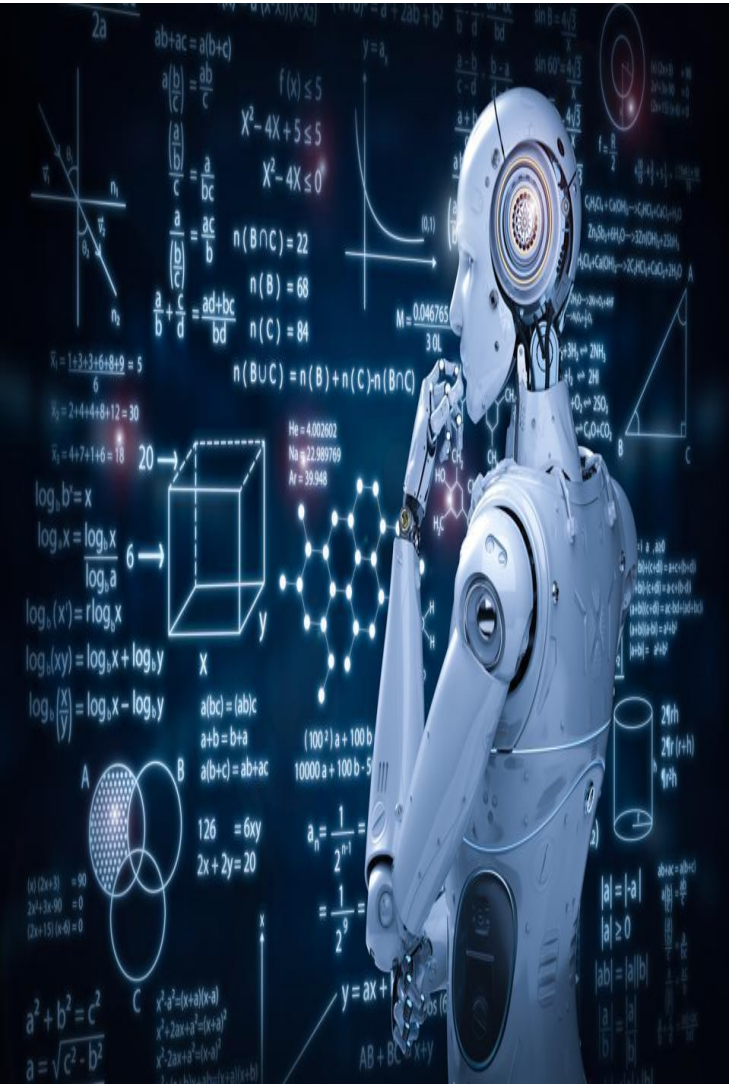


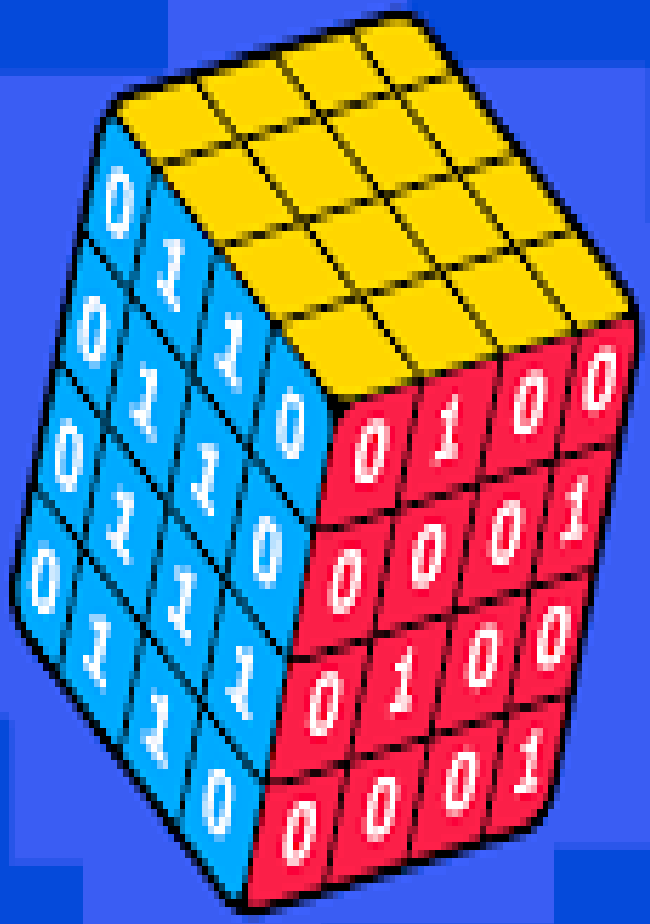
Practical Machine Learning

Day 16: Mar23 DBDA

Kiran Waghmare

Agenda

- Natural Language Processing



Structured Data

The Human Language

LANGUAGE

ALPHABETS

字母 वर्णमाला ALFABETOS

الأبجدية எழுத்துக்கள்

Words form Sentences

The Human Language

LANGUAGE

ALPHABETS

字母 वर्णमाला ALFABETOS

الأبجدية எழுத்துக்கள்

Words form Sentences

What is Text Mining ?

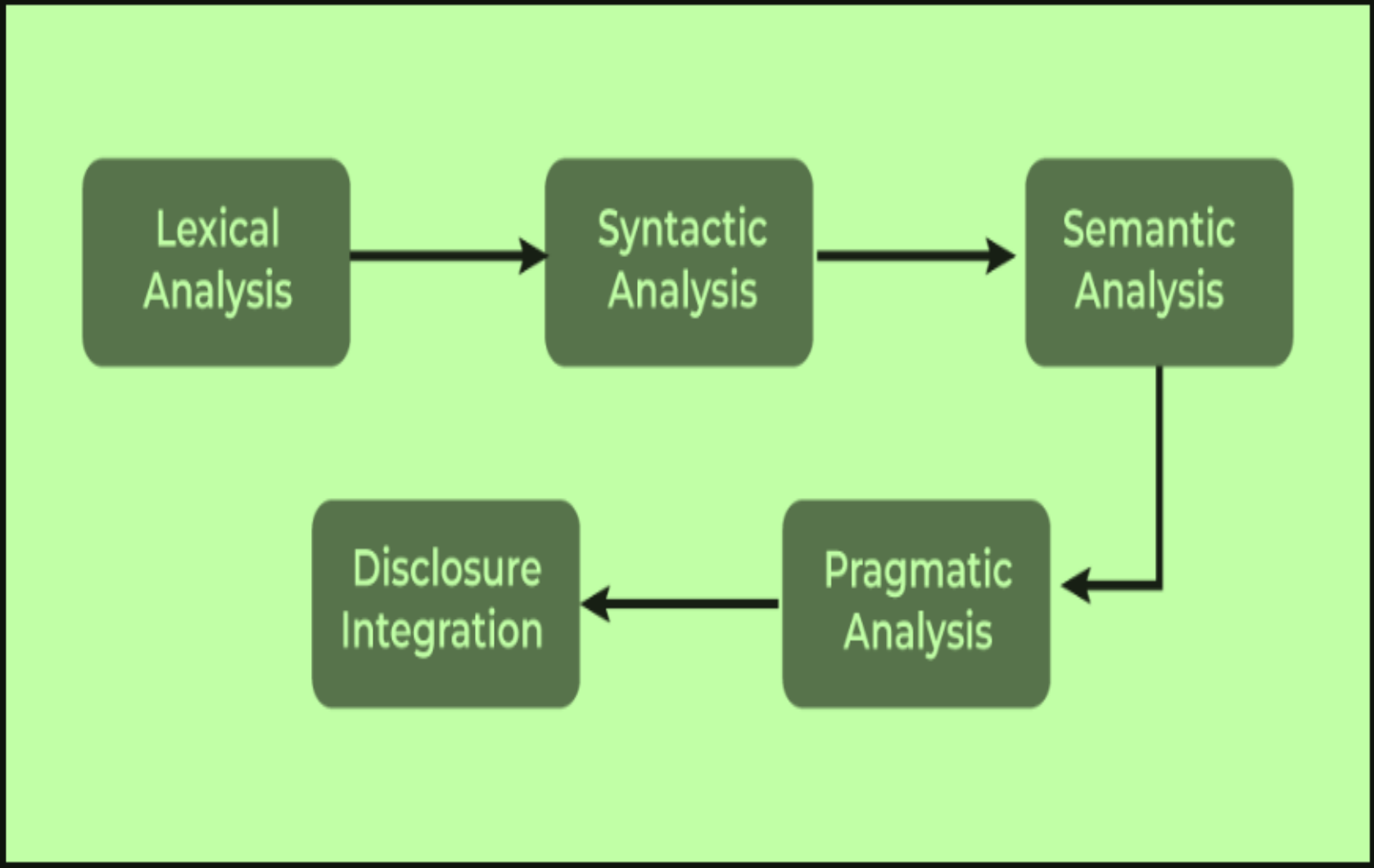
Text Mining / Text Analytics is the process of deriving meaningful information from natural language text

What is NLP?

NLP: Natural Language Processing is a part of computer science and artificial intelligence which deals with human languages.

- NLP stands for Natural Language Processing.
- It is the branch of Artificial Intelligence that gives the ability to machine understand and process human languages.
- Human languages can be in the form of text or audio format.

Phases of Natural Language Processing



Applications of NLP

Sentimental Analysis

Chatbot

Speech Recognition

Machine Translation

Applications of NLP and Text Mining

Spell Checking

Information Extraction

Keyword Search

Advertisement Matching

SENTIMENT ANALYSIS

POSITIVE

"Great service for an affordable price.
We will definitely be booking again."

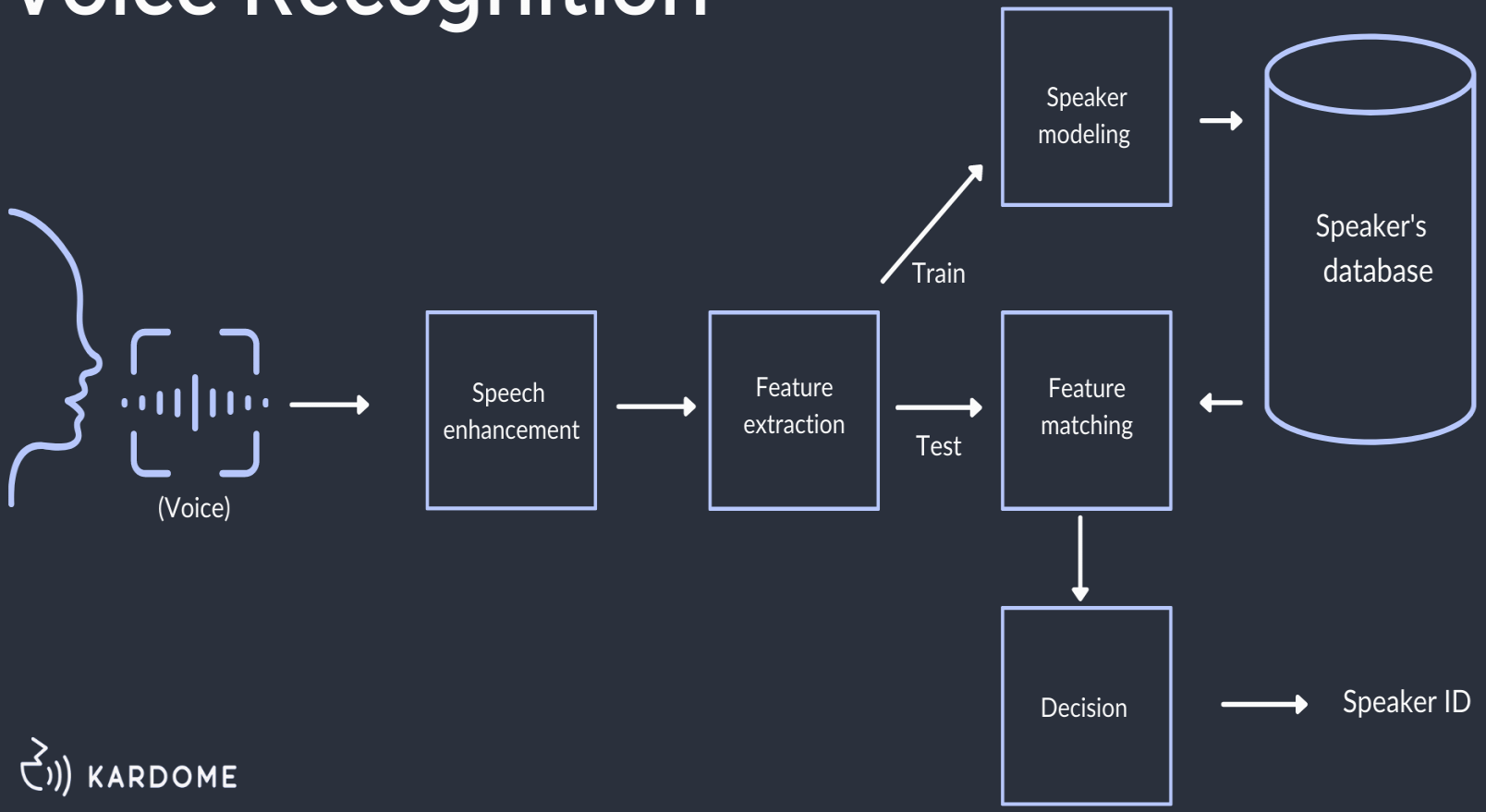
NEUTRAL

"Just booked two nights at this hotel."

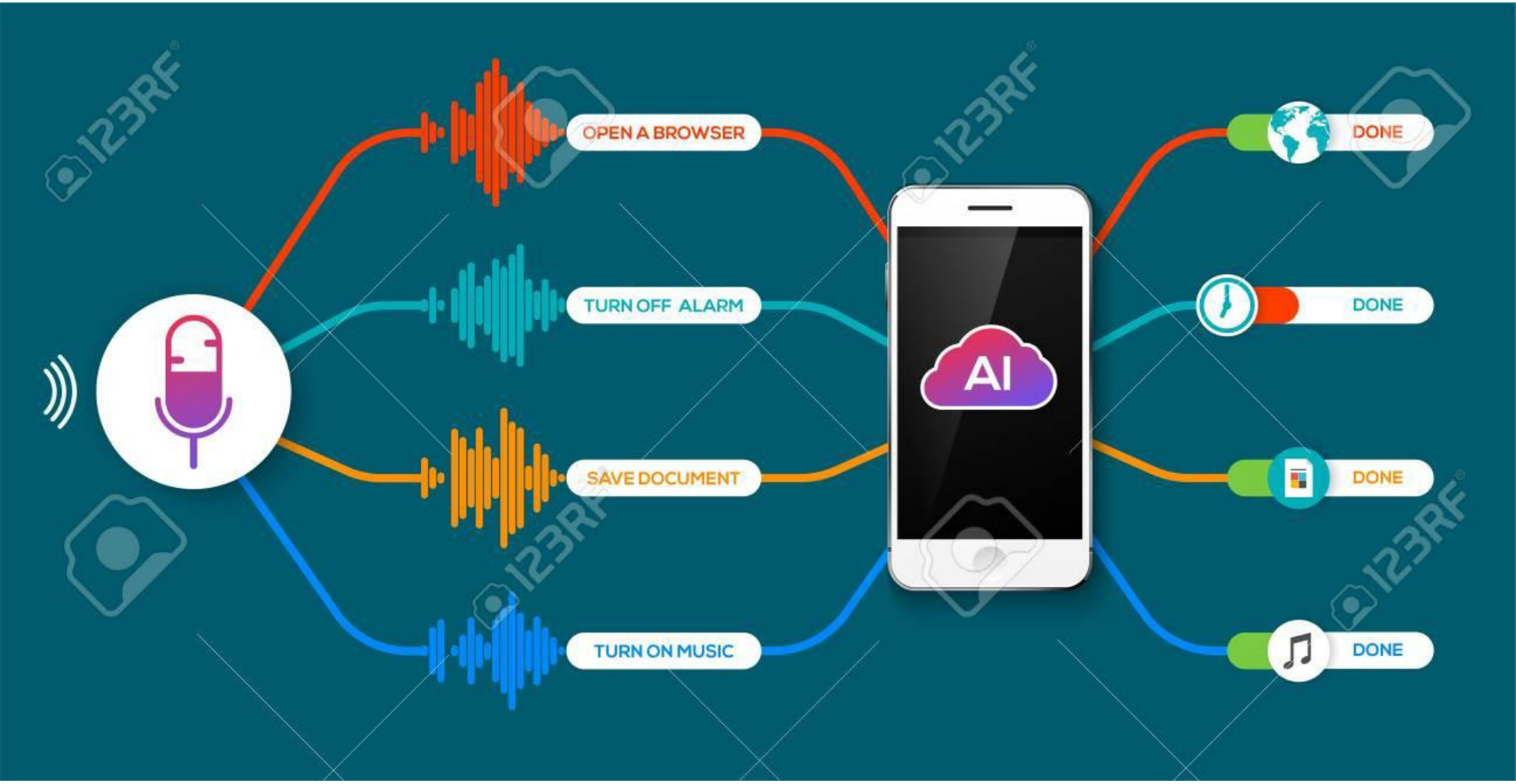
NEGATIVE

"Horrible services. The room was dirty and unpleasant.
Not worth the money."

Voice Recognition



KARDOME


OPEN A BROWSER
TURN OFF ALARM
SAVE DOCUMENT
TURN ON MUSIC
AI
DONE
DONE
DONE
DONE

Answering Questions

- ”What time is the next bus from the city after the 5:00 pm bus ?”
- ”I am a 3rd year CSE student, which classes do I have today ?”
- ”Which gene is associated with Diabetes ?”
- ”Who is Donald Knuth ?”

Information extraction

- Extraction of meaning from email :-

We have decided to meet tomorrow at 10:00am in the lab.

To do : meeting
Time : 10:00 am, 22/3/2012
Venue : Lab

Machine Translation

मेरा नाम रजत है | => My name is Rajat.

Grass is greener on the other side. => दूर के ढोल सुहावने |

Google's Translation : घास दूसरी तरफ हरियाली है |

Information extraction

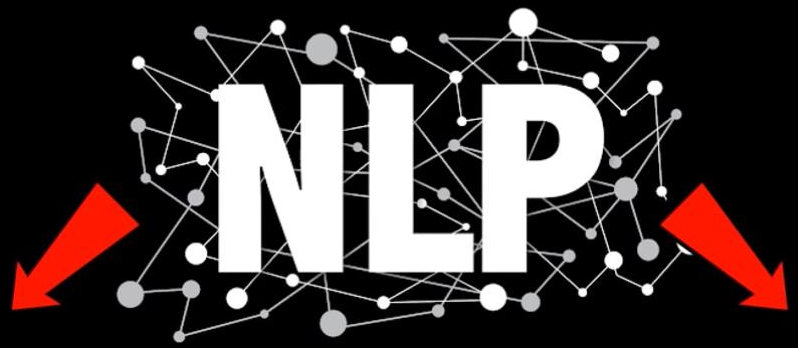
- Extraction of meaning from email :-

  We have decided to meet tomorrow at 10:00am in the lab.

What is natural language processing?

- Process information contained in natural language text
- Also known as Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)

Components of NLP



Natural Language Understanding

Natural Language Generation

SUBSCRIBE

Components of NLP

- Natural Language Understanding
  - Taking some spoken/typed sentence and working out what it means
- Natural Language Generation
  - Taking some formal representation of what you want to say and working out a way to express it in a natural (human) language (e.g., English)

Components of NLP (cont.)

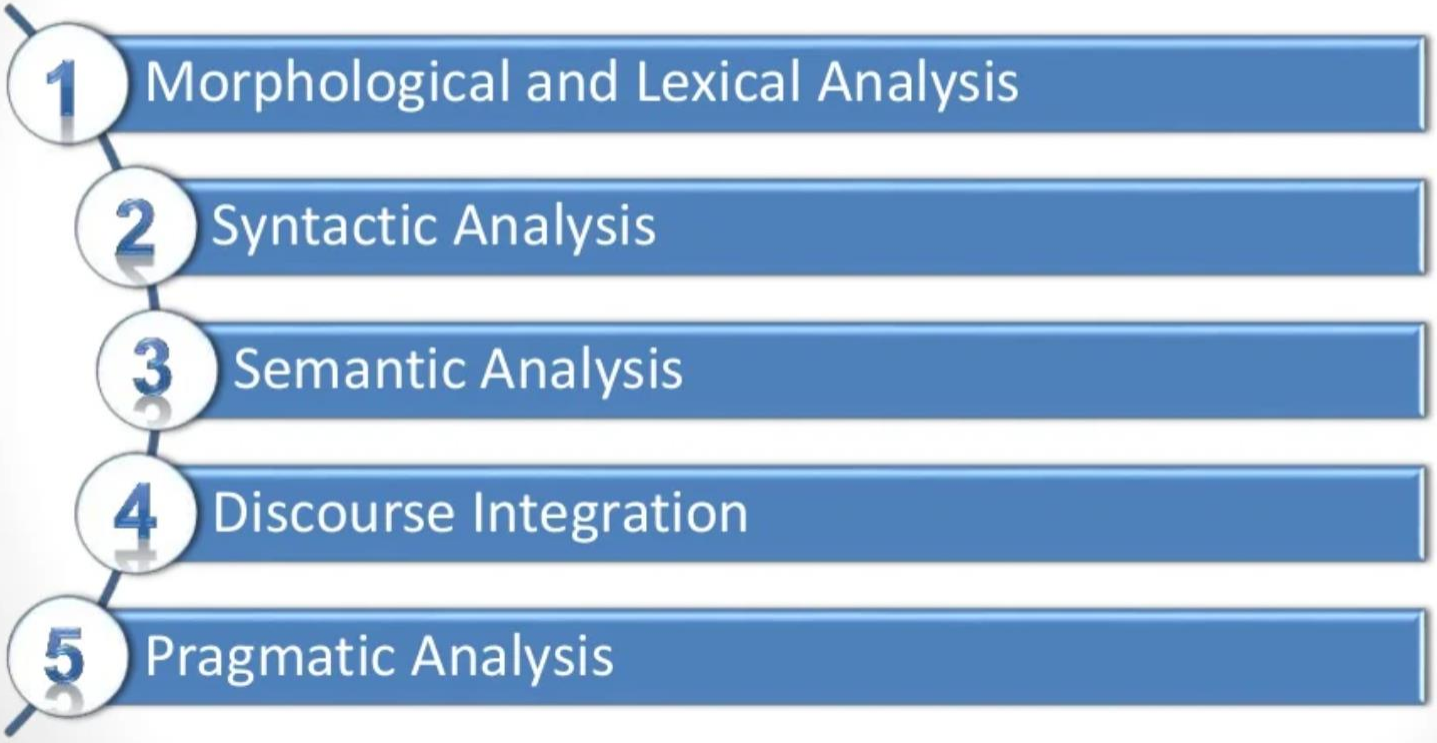
- Natural Language Understanding
  - Mapping the given input in the natural language into a useful representation
  - Different level of analysis required:
    - morphological analysis
    - syntactic analysis
    - semantic analysis
    - discourse analysis

Components of NLP (cont.)

- Natural Language Generation
  - Producing output in the natural language from some internal representation
  - Different level of synthesis required:
    - deep planning (what to say)
    - syntactic generation

- NL Understanding is much harder than NL Generation. But, still both of them are hard

Steps of NLP


1
Morphological and Lexical Analysis
2
Syntactic Analysis
3
Semantic Analysis
4
Discourse Integration
5
Pragmatic Analysis

Morphological and Lexical Analysis

- The lexicon of a language is its vocabulary that includes its words and expressions
- Morphology depicts analyzing, identifying and description of structure of words
- Lexical analysis involves dividing a text into paragraphs, words and the sentences

Syntactic Analysis

- Syntax concerns the proper ordering of words and its affect on meaning
- This involves analysis of the words in a sentence to depict the grammatical structure of the sentence
- The words are transformed into structure that shows how the words are related to each other
- Eg. “the girl the go to the school”. This would definitely be rejected by the English syntactic analyzer

Semantic Analysis

- Semantics concerns the (literal) meaning of words, phrases, and sentences
- This abstracts the dictionary meaning or the exact meaning from context
- The structures which are created by the syntactic analyzer are assigned meaning
- E.g.. “colorless blue idea” .This would be rejected by the analyzer as colorless blue do not make any sense together

Discourse Integration

- Sense of the context
- The meaning of any single sentence depends upon the sentences that precedes it and also invokes the meaning of the sentences that follow it
- E.g. the word “it” in the sentence “she wanted it” depends upon the prior discourse context

Pragmatic Analysis

- Pragmatics concerns the overall communicative and social context and its effect on interpretation
- It means abstracting or deriving the purposeful use of the language in situations
- Importantly those aspects of language which require world knowledge
- The main focus is on what was said is reinterpreted on what it actually means
- E.g. “close the window?” should have been interpreted as a request rather than an order

Tokenization

Stemming

Lemmatization

POS Tags

Named Entity Recognition

Chunking

Tasks in NLP

- Tokenization / Segmentation
- Disambiguation
- Stemming
- Part of Speech (POS) tagging
- Contextual Analysis
- Sentiment Analysis

Segmentation

- **Segmenting** text into words

"The meeting has been scheduled for this Saturday."

"He has agreed to co-operate with me."

"Indian Airlines introduces another flight on the New Delhi–Mumbai route."

"We are leaving for the U.S.A. on 26th May."

"Vineet is playing the role of Duke of Athens in A Midsummer Night's Dream in a theatre in New Delhi."

- **Named Entity Recognition**

Tokenization

Tokenization | is | the | first | step | in | NLP

Stemming

Normalize words into its base form or root form

Affectation | Affects | Affections | Affected | Affection | Affecting

Stemming

Normalize words into its base form or root form

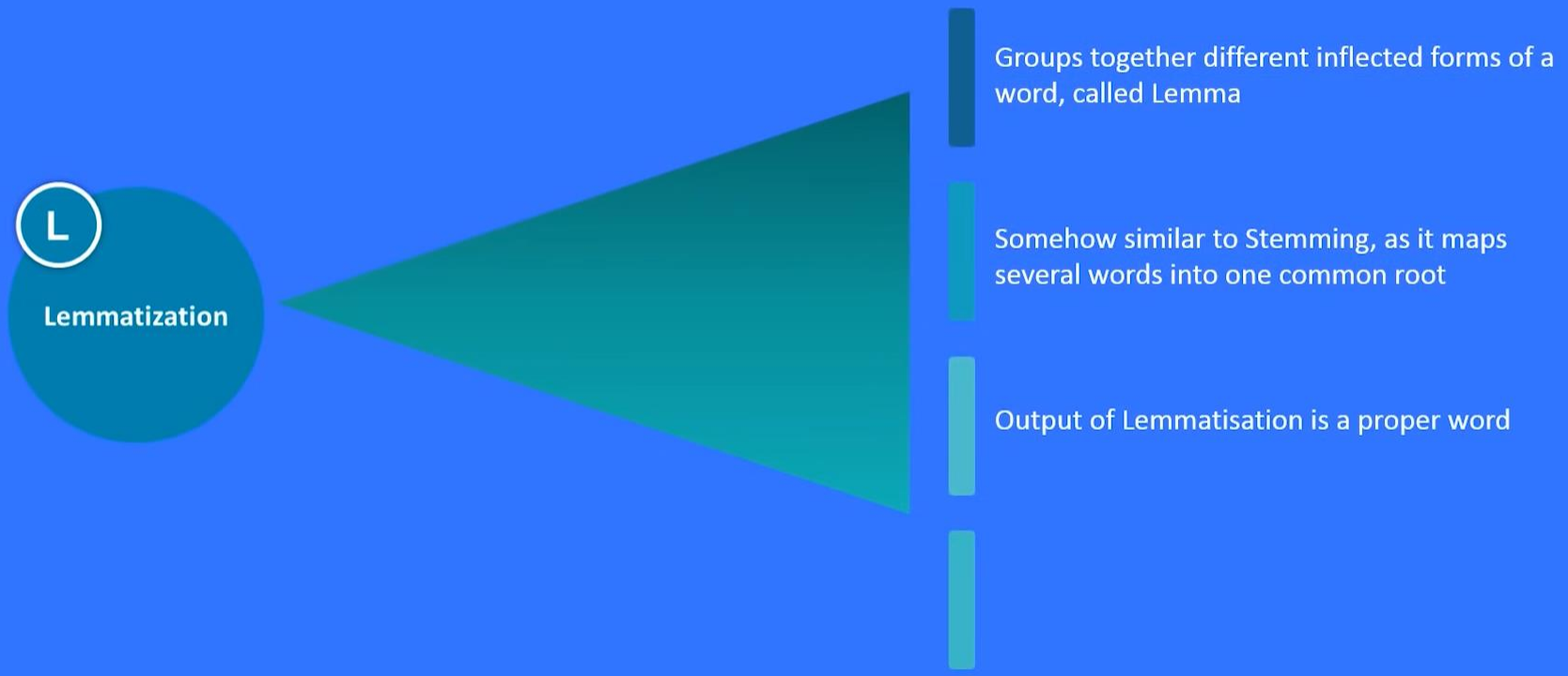
Affect

Stemming

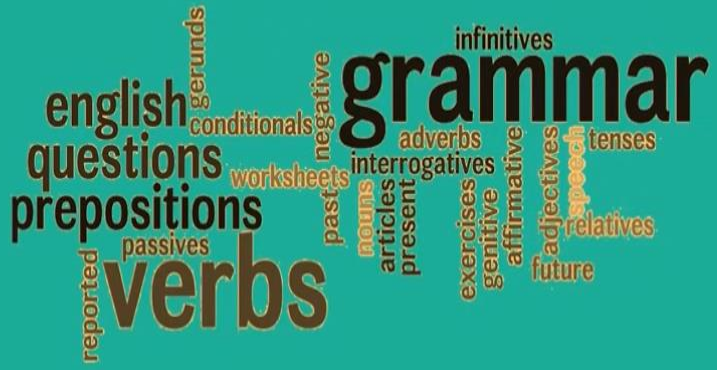
- Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form.

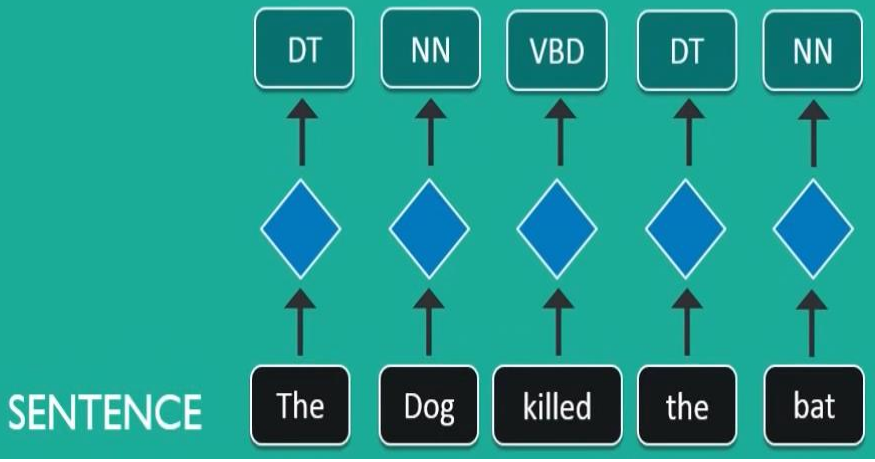
- car, cars -> car
- run, ran, running -> run
- stemmer, stemming, stemmed -> stem

Lemmatization



POS Tags





POS tagging

- **Part of speech (POS) recognition**

" Today is a beautiful day. "

Today	is	a	beautiful	day
Noun	Verb	Article	Adjective	Noun

POS tagging

- **Part of speech (POS) recognition**

" Today is a beautiful day. "

Today	is	a	beautiful	day
Noun	Verb	Article	Adjective	Noun

"Interest rates interest economists for the interest of the nation."

(word sense disambiguation)