INT356:NATURAL LANGUAGE PROCESSING

L:2 T:0 P:3 Credits:4

Course Outcomes: Through this course students should be able to

CO1 :: Define the concept of Lexical processing, Syntactic processing and Semantic processin

CO2 :: Understand the process of Lexical, Syntax formation and Semantic analysis

CO3 :: Apply the Lexical processing, Syntax and Semantic analysis process on different texts.

CO4 :: Analyze the Syntax, Semantics, and Pragmatics of a statement written in a Natural Language

CO5 :: Evaluate the different Lexical, Semantic and Syntax analysis models on various data sets.

CO6 :: Develop Text summarization, Machine translation and Speech-based applications uses for Speech analysis

Unit I

Module 1: Lexical Processing - I : NLP: Areas of Application, Understanding Text, Text Encoding, Regular expressions: Quantifiers, Comprehension: Regular Expressions, Regular Expressions: Anchors and Wildcard, Regular Expressions: Characters Sets, Greedy versus Non-greedy Search, Commonly Used RE Functions, Regular Expressions: Grouping, Regular Expressions: Use Cases

Unit II

Module 1: Lexical Processing - II : Word Frequencies and Stop Words, Tokenisation, Bag-of-Words Representation, Stemming and Lemmatization, Final Bag-of-Words Representation, TF-IDF Representation, Building a Spam Detector, Canonicalisation, Phonetic Hashing, Edit Distance, Spell Corrector, Pointwise Mutual Information

Unit III

Module 2: Syntactic Processing - I : Syntax and Syntactic Processing, Parts of Speech, PoS Tagging, Hidden Markov Model, PoS Tagging Application, PoS Tagging Case Study, Constituency Parsing, Dependency Parsing, Parsing - Python Demonstration

Unit IV

Module 2: Syntactic Processing - II : Named Entity Recognition, How to do Named Entity Recognition, IOB Labeling, NER: Python Demonstration, Conditional Random Field: Overview, CRF: Model Training Part-I, CRF: Model Prediction, Custom NER: Python Implementation

Unit V

Module 4: Semantic Processing - I : Knowledge Graphs, WordNet - Relation Between Senses, WordNet - Code Demo, Word Sense Disambiguation, Word Sense Disambiguation - Code Demonstration, Inspiration for Semantic Processing, Geometric Representation of Meaning, Cosine Similarity, Bag of Words Representation

Unit VI

Module 4: Semantic Processing - II : Intuition of Word2Vec Model, Recall for Forward Pass, Input Data for CBOW, Training of CBOW Model, Weight Matrices, Skip-Gram Model, Code Demonstration - Gensim, Architecture for Binary Classification, Training Data for Negative Sampling, Text Pre-Processing, Code Demonstration, Topic Modeling, Non-Negative Matrix Factorisation, Code Demonstration: Topic Modeling on IMDb Review

List of Practicals / Experiments:

List of Practical / Experiments:

- Tokenizing a Given Text - Print the tokens of the text document.
- Tokenize Text With Stopwords As Delimiters - Tokenize the text with stop words (“is”,”the”,”was”) as delimiters. Tokenizing this way should identifies meaningful phrases.
- Stop Words Removal - Remove all the stopwords (‘a’ , ‘the’, ‘was’…) from the text.
- Adding Stop Words - Add the custom stopwords “NIL” and “JUNK” in spaCy and remove the stopwords in text.
- Stemming - Perform stemming/ convert each token to it’s root form in the text.
- Lemmatization - Perform lemmatization on the text.

- Extracting Usernames - Extract the usernames from the email addresses present in the text.
- Extracting Top Common Words - Extract the top 10 most common words in the text excluding stopwords.
- Tokenize Tweets - Clean the tweet and tokenize them.
- Extracting Nouns - Extract and print all the nouns present in the text.
- Cost Function - Find the similarity between any two words.
- Word Mover Distance - Compute the word mover distance between two texts.
- Replacing Pronouns - Replace the pronouns in text by the respective object names.
- Topic Extraction - Extract the topics from the texts with the help of NMF (Non-negative Matrix Factorization method).
- Word Vector Extraction - Extract the word vector representation of the word using word2vec model.
- Implementing Word embedding - Implement Word embedding on the texts and visualize it.
- TF-IDF Matrix Extraction - Extract the TF-IDF (Term Frequency -Inverse Document Frequency) Matrix for the list of text documents.
- Merging Two Tokens As One - Merge the first name and last name as single token in the sentence.
- Replace All Names - Identify and replace all the person names in the news article with UNKNOWN to keep privacy.
- Visualizing the Dependency Tree - Visualize the dependencies of various tokens of the text using spaCy.

Text Books:

1. FUNDAMENTALS OF NATURAL LANGUAGE PROCESSING by FIROS A, NOTION PRESS CHENNAI

References:

1. NATURAL LANGUAGE PROCESSING by ELA KUMAR, DREAMTECH PRESS