:**

**Title: Do text mining on extract data and Accessing text corpus**

1. Calculate word count of a given specific document and show top 10 frequent words with their frequency and Create world cloud and show graphically.
2. Study and Implementation of NER (Name Entity Recognition)

**EXPERIMENT-5:**

**Title: POS-Tagging and Tagging and Parsing**

1. Categorizing and tagging words in Twitter Data.
2. Study and implementation of POS Tagging and Chunking in a sentence.

**EXPERIMENT-6:**

**Title: Language Processor**

1. Implement N–Gram Language Mode and Smoothing.

**EXPERIMENT-7:**

**Title: Do sentimental analysis**

1. Analysis of Sentiment and Subjectivity
2. Implement sentimental analysis on IMDB Movie Reviews Dataset.

**EXPERIMENT-8:**

**Title: Do Text Summarization**

1. Analysing Meaning of Sentences
2. Implement Text Summarization on IMDB Movie Reviews Dataset.

**EXPERIMENT-9 & 10:**

**Title: Mini-Project on NLP**

Implement Mini-Project on NLP applications.

**Total Lab hours 30**

# References\*

|  |  |
| --- | --- |
| **Textbooks** | 1. Daniel Jurafsky, and James H. Martin, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition”, 2nd Edition, Pearson, 2013. |
| **Reference books** | 1. David A. Grossman, and Ophir Frieder, “Information Retrieval: Algorithms and Heuristics”, Springer, 2004. 2. Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal, “Natural Language Processing: A paninian perspective”, Prentice Hall, New Delhi, 1995. |
| **Web Resources** |  |
| **Journals** |  |
| **MOOCs, online courses** |  |

**Modes of Evaluation: Quiz/Assignment/ presentation/ extempore/ Written Examination etc.**

**Examination Scheme:** Continuous Assessment

|  |  |  |
| --- | --- | --- |
| **Components** | **Quiz & Viva** | **Performance & Lab Report** |
| Weightage (%) | 50 % | 50 % |