NATURAL LANGUAGE PROCESSING
[Subject Code: DSE 3155]

Class:

Faculty: Mrs. Rashmi M, Dr Savitha G

What is Natural Language Processing (NLP)?

- It is a subfield or branch of Artificial intelligence (AI) that enables computers to understand human languages and process them in a manner that is valuable. It concerns the interactions between human spoken (natural) languages like English and computers.

- It is also known as Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)

What is Natural Language Processing (NLP)?

- NLP is an interdisciplinary field that uses computational methods to:
  - Investigate the properties of written human language and model the cognitive mechanisms underlying the understanding and production of written language.
  - Develop novel practical applications involving the intelligent processing of written human language by computer.

Goal of NLP

- Find new methods of communication between humans and computers, as well as to grasp human speech as it is uttered.
- Analyze, understand and generate human languages just like humans do
- Applying computational techniques to language domain
- To explain linguistic theories, to use the theories to build systems that can be of social use
- Make computers learn our language rather than we learn theirs
- Combine machine learning with computational linguistics, Cognitive Science, statistics, and deep learning models so that computers can process human language from voice or text data and grasp its entire meaning, as well as the writer or speaker's intentions.

Steps in Interaction between Human and Machine using NLP

1. A human talks to the machine
2. The machine captures the audio
3. Audio to text conversion takes place
4. Processing of the text’s data
5. Data to audio conversion takes place
6. The machine responds to the human by playing the audio file

History of NLP

- In 1950, Alan Turing published an article titled "Machine and Intelligence" which advertised what is now called the Turing test as a subfield of intelligence
- Some beneficial and successful Natural language systems were developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks of words" with restricted vocabularies was written between 1964 to 1966

Main techniques used in NLP

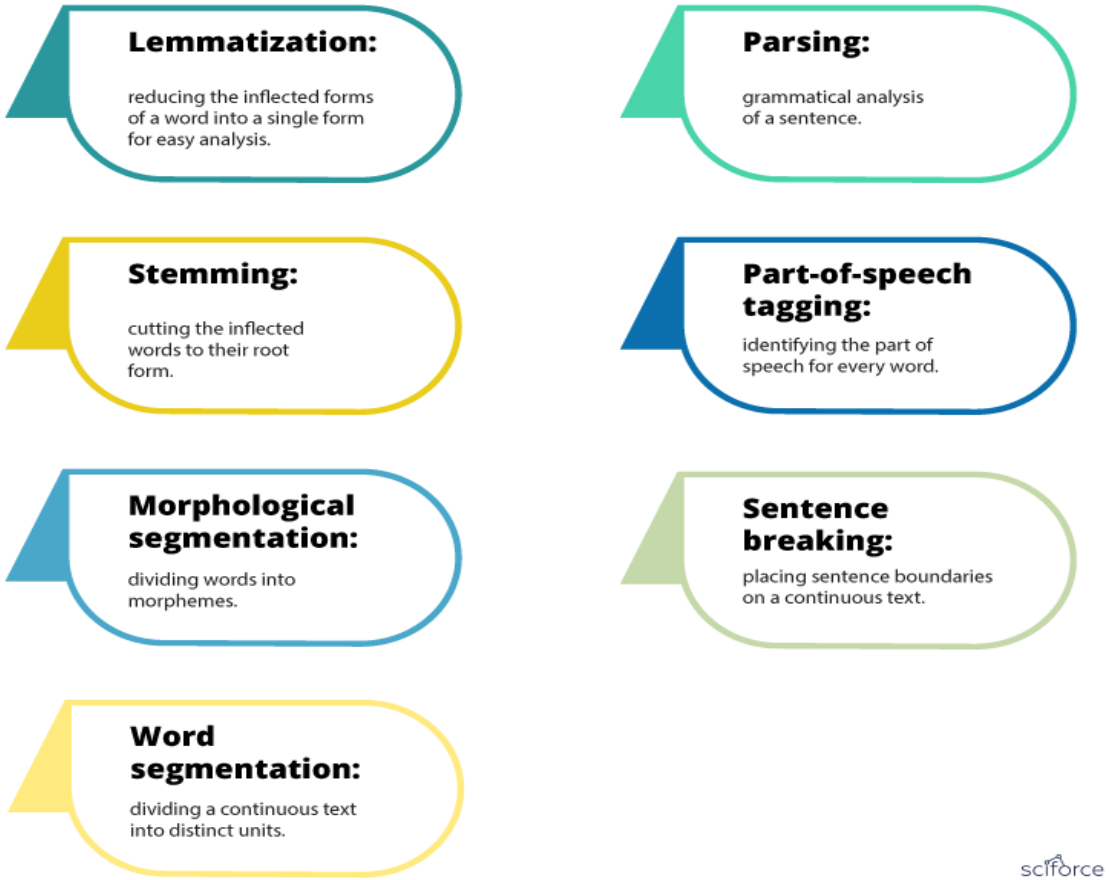

Fig credit:https://towardsdatascience.com/

- **Lemmatization**: It entails reducing the various inflected forms of a word into a single form for easy analysis.

  Ex: *“running”, “runs”* and *“ran”* are all forms of the word *“run”*, so *“run”* is the lemma of all these words.
- **Morphological segmentation**: It involves dividing words into individual units called morphemes.
- **Word segmentation**: It involves dividing a large piece of continuous text into distinct units.
- **Part-of-speech tagging**: It involves identifying the part of speech for every word.
- **Parsing**: It involves undertaking a grammatical analysis for the provided sentence.
- **Sentence breaking**: It involves placing sentence boundaries on a large piece of text.
- **Stemming**: It involves cutting the inflected words to their root form.

Syntax Analysis

- Syntax refers to the arrangement of words in a sentence such that they make grammatical sense.
- In NLP, syntactic analysis is used to assess how the natural language aligns with the grammatical rules.
- Computer algorithms are used to apply grammatical rules to a group of words and derive meaning from them.

Semantic Analysis

- Semantics refers to the meaning that is conveyed by a text.
- It involves applying computer algorithms to understand the meaning and interpretation of words and how sentences are structured.

Techniques in semantic analysis:

- **Named entity recognition (NER):** It involves determining the parts of a text that can be identified and categorized into preset groups. Examples of such groups include names of people and names of places.
- **Word sense disambiguation:** It involves giving meaning to a word based on the context.
- **Natural language generation**: It involves using databases to derive semantic intentions and convert them into human language.

Sample Applications of NLP

- Language translation application such as google translate.
- Word processors such as Microsoft word Grammarly that employ NLP to check grammatical accuracy of the text.
- Interactive Voice Response (IVR) applications used in call centers to respond to certain users' requests.
- A personal assistant application such as OK Google. Hay siri, and Alexa.

Course Objectives

- CO1: To understand the concepts for the processing of linguistic information and computational properties of natural languages.
- CO2:To conceive basic knowledge on various morphological and lexical NLP tasks
- CO3:To familiarize the essential probabilistic architecture and word prediction model
- CO4:To understand the probabilistic models for Syntactic Analysis in NLP

Topics to study

L0	Introduction:Natural Language Processing
L1	Knowledge in Speech and Language Processing, Ambiguity, Models and Algorithm
L2	Basics of Finite State Automata
L3	Regular Expressions
L4	Survey of English Morphology-Inflectional and Derivational
L5	Finite-State Morphological Parsing-Lexicon and Morphotactics
L6	Morphological Parsing with FST
L7	Orthographic Rules and FST
L8	Combining FST Lexicon and Rules
L9	Lexicon -Free FSTs:The Porter Stemmer
L10	Words and Sentence tokenization

Topics to study

L11	Text Normalization
L12	Segmentation
L13	Probablistic Models of Pronunciation and Spelling -Dealing with Spelling errors
L14	Spelling Error Patterns
L15	Probablistic Models-Noisy Channel Model
L16	Applying the Bayesian Method to Spelling
L17	Minimum Edit Distance
L18	NGrams-Counting words in Corpora
L19	Unsmoothed N-Grams
L20	More on N-grams and Their Sensitivity

Topics to study

L21	Smoothing
L22	Backoff
L23	Deleted Interpolation
L24	English Word Classes
L25	Tag-sets for English
L26	Part-of-Speech Tagging
L27	Case Study:Automatic Tagging
L28	Context Free Grammars for English
L29	Consistuency
L30	Context Free Rules and Trees
L31	The Penn Treebank Project
L32	Grammar Equivalence and Normal form
L33	Parsing with Context Free Grammars
L34	Probablistic Context Free Grammar-CYK
L35	Dependency Grammar
L36	Statistical Parsing

Text Books

1. J.E.Hopcroft, R.Motwani & J.D.Ullman , Introduction to Automata Theory Languages, and Computation, (3rd Edition) , Pearson Education.

2. Daniel Jurafsky & James H. Martin, Speech and Language Processing, (2e), Pearson, 2009.

3. Steven Bird, Ewan Klein and Edward Loper, Natural Language Processing with Python, (1e), O'Reilly Media, 2009

4. Akshar Bharati, Rajeev Sangal and Vineet Chaitanya, Natural Language Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi, 1995

5. Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python – Analysing Text with natural language toolkit, O'Reilly Media, 2009

6. Chris Manning, Hinrich Schutze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, 1999