

## Basics of Text Processing

### Learning Objectives:

Learn about the interaction between computers and human beings which gives computers the ability to understand human speech with the help of machine learning. Understand the concept behind tokenization and normalization

### Topics

- Introduction to Regular Expressions
- Tokenization of text
- Normalization of text
- Substituting and correcting tokens
- Applying Zipf's law to text
- Applying similarity measures using the Edit Distance Algorithm
- Applying similarity measures using Jaccard's Coefficient
- Applying similarity measures using Smith Waterman

### Hands-on:

Apply various similarity measures to strings using NLTK

## Statistical Language Modeling

### Learning Objectives:

Understand the preprocessing tasks or the computations that can be performed on natural language text. Learn about the ways to calculate word frequencies, the Maximum Likelihood Estimation (MLE) model, interpolation on data, and so on

### Topics

- Understanding word frequency
- Applying smoothing on the MLE model
- Develop a backup mechanism for MLE
- Data Interpolation
- Language modelling using metropolis hastings
- Gibbs sampling in language processing

### Hands-on:

Implement Maximum Likelihood Estimation in NLTK and perform language modeling

## Morphological Modeling

### Learning Objectives:

Learn about stemming and lemmatization, stemmer and lemmatizer for non-English languages, developing a morphological analyzer and morphological generator using machine learning tools, search engines, and many such concepts

### **Topics**

- Introducing Morphology
- Understanding stemmer
- Lemmatization
- Morphological analyzer
- Morphological generator

### **Hands-on:**

Perform preprocessing on the original text in order to implement or build an application. Implement stemming, lemmatization, and morphological analysis and generation in NLTK

## **Syntactic Analysis**

### **Learning Objectives:**

Understand the process of finding whether a character sequence, written in natural language, is in accordance with the rules defined in formal grammar. Also, learn about the process of breaking the sentences into words or phrase sequences and providing them a particular component category (noun, verb, preposition, and so on)

### **Topics**

- Introducing Parsing
- Treebank construction
- Extracting Context Free Grammar (CFG) rules from Treebank
- CYK chart parsing algorithm
- Earley chart parsing algorithm

### **Hands-on:**

Implement Context-free Grammar, Probabilistic Context-free Grammar, the CYK algorithm and the Earley algorithm

## **Semantic Analysis**

### **Learning Objectives:**

Understand the process of determining the meaning of character sequences or word sequences which may be used for performing the task of disambiguation

### **Topics**

- Introducing semantic analysis
- Named-entity recognition (NER)

- NER system using the HMM
- Training NER using machine learning toolkits
- NER using POS tagging
- Generation of the synset id from Wordnet
- Disambiguating senses using Wordnet

## **Sentiment Analysis**

### **Learning Objectives:**

Understand the process of determining the sentiments behind a character sequence. It may be used to determine whether the speaker or the person expressing the textual thoughts is in a happy or sad mood, or it represents a neutral expression

### **Topics**

- Introducing sentiment analysis
- Sentiment analysis using NER
- Sentiment analysis using machine learning
- Evaluation of the NER system

## **Information Retrieval**

### **Learning Objectives:**

Understand the process of retrieving information (for example, the number of times the word "Analysis" has appeared in the document) corresponding to a query that has been made by the user

### **Topics**

- Introducing information retrieval
- Stop word removal
- Information retrieval using a vector space model
- Vector space scoring and query operator interactions
- Text summarization

### **Hands-on:**

Implement text summarization, question-answering systems, and vector space models

## **Discourse Analysis**

### **Learning Objectives:**

Understand the process of determining contextual information that is useful for performing other tasks, such as anaphora resolution (AR), NER, and so on

## **Topics**

- Introducing discourse analysis
- Discourse analysis using Centering Theory
- Anaphora resolution

## **Hands-on:**

Use NLTK to implement first order predicate logic using UML diagrams

## **Evaluation of NLP Systems – Analyzing Performance**

### **Learning Objectives:**

Learn to analyze whether a given NLP system produces the desired result or not and the desired performance is achieved or not which may be performed automatically using predefined metrics, or it may be performed manually by comparing human output with the output obtained by an NLP system

### **Topics**

- The need for the evaluation of NLP systems
- Evaluation of IR Systems
- Metrics for error identification
- Metrics based on lexical matching
- Metrics based on syntactic matching
- Metrics using shallow semantic matching