Contents:

- *NLP Introduction*
- *Applications*
- *NLP Approaches*
- *NLP Components*
- *Phases in NLP*
- *Ambiguities*
- *NLP Preprocessing techniques*
- *POS Tagging*
- *Regular Expressions*

The Human Language

LANGUAGE

ALPHABETS

字母 वर्णमाला

ALFABETOS

الأبجدية எழுத்துக்கள்

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

The Human Language

LANGUAGE

ALPHABETS

字母 वर्णमाला ALFABETOS

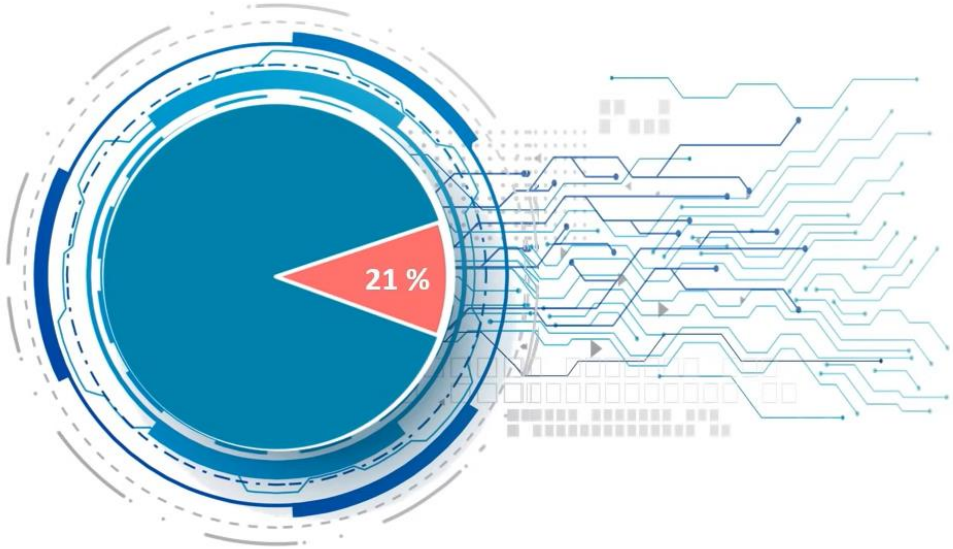
الأبجدية எழுத்துக்கள்

Words form Sentences

A language is not just words. It's a culture, a tradition, a unification of a community, a whole history that creates what a community is. It's all embodied in a language.

Noam Chomsky

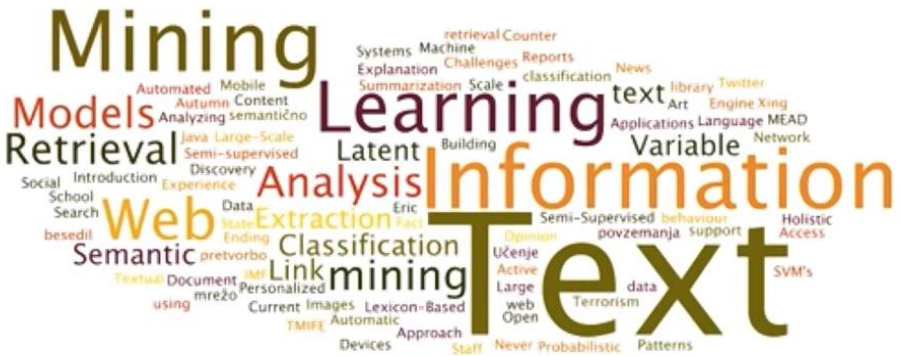
The 21st Century

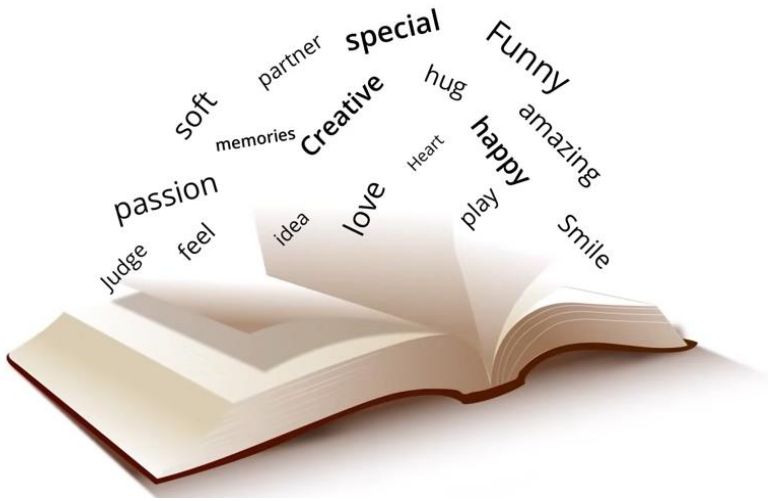


STRUCTURED DATA

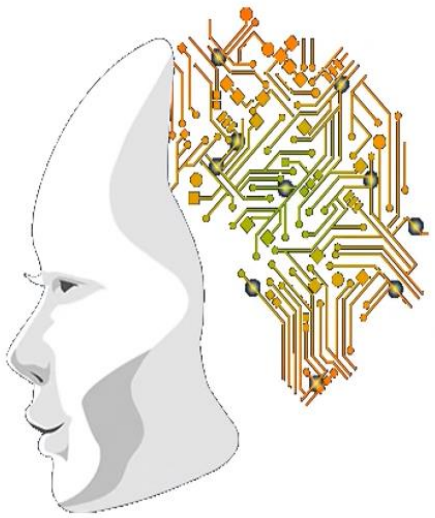
What is Text Mining ?

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





Text Mining and NLP

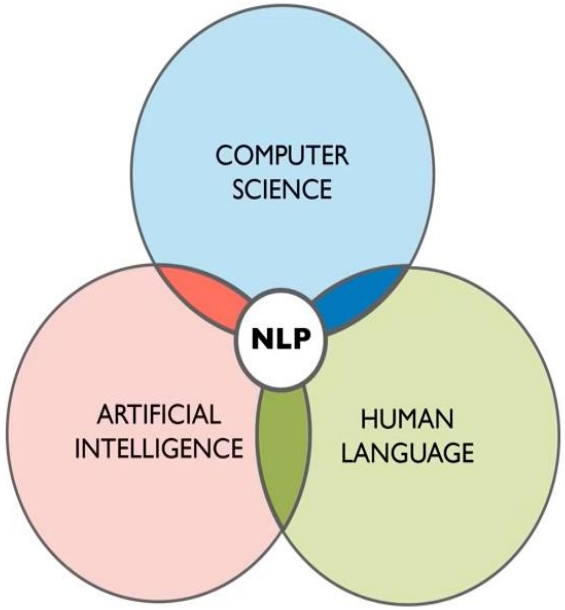


As, Text Mining refers to the process of deriving high quality information from the text .

The overall goal is, essentially to turn text into data for analysis, via application of Natural Language Processing (NLP)

What is NLP?

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



Applications of NLP

Sentimental Analysis

Chatbot

Speech Recognition

Machine Translation

Applications of NLP and Text Mining

Spell Checking

Information Extraction

Keyword Search

Advertisement Matching


Core Tasks
Text Classification
Information Extraction
Conversational Agent
Information Retrieval
Question Answering Systems
General Applications
Spam Classification
Calendar Event Extraction
Personal Assistants
WWW
Search Engines
JEOPARDY!
Jeopardy!
Industry Specific
Social Media Analysis
Retail Catalog Extraction
Health Records Analysis
Financial Analysis
Legal Entity Extraction

Main approaches in NLP

1. Rule-based Approach:

- This approach relies on predefined linguistic rules and patterns to process text.
- Linguistic experts and programmers manually create rules that encode knowledge about language.
- These rules are used to perform tasks such as tokenization, part-of-speech tagging, parsing, and information extraction.
- Rule-based systems are based on explicit, handcrafted rules and are effective for domains with well-defined rules.

2. Statistical Approach:

- Statistical NLP, also known as data-driven or machine learning-based NLP, utilizes statistical models and algorithms to learn patterns and structures from large amounts of annotated text data.
- These models use probabilistic techniques to make predictions about linguistic phenomena based on observed patterns in the training data.
- Statistical NLP techniques include machine translation, named entity recognition, sentiment analysis, and text classification.
- Statistical models require large amounts of annotated training data and can automatically extract relevant features from the data.

3. Neural Network Approach:

- Neural networks have revolutionized NLP in recent years.
- Deep learning models, such as recurrent neural networks (RNNs) and transformers, have shown remarkable performance in various NLP tasks.
- These models can learn hierarchical representations of text data and capture complex linguistic patterns.
- They excel in tasks such as language modeling, machine translation, sentiment analysis, and question answering.
- Neural network approaches require substantial computational resources and large amounts of annotated data for training.
- Distributional Semantics and Word Embeddings: Word embeddings, such as Word2Vec and GloVe, are vector representations of words that capture their semantic relationships and contextual information.

- **Pre-trained Language Models**: Pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have gained popularity in recent years.
- These models are trained on large-scale text data and capture rich linguistic representations.
- They can be fine-tuned for specific NLP tasks, requiring less task-specific data for training and achieving state-of-the-art performance in various tasks.

NLP : Components

Natural Language Understanding

Natural Language Generation

NLP : Components

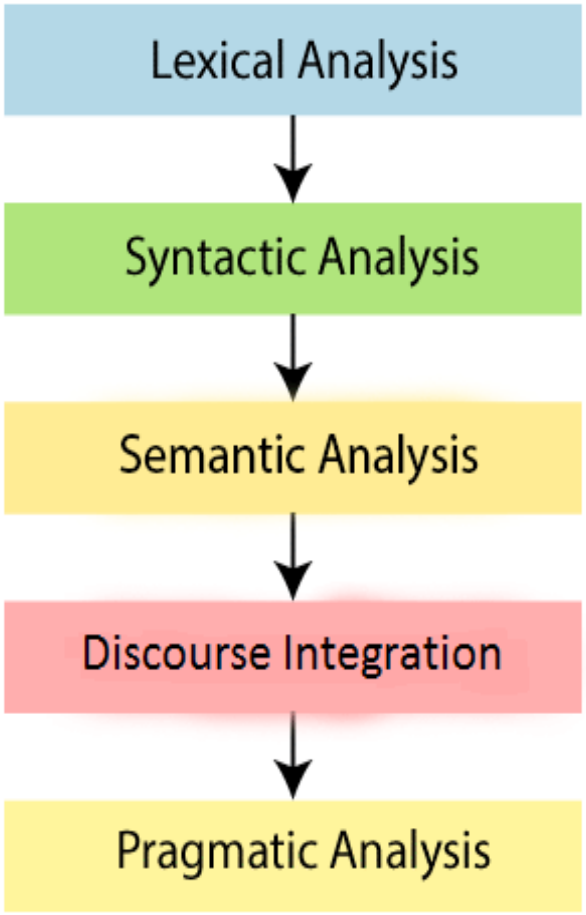
Natural Language Understanding

- Mapping input to useful representations
- Analyzing different aspects of the language

Natural Language Generation

- Text Planning
- Sentence Planning
- Text Realization

Phases of NLP



Morphological Analysis/Lexical Analysis

- This phase scans the source code as **a stream of characters** and converts it into meaningful **lexemes**.
- It divides the whole text into **paragraphs**, **sentences**, and **words**.
- For example, irrationally can be broken into ir (prefix), rational (root) and -ly (suffix).
- It also assigns the possible Part-Of-Speech (**POS**) to the word.
- For example, the word “character” can be used as a noun or a verb.

The Two relationships

Two different kind of relationship among words

Inflectional morphology

Grammatical: number, tense, case, gender

Creates new forms of the same word: *bring, brought, brings, bringing*

Derivational morphology

Creates new words by changing part-of-speech: *logic, logical, illogical, logically, logician, logicize.*

Morphological Analysis

Input	Morphological Parsed Output
cats	cat +N +PL
cat	cat +N +SG
cities	city +N +PL
geese	goose +N +PL
goose	(goose +N +SG) or (goose +V)
gooses	goose +V +3SG
merging	merge +V +PRES-PART
caught	(catch +V +PAST-PART) or (catch +V +PAST)

Goal

To take input forms like those in the first column and produce output forms like those in the second column.

Output contains stem and additional information; +N for noun, +SG for singular, +PL for plural, +V for verb etc.

Syntactic Analysis (Parsing)

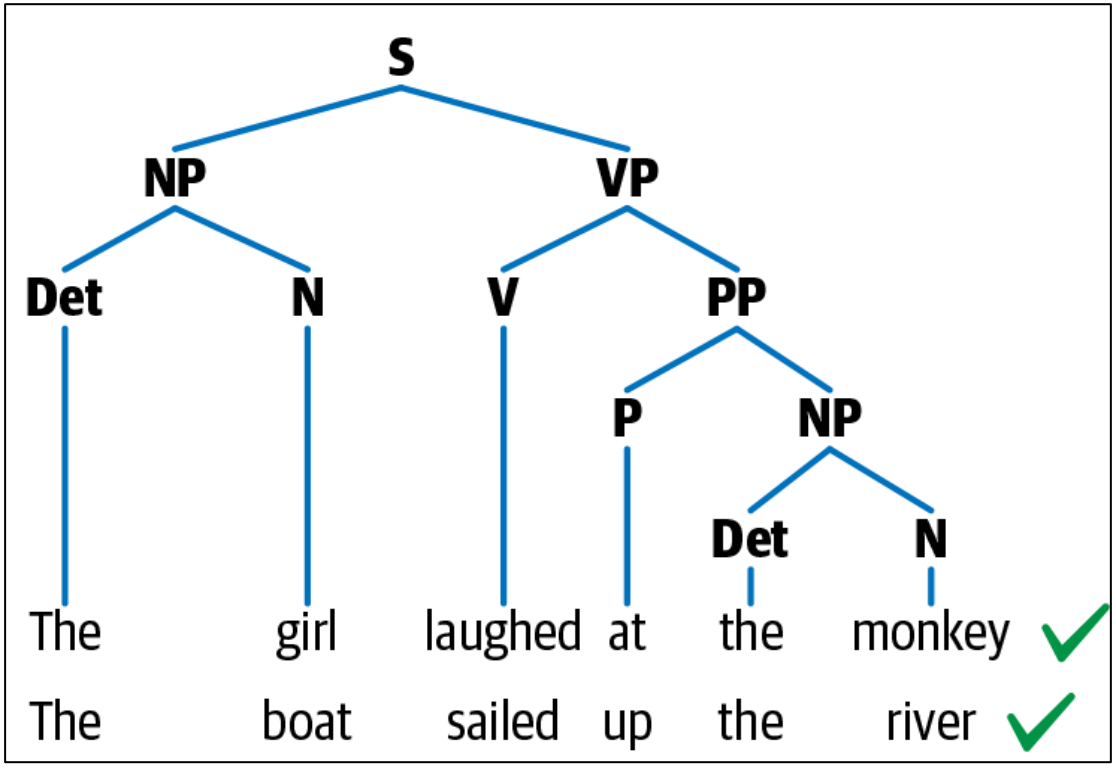
- Syntax is a set of rules to construct grammatically correct sentences out of words and phrases in a language.
- Syntax Analysis ensures that a given piece of text is correct structure.
- It tries to parse the sentence to check correct grammar at the sentence level.
- It also assigns the possible Part-Of-Speech (POS) to the word.
- For example:

  Correct Syntax: Sun rises in the east.

  Incorrect Syntax: Rise in sun the east.

Syntactic Analysis (Parsing)

- Syntactic structure in linguistics is represented in many different ways. A common approach to representing sentences is a parse tree.



Syntactic structure of two syntactically similar sentences

Semantic Analysis

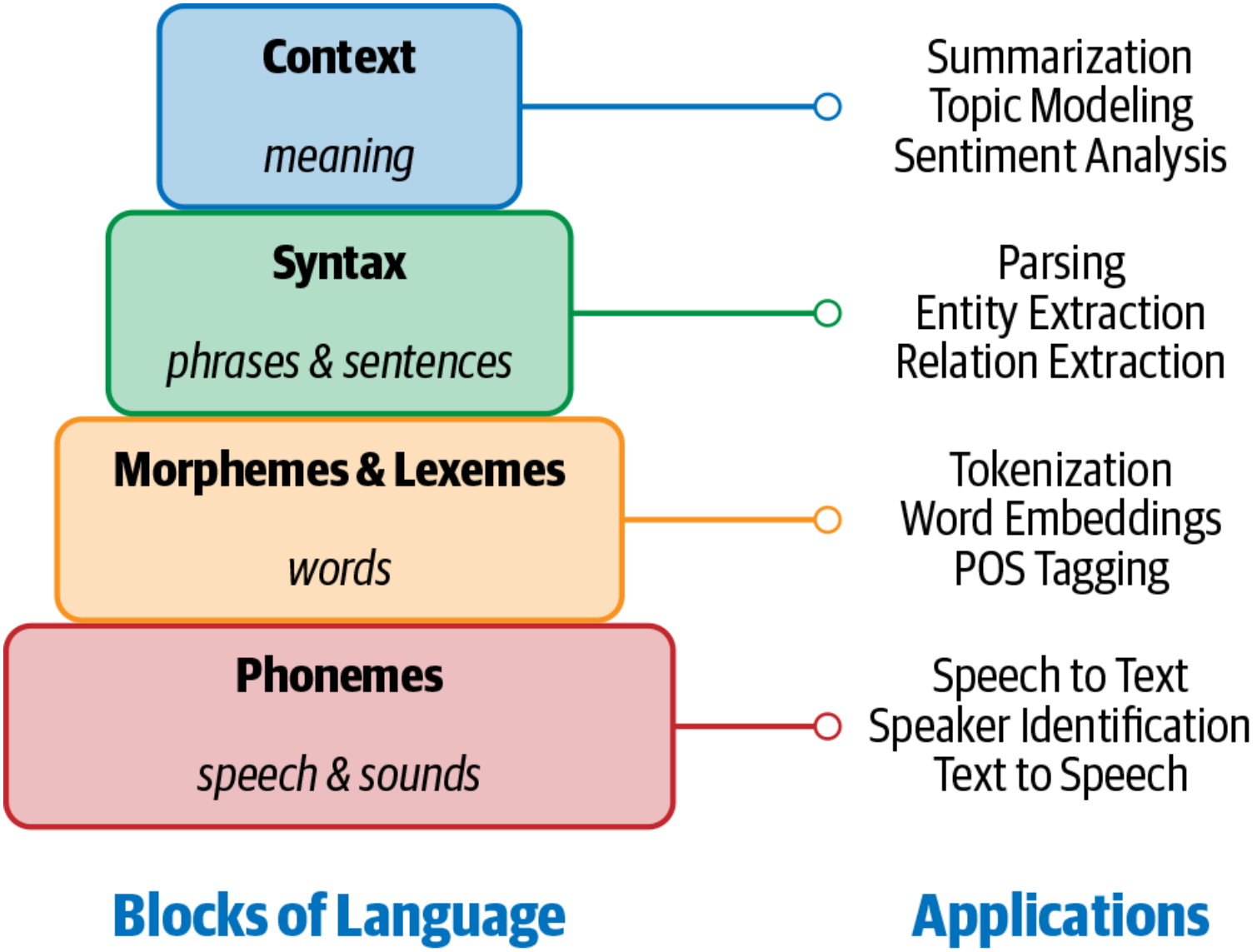
- Semantic analysis is concerned with the **meaning representation**.
- It mainly focuses on the literal meaning of words, phrases, and sentences.
- Consider the sentence: “The apple ate a banana”. Although the sentence is syntactically correct.
- It doesn’t make sense because apples can’t eat.
- Knowledge of wordnet

Discourse Integration

- Discourse deals with the effect of a previous sentence on the sentence in consideration.
- In the text, “Jack is a bright student. He spends most of the time in the library.”
- Here, discourse assigns “he” to refer to “Jack”.

Pragmatic Analysis

- Pragmatic is the fifth and last phase of NLP.
- It helps you to discover the intended effect by applying a set of rules that characterize cooperative dialogues.
- For Example: "Open the door" is interpreted as a request instead of an order.


Context
meaning
Syntax
phrases & sentences
Morphemes & Lexemes
words
Phonemes
speech & sounds
Summarization
Topic Modeling
Sentiment Analysis
Parsing
Entity Extraction
Relation Extraction
Tokenization
Word Embeddings
POS Tagging
Speech to Text
Speaker Identification
Text to Speech
Blocks of Language
Applications

NLP : Ambiguity

Natural Language Understanding

Ambiguity

- Lexical Ambiguity
- Syntactic Ambiguity
- Referential Ambiguity

NLU : Ambiguity

- **Lexical Ambiguity**
- Syntactic Ambiguity
- Referential Ambiguity

She is looking for a ***match***.

The fisherman went to the ***bank***.

NLU : Ambiguity

- Lexical Ambiguity
  - She is looking for a ***match***.
  - The fisherman went to the ***bank***.
- **Syntactic Ambiguity**
  - The chicken is ready to eat.
  - Visiting relatives can be boring.
  - I saw the man with the binoculars.
- Referential Ambiguity

NLU : Ambiguity

Lexical Ambiguity

She is looking for a ***match***.

The fisherman went to the ***bank***.

Syntactic Ambiguity

The chicken is ready to eat.

Visiting relatives can be boring.

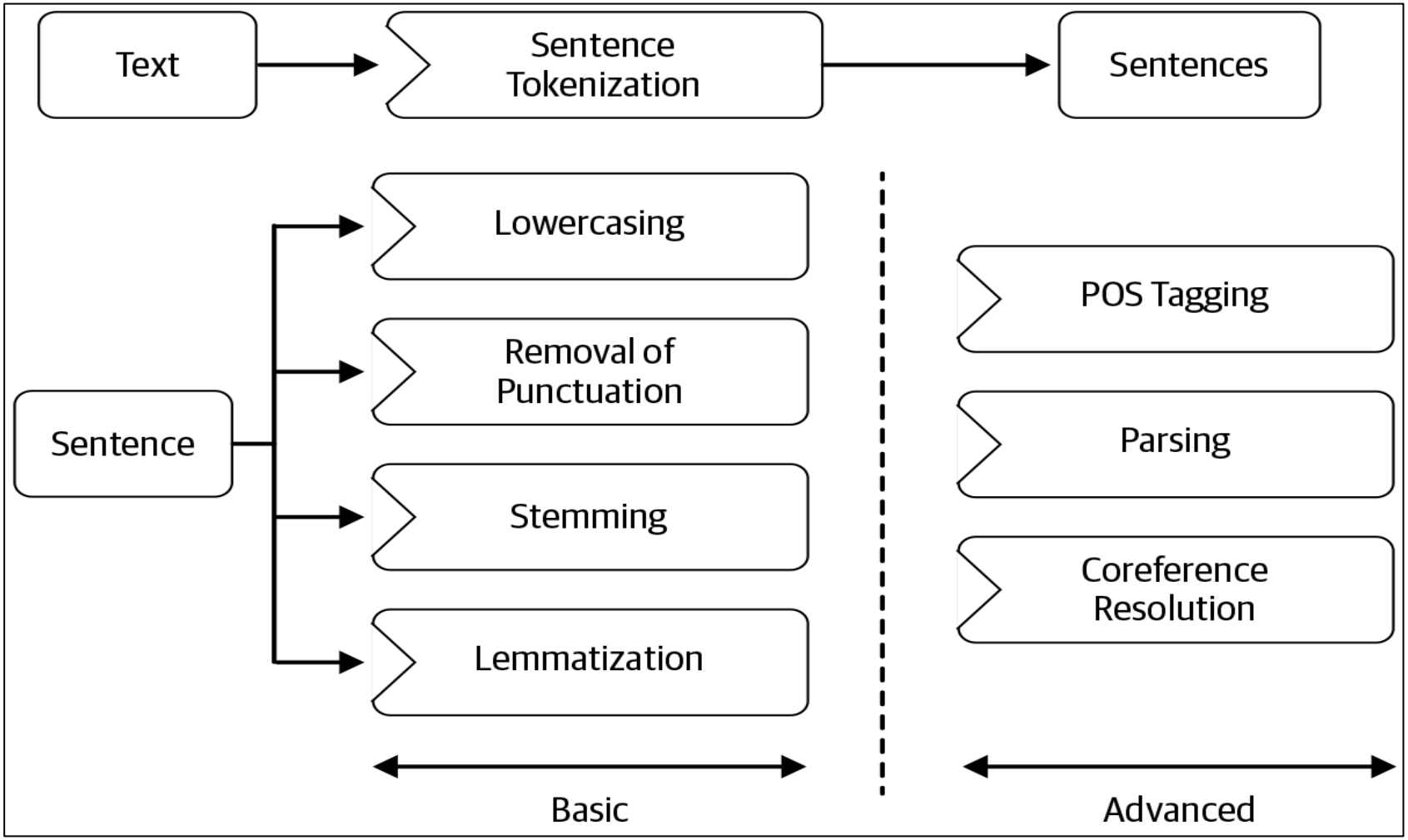
I saw the man with the binoculars.

Referential Ambiguity

The boy told his father the theft. ***He*** was very upset

NLP Pre-Processing Techniques

Pre-Processing Steps



Tokenization

Tokenization is the first step in NLP

01 Break a complex sentence into words

02 Understand the importance of each of the words with respect to the sentence

03 Produce a structural description on an input sentence

Tokenization

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

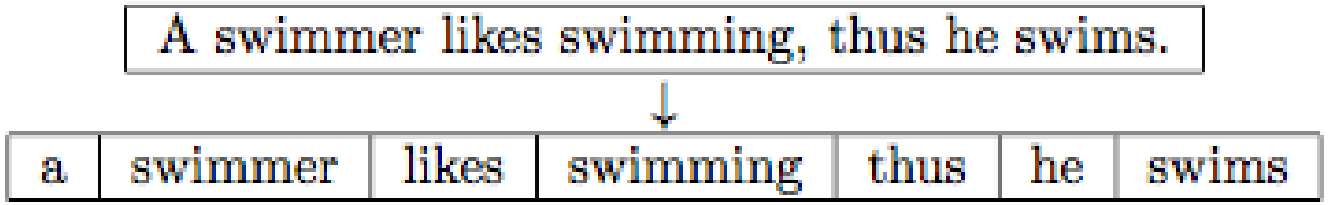
01 Break a complex sentence into words

02 Understand the importance of each of the words with respect to the sentence

03 Produce a structural description on an input sentence

Word Tokenization

Tokenization is the process of segmenting a string of characters into tokens (words).



An example

I have a can opener; but I can't open these cans.

- Word Tokens: 11
- Word Types: 10

Several tokenization libraries

- NLTK Toolkit (Python)
- Spacy (Python)
- Polyglot (Python)
- Stanford CoreNLP (Java)
- Unix Commands

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

Lemmatization example

WordNet lemmatizer

- Uses the WordNet Database to lookup lemmas
- nltk.stem.WordNetLemmatizer
- Examples:
  - feet → foot
  - wolves → wolf
  - cats → cat
  - talked → talked

- Problems: not all forms are reduced

- Takeaway: we need to try stemming or lemmatization and choose best for our task

Stop Words

Are they Helpful or Not ?

What are Named Entity Recognition ?

MOVIE

MONETARY VALUE

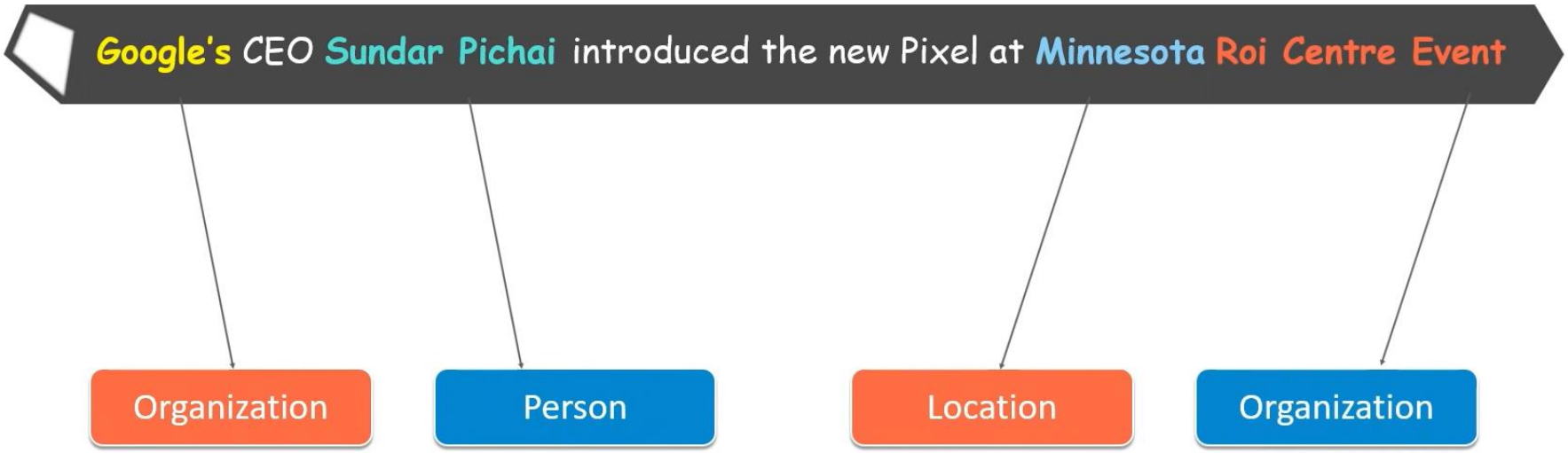
ORGANIZATION

LOCATION

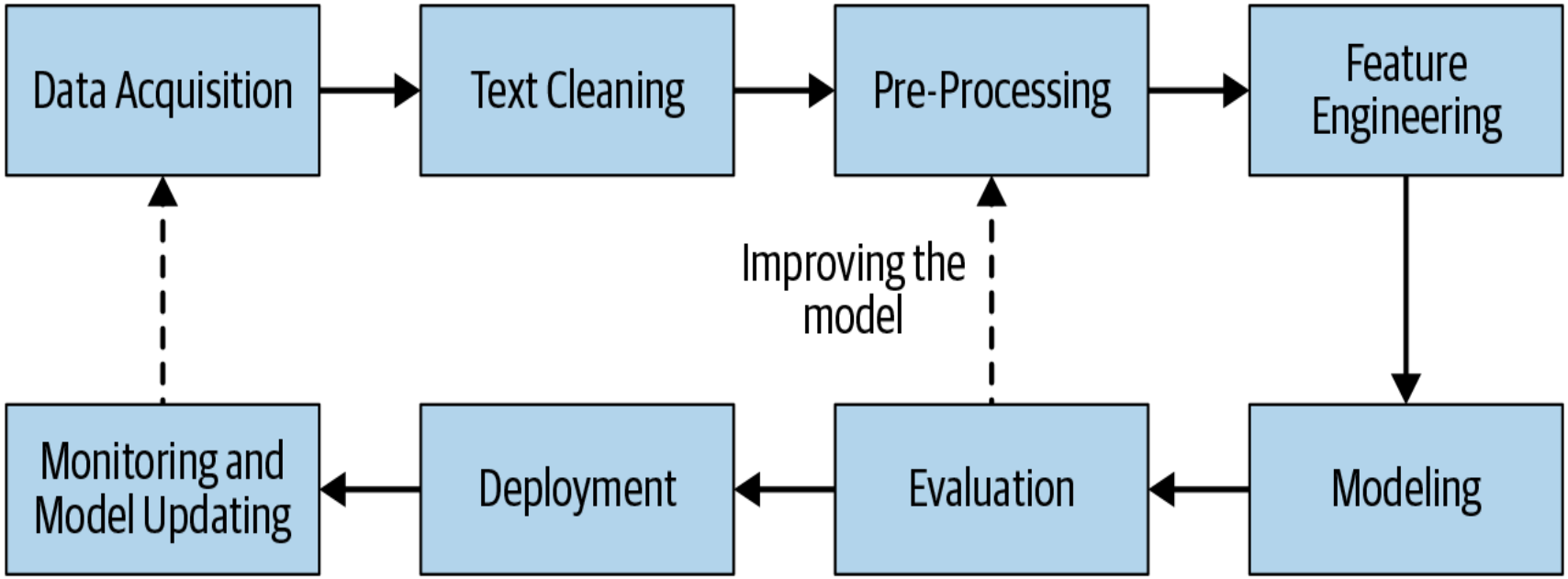
QUANTITIES

PERSON

NER : Named Entity Recognition



Generic NLP Pipeline



References:

- Speech and Language Processing (3rd ed. draft) by Dan Jurafsky and James H. Martin

  https://web.stanford.edu/~jurafsky/slp3/
- Practical Natural Language Processing (O'Reilly)

  A Comprehensive Guide to Building Real-World NLP Systems

  By Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana

Thank you