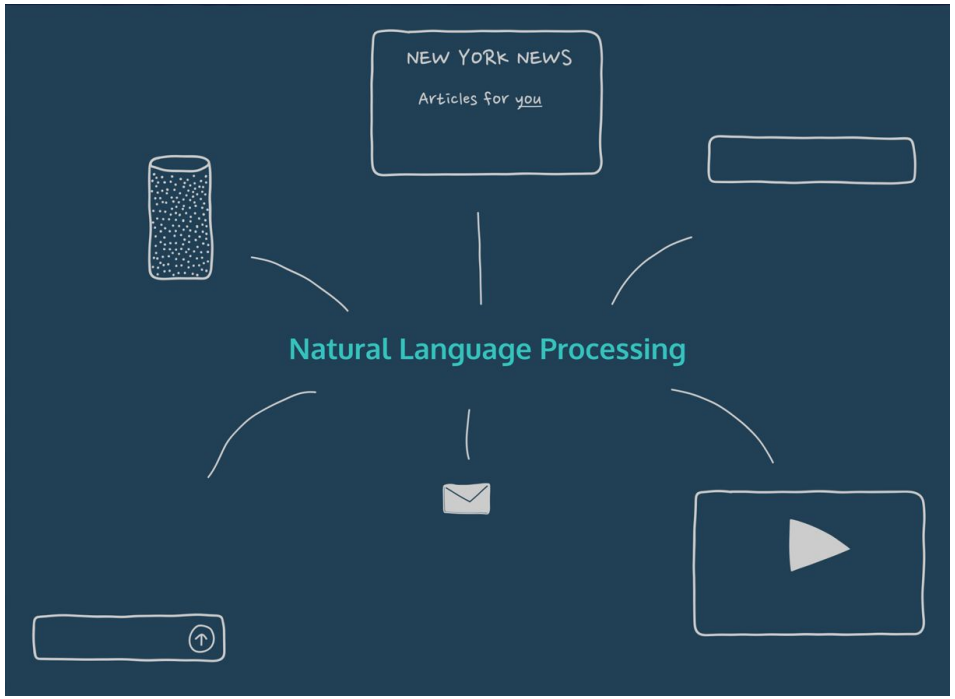
Session 0

NATURAL LANGUAGE PROCESSING - SUBJECT OVERVIEW



Prof. Manan Thakkar

Computer Engineering
(Artificial Intelligence)

Ganpat University
॥ विद्यया समाजोत्कर्षः ॥

Vision and Mission

Vision and Mission of Ganpat University	
Vision	It shall be the constant endeavour of Ganpat University to meet the educational needs of the youth in the areas of professional studies and provide state-of the art learning opportunities along with inculcation of values of commitment and uprightness.
Mission	Seek, search and offer programs that lead to symbiotic emergence of 'academic excellence' and 'industrial relevance' in education and research.

Course Outcomes

- To understand natural language processing and importance of word representation.
- Apply deep learning to solve natural language problems such as language modelling, machine translation, POS tagging, Seq2Seq generation.
- Solve NLP problem in Indian context (Indian languages).

Program Outcomes

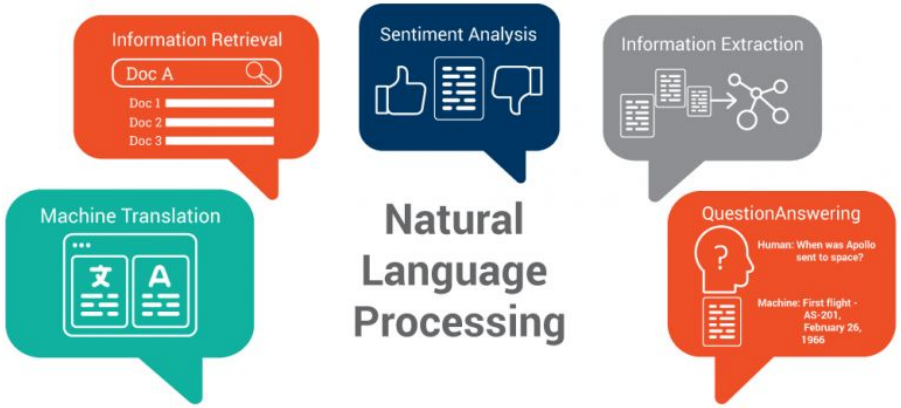
Program Outcomes (POs)		
Program: **B.Tech.- CE** **B.Tech.- IT**	**Name of PO**	**Description**
	PO1	**Engineering Knowledge:** Apply knowledge of mathematics, science, engineering fundamentals and an engineering specialization to the solution of complex engineering problems.
	PO2	**Problem Analysis:** Identify, formulate, research literature and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences and engineering sciences.
	PO3	**Design/ Development of Solutions:** Design solutions for complex engineering problems and design system components or processes that meet specified needs with appropriate consideration for public health and safety, cultural, societal and environmental considerations.
	PO4	**Conduct investigations of complex problems** using research-based knowledge and research methods including design of experiments, analysis and interpretation of data and synthesis of information to provide valid conclusions.
	PO5	**Modern Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
	PO6	**The Engineer and Society**: Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice.
	PO7	**Environment and Sustainability:** Understand the impact of professional engineering solutions in societal and environmental contexts and demonstrate knowledge of and need for sustainable development.
	PO8	**Ethics:** Apply ethical principles and commit to professional ethics, responsibilities, and norms of engineering practice.
	PO9	**Individual and Team Work:** Function effectively as an individual, and as a member or leader in diverse teams and in multi-disciplinary settings.
	PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as comprehending and writing effective reports and design documentation, making effective presentations, and giving and receiving clear instructions.
	PO11	**Project Management and Finance:** Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
	PO12	**Life-long Learning:** Recognize the need for and have the preparation and ability to Engage in independent and life- long learning in the broadest context of technological Change.

Program Specific Outcomes (PSOs)		
Program: **B.Tech.- CE** **B.Tech.- IT**	**Name of PSO**	**Description**
	PSO1	Acquire skills to design, **analyse and develop** algorithms and implement those using high-level programming languages.
	PSO2	Contribute their engineering **skills** in computing and information engineering domains like network design and administration, database design and knowledge engineering.
	PSO3	Develop strong skills in systematic planning, developing, testing, implementing and providing IT solutions for different domains which helps in the **betterment of life.**

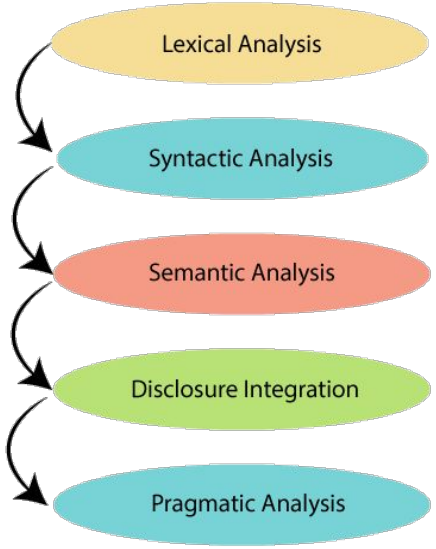
Bloom’s Taxonomy

- R: Remembering,
- U: Understanding,
- A: Applying,
- N: Analyzing,
- E: Evaluating,
- C: Creating

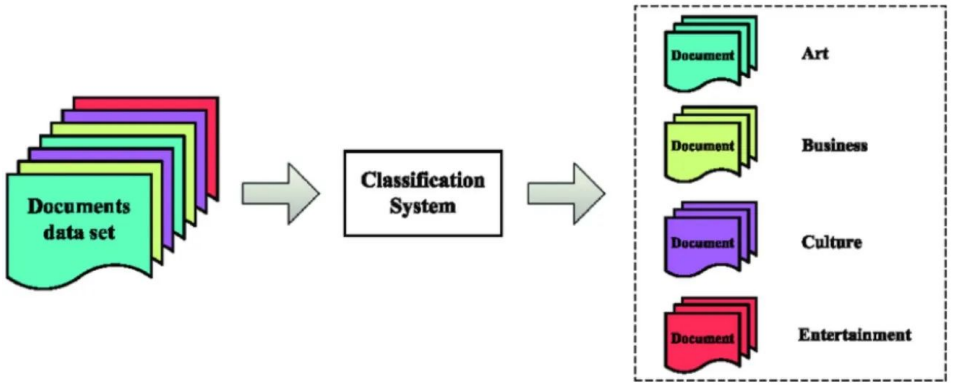
TECHNOLOGY STACKS TO STUDY



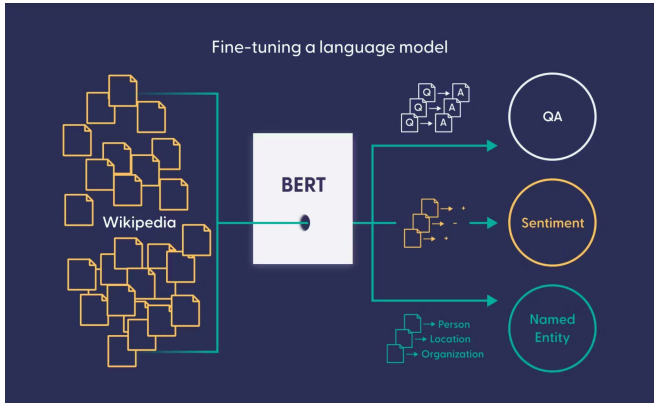
Introduction to NLP



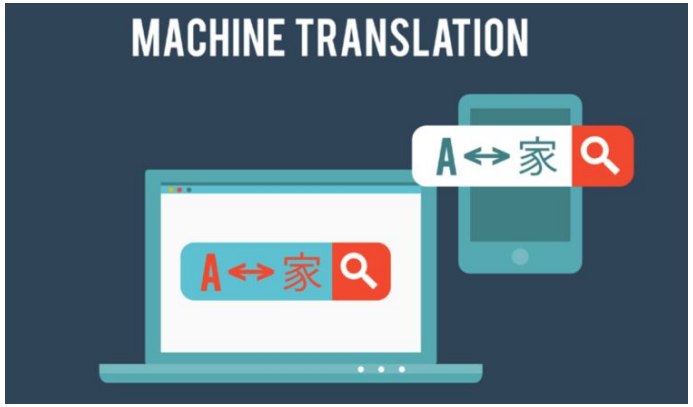
Text Processing in NLP



Text Classification



Language Modelling



Machine Translation

Syllabus

https://d2z4x7fn3a0wyp.cloudfront.net/subject/natural-language-processing/natural-language-processing.pdf

GANPAT UNIVERSITY

FACULTY OF ENGINEERING & TECHNOLOGY

Programme	Bachelor of Technology	Branch/Spec.	Computer Engineering (Artificial Intelligence)
Semester	VI	Version	1.0.0.0
Effective from Academic Year	2021-22	Effective for the batch Admitted in	July 2019
Subject code	2CEAI602	Subject Name	Natural Language Processing

Teaching scheme						Examination scheme (Marks)			
(Per week)	Lecture(DT)		Practical(Lab.)		Total		CE	SEE	Total
	L	TU	P	TW					
Credit	3	0	1	-	4	Theory	40	60	100
Hours	3	0	2	-	5	Practical	30	20	50

Pre-requisites:

Basic knowledge of python programming

Objectives of the course:

- To introduce the fundamental concepts and techniques of Natural language Processing for analyzing words based on Morphology and CORPUS
- To examine the NLP models and interpret algorithms for classification of NLP sentences by using both the traditional, symbolic and the more recent statistical approach
- To get acquainted with the algorithmic description of the main language levels that includes morphology, syntax, semantics, and pragmatics for information retrieval and machine translation applications

Theory syllabus:

Unit	Content	Hrs
1.	**Introduction to NLP:** Introduction to various levels of natural language 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.	03
2.	**Text Processing:** Ambiguity in language, Segmentation, Stemming, Tokenization, Representation of word, Sentence, Word embedding, Word Senses, Linguistic Structure: Dependency Parsing.	09
3.	**Text Classification:** Word Window Classification, Neural Networks for text, N-gram Language Models, Perplexity, Viterbi algorithm, Recurrent Neural network, Vanishing Gradients and exploding gradients.	13
4.	**Language Modelling:** The role of language models, Estimating parameters and smoothing. Evaluating language models, LSTM (Long sort 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.	10
5.	**Machine Translation:** Statically Machine Translation, Neural Machine Translation, Seq2Seq Modelling, Attention, Question Answering Bot, Natural Language Generation, Neural Machine Translation.	07
6.	**NLP Case Study:**	03

Practical content:

- Experiments/ Practicals /Simulations would be carried out based on syllabus

Text Book:

1.	Daniel Jurafsky and James H. Martin, Speech and Language processing an introduction to Natural Language Processing,Computational Linguist (Latest Edition), Prentice Hall, 2008. ISBN 978-0131873216.

Reference Books:

1.	Steven Bird, Ewan Klein and Edward Lopper, Natural Language Processing with Python (Latest Edition), O'Reilly, 2009. ISBN 978-0596516499.
2.	Siddiqui and Tiwari, Natural Language Processing and Information Retrieval (Latest Edition), Oxford University Press, 2008. ISBN 978-0195692327.
3.	Nitin Indurkhya, Fred J. Damerau and Fred J. Damerau, Handbook of Natural Language Processing (Latest Edition), Taylor and Francis, 2010. ISBN 978-1420085921.
4.	Allen J., Natural Language understanding (Latest Edition), Pearson, 1994. ISBN 978-0805303346.

ICT/MOOCs Reference:

1.	https://nptel.ac.in/courses/106/105/106105158/
2.	https://nptel.ac.in/courses/106/106/106106211/

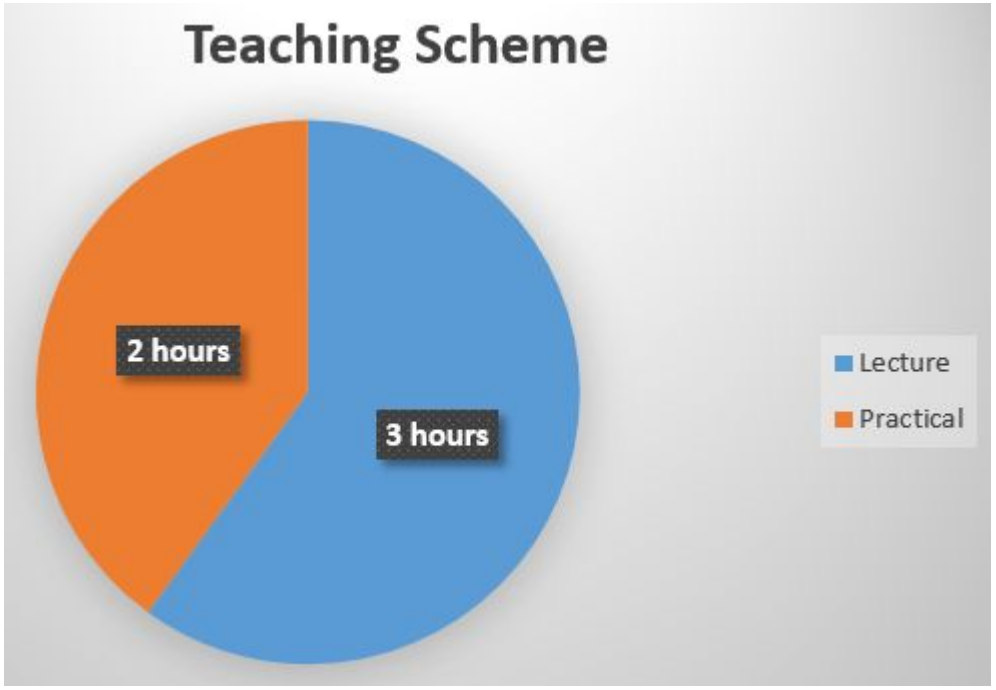
Course Outcomes:

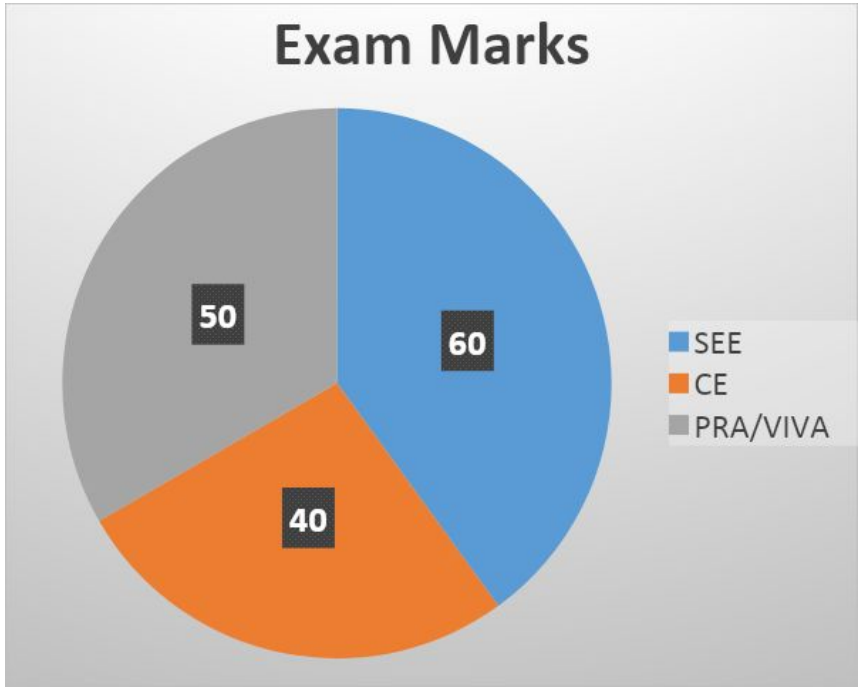
COs	Description
CO1	To understand natural language processing and importance of word representation
CO2	Apply deep learning to solve natural language problems such as language modelling, machine translation, POS tagging, Seq2Seq generation
CO3	Solve NLP problem in Indian context (Indian languages)

Mapping of CO and PO:

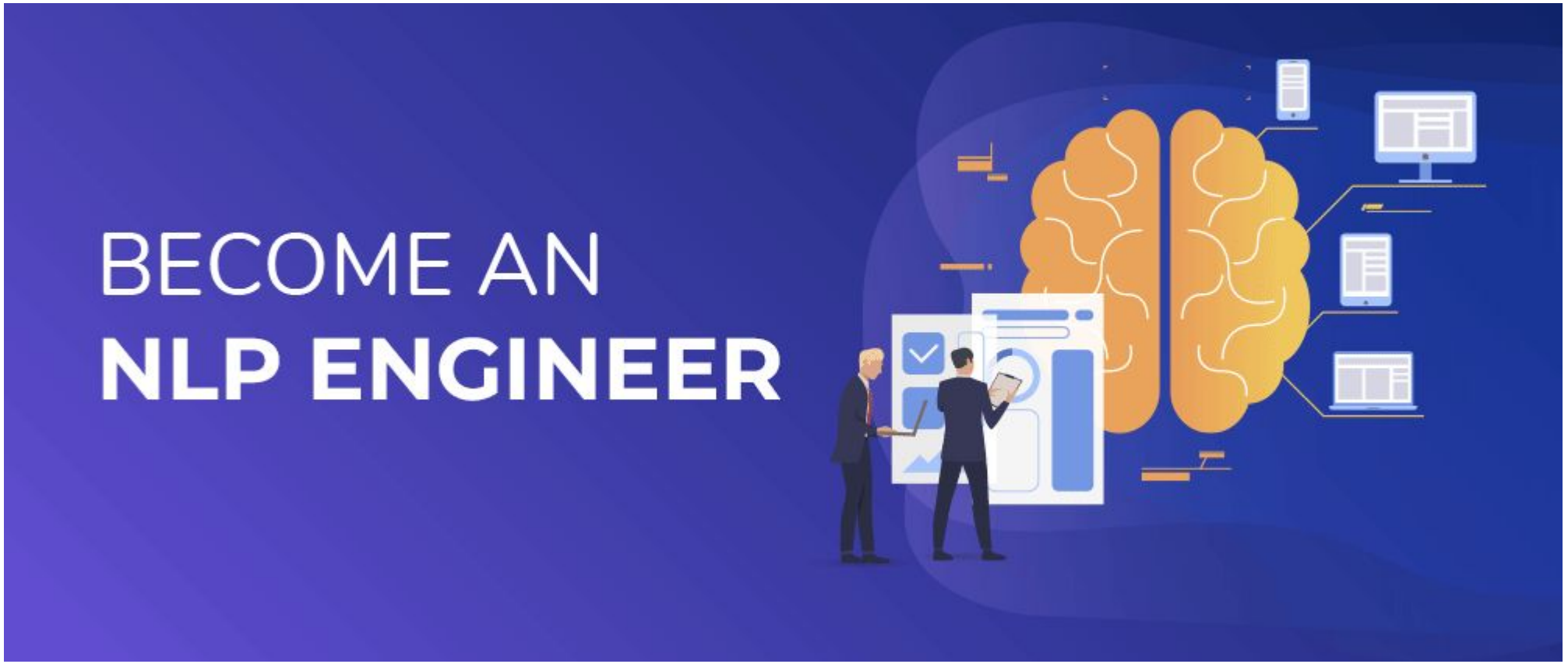
COs	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	3	3	1	2	2	0	0	0	0	0	0	0
CO2	3	2	3	2	2	0	1	0	1	0	1	1
CO3	3	1	3	3	3	0	2	0	1	0	0	1

Teaching Scheme





Employment Opportunities



NLP Engineer

Tasks

- Design and develop NLP systems
- Define appropriate datasets for language learning
- Use effective text representations to transform natural language into useful features
- Develop NLP systems according to requirements
- Train the developed model and run evaluation experiments
- Find and implement the right algorithms and tools for NLP tasks

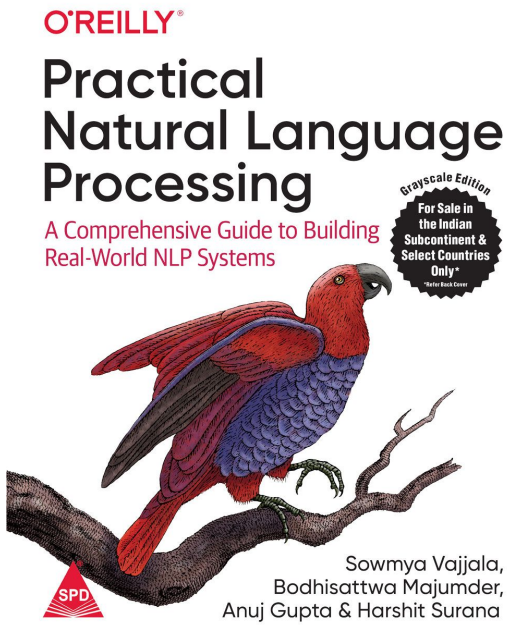
NLP Engineer

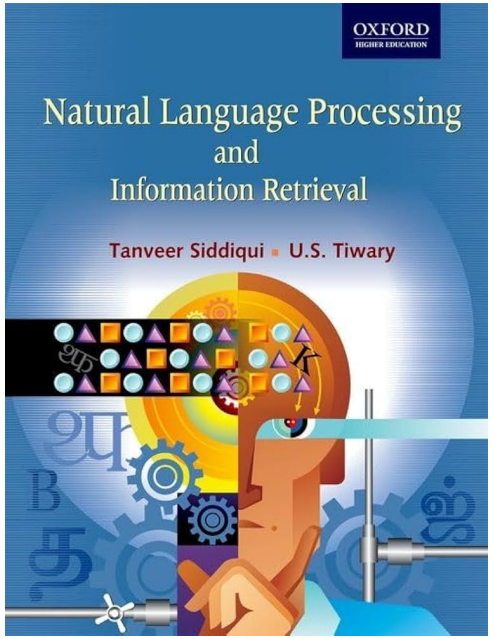
Skills

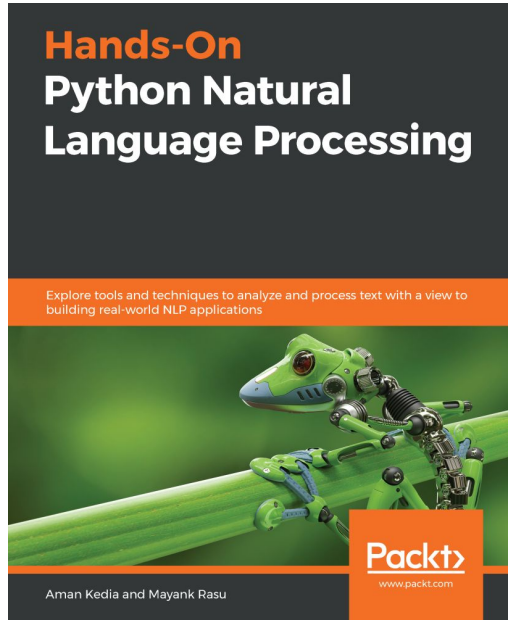
- Deep understanding of text representation techniques
- Machine Translation & Compilers
- Experience with machine learning frameworks and libraries
- Text Classification & Clustering
- Knowledge of Python, Java and R
- An analytical mind with problem-solving abilities

Industries Involved

Books and Links – For Reference







https://www.youtube.com/@DigitalSreeni

https://www.youtube.com/@CodeWithAarohi

https://www.youtube.com/watch?v=SMZQrJ_L1vo&t=2791s

UNIT 1: Introduction to NLP

- Introduction to various levels of natural language 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.

Introduction

- Natural languages are different from computer programming languages.
- Natural languages are what humans use to share information with each other. We don't use programming languages to tell each other about our day or to give directions to the grocery store.
- A computer program written with a programming language tells a machine exactly what to do. But there are no compilers or interpreters for natural languages such as English and French.

Introduction

- Natural language processing is an area of research in computer science and artificial intelligence (AI) concerned with processing natural languages such as English or Mandarin.
- This processing generally involves translating natural language into data (numbers) that a computer can use to learn about the world.
- And this understanding of the world is sometimes used to generate natural language text that reflects that understanding.

Introduction

- **DEFINITION:** A natural language processing system is often referred to as a pipeline because it usually involves several stages of processing where natural language flows in one end and the processed output flows out the other.

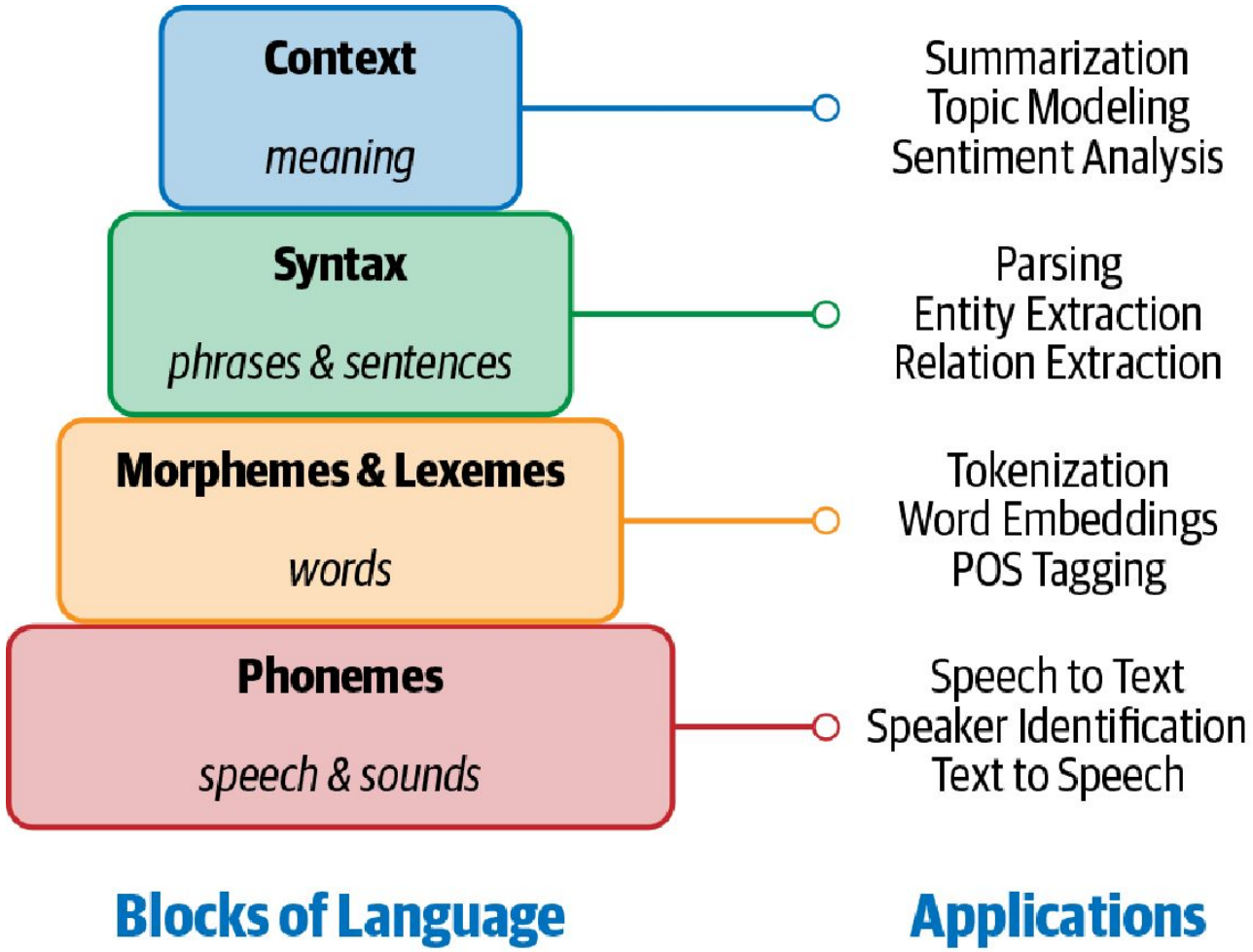
Natural Language Processing Stages

- Natural language simply refers to the way we communicate with each other: speech and text.
- Processing refers to making natural language usable for computational tasks.
- Natural language processing, also referred to as text analytics, plays a very vital role in today's era because of the sheer volume of text data that users generate around the world on digital channels such as social media apps, e-commerce websites, blog posts, etc.

Natural Language Processing Stages

- We can think of human language as composed of four major building blocks:
  - phonemes,
  - morphemes and lexemes,
  - syntax, and
  - context.
- NLP applications need knowledge of different levels of these building blocks, starting from the basic sounds of language (phonemes) to texts with some meaningful expressions (context).

Natural Language Processing Stages



Building blocks of language and their applications

Stage 1: Phonemes

- The very first component of language is phonology. It is the study of the speech sounds of a particular language. The origin of the word can be traced to Greek language, where ‘phone’ means sound or voice. Phonetics, a subdivision of phonology is the study of the speech sounds of human language from the perspective of their production, perception or their physical properties.
- IPA (International Phonetic Alphabet) is a tool that represents human sounds in a regular way while studying phonology. In IPA, every written symbol represents one and only one speech sound and vice-versa.

Stage 1: Phonemes

- Phonemes may be defined as one of the units of sound that differentiate one word from other in a language. In linguistic, phonemes are written between slashes. For example, phoneme /k/ occurs in the words such as kit, skit.
- Standard English has 44 phonemes, which are either single letters or a combination of letters.
- Phonemes are particularly important in applications involving speech understanding, such as speech recognition, speech-to-text transcription, and text-to-speech conversion.

Stage 1: Phonemes

Consonant phonemes, with sample words		Vowel phonemes, with sample words	
1. /b/ - bat	13. /s/ - sun	1. /a/ - ant	13. /oi/ - coin
2. /k/ - cat	14. /t/ - tap	2. /e/ - egg	14. /ar/ - farm
3. /d/ - dog	15. /v/ - van	3. /i/ - in	15. /or/ - for
4. /f/ - fan	16. /w/ - wig	4. /o/ - on	16. /ur/ - hurt
5. /g/ - go	17. /y/ - yes	5. /u/ - up	17. /air/ - fair
6. /h/ - hen	18. /z/ - zip	6. /ai/ - rain	18. /ear/ - dear
7. /j/ - jet	19. /sh/ - shop	7. /ee/ - feet	19. /ure/[4] - sure
8. /l/ - leg	20. /ch/ - chip	8. /igh/ - night	20. /ə/ - corner (the 'schwa' - an unstressed vowel sound which is close to /u/)
9. /m/ - map	21. /th/ - thin	9. /oa/ - boat	
10. /n/ - net	22. /**th**/ - then	10. /**oo**/ - boot	
11. /p/ - pen	23. /ng/ - ring	11. /oo/ - look	
12. /r/ - rat	24. /zh/[3] - vision	12. /ow/ - cow	

Phonemes and examples

Stage 2: Morphemes and Lexemes

- A morpheme is the smallest unit of language that has a meaning. It is formed by a combination of phonemes.
- Not all morphemes are words, but all prefixes and suffixes are morphemes.
- For example, in the word "multimedia," "multi-" is not a word but a prefix that changes the meaning when put together with "media." "Multi-" is a morpheme.

Stage 2: Morphemes and Lexemes

- Figure illustrates some words and their morphemes. For words like "cats" and "unbreakable," their morphemes are just constituents of the full word, whereas for words like "tumbling" and "unreliability," there is some variation when breaking the words down into their morphemes.

unbreakable
un + break + able

cats
cat + s

tumbling
tumble + ing

unreliability
un + rely + able + ity

Stage 2: Morphemes and Lexemes

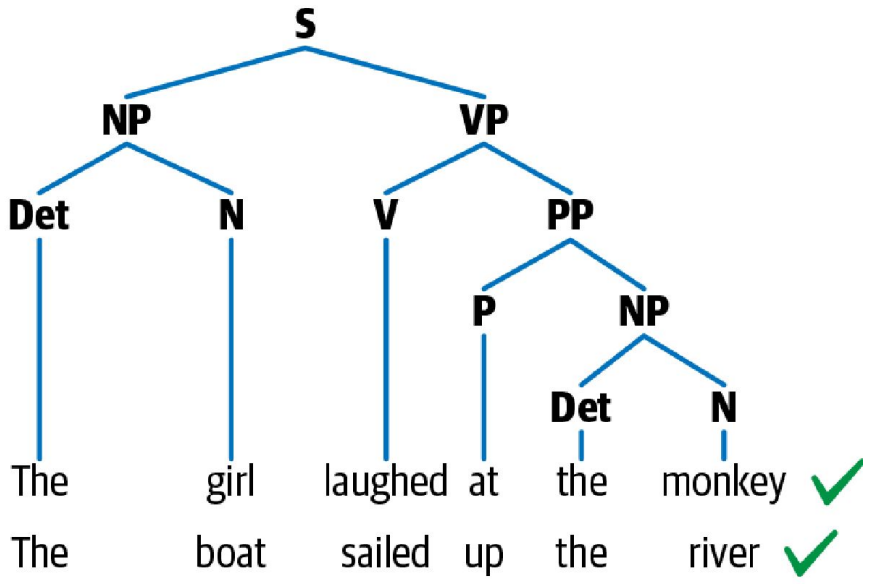
- Lexemes are the structural variations of morphemes related to one another by meaning. For example, “run” and “running” belong to the same lexeme form. Morphological analysis, which analyzes the structure of words by studying its morphemes and lexemes, is a foundational block for many NLP tasks, such as tokenization, stemming, learning word embeddings, and part-of-speech tagging.

Stage 2: Morphemes and Lexemes

- It involves identifying and analyzing the structure of words and parts of speech. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words. This level of linguistic processing utilizes a language's lexicon, which is a collection of individual lexemes.
- For example, the sentences "My cat ate its third meal" and "My third cat ate its meal", have very different meanings. However, lexical processing will treat the two sentences as equal, as the "group of words" in both sentences is the same. Hence, we clearly need a more advanced system of analysis.

Stage 3: Syntax

- Syntax is a set of rules to construct grammatically correct sentences out of words and phrases in a language. Syntactic structure in linguistics is represented in many different ways. A common approach to representing sentences is a parse tree.
- Figure shows an example parse tree for two English sentences.



Stage 3: Syntax

- This has a hierarchical structure of language, with words at the lowest level, followed by part-of-speech tags, followed by phrases, and ending with a sentence at the highest level.
- Both sentences have a similar structure and hence a similar syntactic parse tree. In this representation, N stands for noun, V for verb, and P for preposition. Noun phrase is denoted by NP and verb phrase by VP.

Stage 3: Syntax

- The two noun phrases are “The girl” and “The boat,” while the two verb phrases are “laughed at the monkey” and “sailed up the river.”
- The syntactic structure is guided by a set of grammar rules for the language (e.g., the sentence comprises an NP and a VP), and this in turn guides some of the fundamental tasks of language processing, such as parsing.
- Parsing is the NLP task of constructing such trees automatically.

Stage 4: Context

- Context is how various parts in a language come together to convey a particular meaning. Context includes long-term references, world knowledge, and common sense along with the literal meaning of words and phrases.
- The meaning of a sentence can change based on the context, as words and phrases can sometimes have multiple meanings.

Stage 4: Context

- Generally, context is composed from semantics and pragmatics. Semantics is the direct meaning of the words and sentences without external context. Pragmatics adds world knowledge and external context of the conversation to enable us to infer implied meaning.
- Complex NLP tasks such as sarcasm detection, summarization, and topic modeling are some of tasks that use context heavily.

Challenge with NLP

- What makes NLP a challenging problem domain?
- The ambiguity and creativity of human language are just two of the characteristics that make NLP a demanding area to work in.

Challenge with NLP - Ambiguity

- Ambiguity means uncertainty of meaning. Most human languages are inherently ambiguous.
- Consider the following sentence: "I made her duck." This sentence has multiple meanings. The first one is: I cooked a duck for her. The second meaning is: I made her bend down to avoid an object.
- Here, the ambiguity comes from the use of the word "made." Which of the two meanings applies depends on the context in which the sentence appears.

Challenge with NLP - Ambiguity

- When it comes to figurative language—i.e., idioms—the ambiguity only increases. For example, "He is as good as John Doe." Try to answer, "How good is he?" The answer depends on how good John Doe is.

Challenge with NLP - Ambiguity

- When it comes to figurative language—i.e., idioms—the ambiguity only increases. For example, "He is as good as John Doe." Try to answer, "How good is he?" The answer depends on how good John Doe is.
- Consider the examples in an image in next slide - These examples are easily disambiguated by a human but are not solvable using most NLP techniques.

Challenge with NLP - Ambiguity

The man couldn't lift his son because he was so **weak**. — Who was weak?

The man couldn't lift his son because he was so **heavy**. — Who was heavy?

Mary and Sue are **sisters**.

Mary and Sue are **mothers**. — How are Mary and Sue related?

Joan made sure to thank Susan for all the help she had **received**. — Who had received help?

Joan made sure to thank Susan for all the help she had **given**. — Who had given help?

John **promised** Bill to leave, so an hour later he left.

John **ordered** Bill to leave, so an hour later he left. — Who left an hour later?

Challenge with NLP – Common Knowledge

- A key aspect of any human language is "common knowledge." It is the set of all facts that most humans are aware of. In any conversation, it is assumed that these facts are known, hence they're not explicitly mentioned, but they do have a bearing on the meaning of the sentence.
- For example, consider two sentences: "man bit dog" and "dog bit man."

Challenge with NLP – Common Knowledge

- We all know that the first sentence is unlikely to happen, while the second one is very possible. Why do we say so? Because we all "know" that it is very unlikely that a human will bite a dog. Further, dogs are known to bite humans. This knowledge is required for us to say that the first sentence is unlikely to happen while the second one is possible.
- Note that this common knowledge was not mentioned in either sentence.

Challenge with NLP – Common Knowledge

- Humans use common knowledge all the time to understand and process any language. In the above example, the two sentences are syntactically very similar, but a computer would find it very difficult to differentiate between the two, as it lacks the common knowledge humans have. One of the key challenges in NLP is how to encode all the things that are common knowledge to humans in a computational model.

Challenge with NLP – Creativity

- Language is not just rule driven; there is also a creative aspect to it. Various styles, dialects, genres, and variations are used in any language. Poems are a great example of creativity in language. Making machines understand creativity is a hard problem not just in NLP, but in AI in general.

Real world Applications of NLP

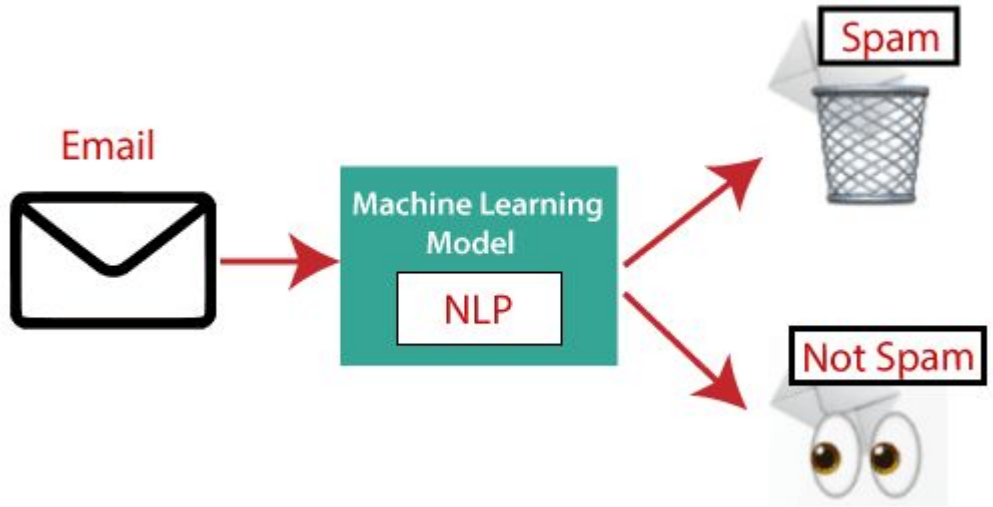
1. Question Answering

- Question Answering focuses on building systems that automatically answer the questions asked by humans in a natural language.

Real world Applications of NLP

2. Spam Detection

- Spam detection is used to detect unwanted e-mails getting to a user's inbox.



Real world Applications of NLP

3. Sentiment Analysis

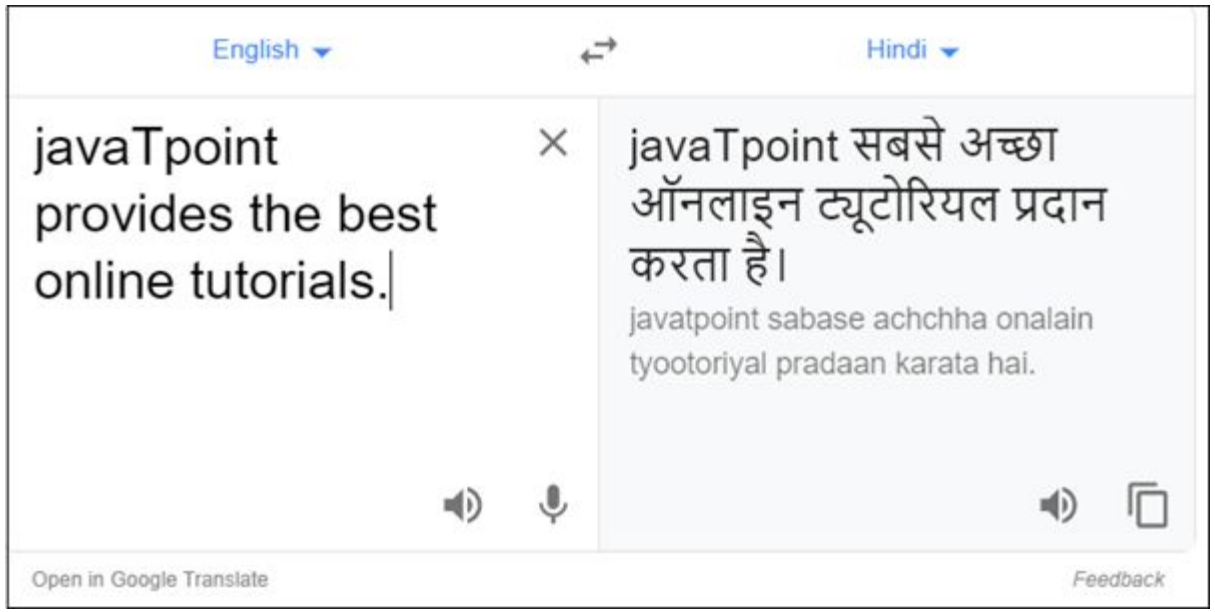
- Sentiment Analysis is also known as opinion mining. It is used on the web to analyse the attitude, behaviour, and emotional state of the sender. This application is implemented through a combination of NLP (Natural Language Processing) and statistics by assigning the values to the text (positive, negative, or natural), identify the mood of the context (happy, sad, angry, etc.)

Real world Applications of NLP

4. Machine Translation

•Machine translation is used to translate text or speech from one natural language to another natural language.

•**Example:** Google Translator



Real world Applications of NLP

5. Spelling correction

- Microsoft Corporation provides word processor software like MS-word, PowerPoint for the spelling correction.

Real world Applications of NLP

6. Speech Recognition

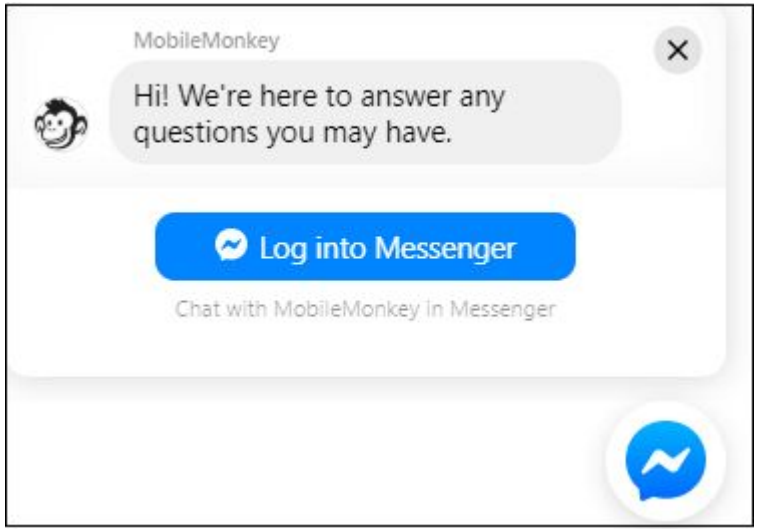
- Speech recognition is used for converting spoken words into text.

  It is used in applications, such as mobile, home automation, video recovery, dictating to Microsoft Word, voice biometrics, voice user interface, and so on.

Real world Applications of NLP

7. Chatbot

- Implementing the Chatbot is one of the important applications of NLP. It is used by many companies to provide the customer's chat services.



Real world Applications of NLP

8. Information extraction

- Information extraction is one of the most important applications of NLP. It is used for extracting structured information from unstructured or semi-structured machine-readable documents.

9. Natural Language Understanding (NLU)

- It converts a large set of text into more formal representations such as first-order logic structures that are easier for the computer programs to manipulate notations of the natural language processing.