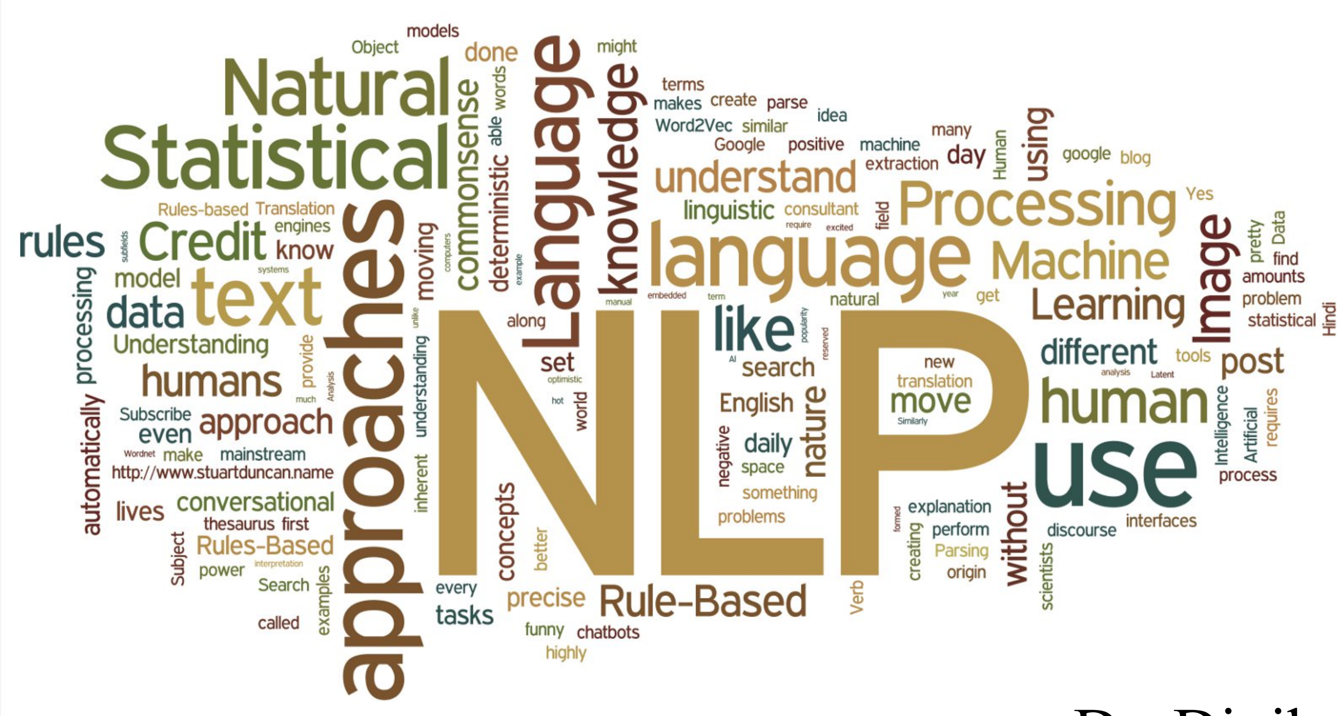
# CSET 346: Natural Language Processing



Dr. Dipika Jain
Assistant Professor
School of Computer Science
Engineering and Technology,
Bennett University
Email: Dipika.jain@bennett.edu.in

## Course Work and Grading Policy

Relevant MOOC Courses being Referred: Coursera

https://www.coursera.org/specializations/natural-language-processing

- Natural Language Processing with Classification and Vector Spaces
- Natural Language Processing with Probabilistic Models
- Natural Language Processing with Sequence Models
- Natural Language Processing with Attention Models.

# Syllabus

**Module 1:** Natural Language Processing: Need, applications, industry demand, Challenges in NLP: Ambiguity in language, Text-preprocessing methods, Stemming, Lemmatization, Tokenization, N-grams, Stops Words, WordNet, Language Corpus, N-gram Language Models, Hidden Markov Models

**Module 2: i**NLTK (Natural Language Toolkit for Indic Languages), Word representation, Sentence representation, Word embedding, Vector space model, Term Frequency, TF-IDF Representation, Distributional representation, Word2vec: CBOW, GloVe

**Module 3:** Neural Networks for text, Recurrent Neural Networks, Vanishing Gradients, exploding gradient, LSTM (Long sort term memory), GRU (Gated recurrent Unit), Seq2Seq Modelling, Bidirectional Model, Contextual Representations, Transformers, BERT, Transfer Learning in Word Embeddings, POS tagging, Named Entity Recognition, Sentiment Analysis

**Module 4:** Topic Modeling, Latent Semantic Analysis, Neural Machine Translation, Self-Attention for Generative Models, Natural Language Generation, Attention, Question Answering Bot, OpenAI’s GPT, Google’s ALBERT, ULMFiT, Facebook’s RoBERTa, Text Summarization, Extractive, Abstractive Text summarization, Transformer models for Text Summarization.

**Book References:** Manning, Christopher, and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

Jurafsky, Dan. *Speech & language processing*. Pearson Education India, 2000.

# High-Level Plan for Lab Assignments (to be completed individually!)

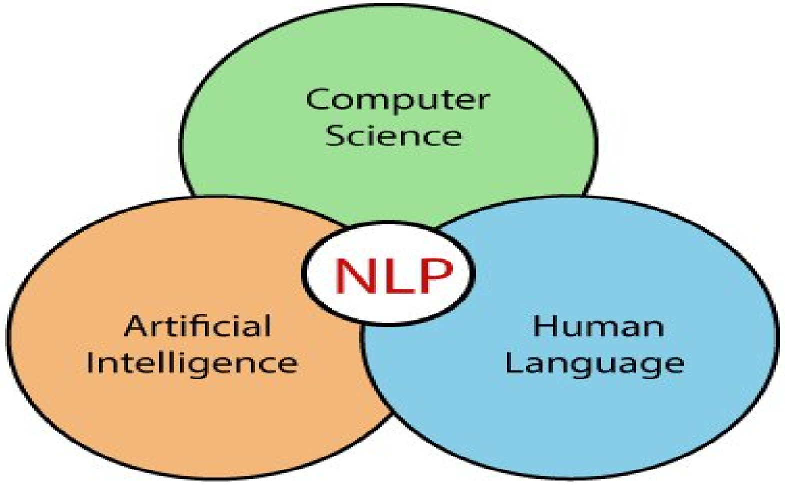
1. Assignments will be done using Python language.
2. Use package such such as nltk, spacy for text preprocessing.
3. Use pytorch/ tensorflow framework for implementation
4. For Final Project will be a research project. Rubrics of the project is as provided.

# What is NLP?

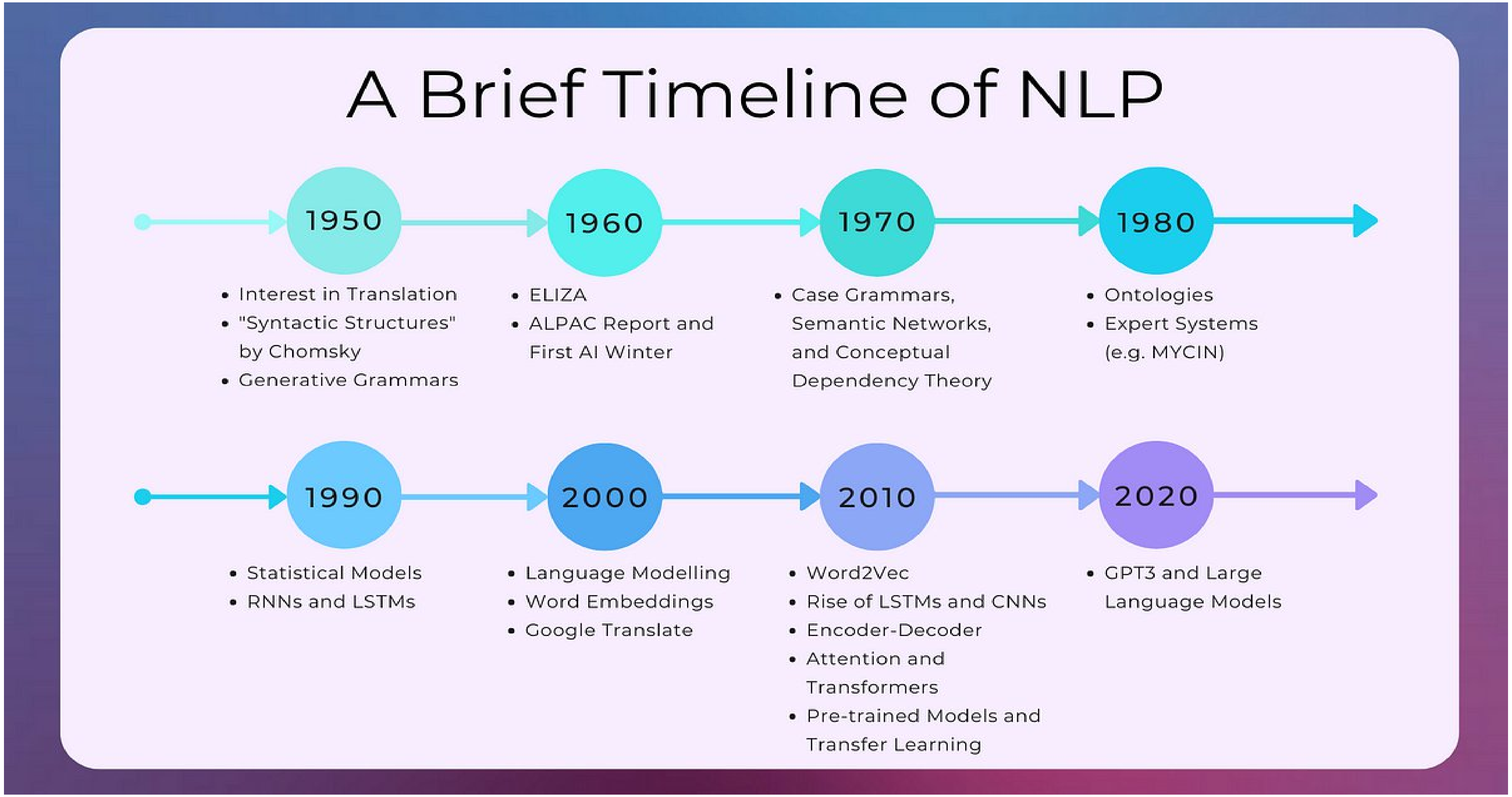
Wiki: Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

# What is NLP?

- NLP stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence.
- It is the technology that is used by machines to understand, analyze, manipulate, and interpret human's languages.



# A Brief History of NLP



A Brief Timeline of NLP

1950
- Interest in Translation
- "Syntactic Structures" by Chomsky
- Generative Grammars

1960
- ELIZA
- ALPAC Report and First AI Winter

1970
- Case Grammars, Semantic Networks, and Conceptual Dependency Theory

1980
- Ontologies
- Expert Systems (e.g. MYCIN)

1990
- Statistical Models
- RNNs and LSTMs

2000
- Language Modelling
- Word Embeddings
- Google Translate

2010
- Word2Vec
- Rise of LSTMs and CNNs
- Encoder-Decoder
- Attention and Transformers
- Pre-trained Models and Transfer Learning

2020
- GPT3 and Large Language Models

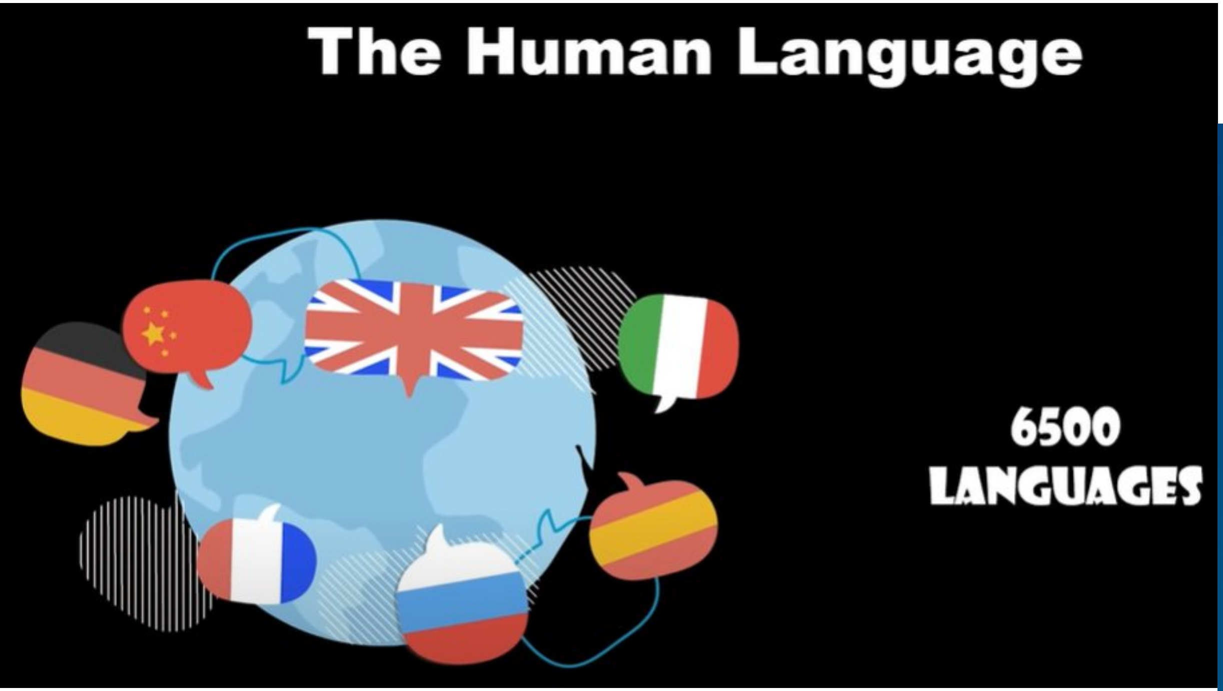
## Why study NLP ?

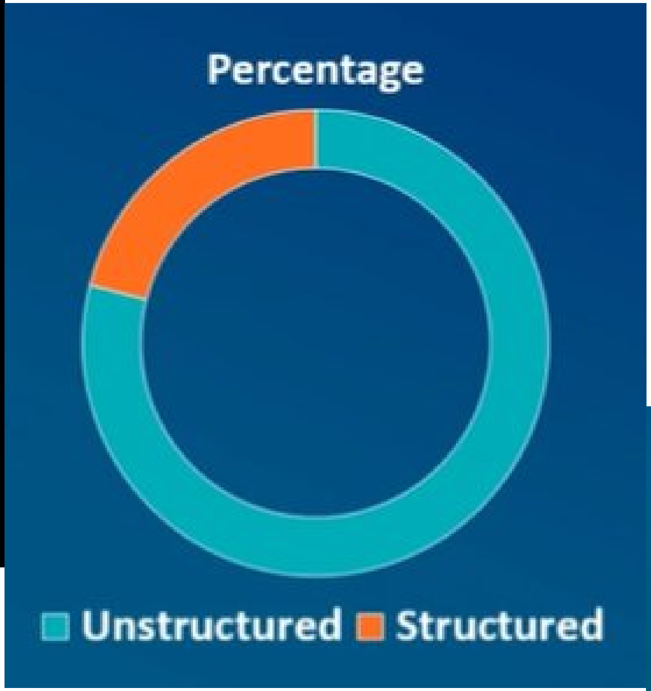
- Text is the largest repository of human knowledge (E.g. News articles, web pages, scientific documents) and is growing quickly
- Computer programs that understood text or speech
- To access (large amount of) information and knowledge stored in the form of human languages quickly.
- To interact with computing devices using human (natural) languages. For example, Building intelligent robots (AI), Enabling voice-controlled operation.
- We could not understand the majority of languages

  For e.g., Chinese people could not understand Spanish or Arabic or

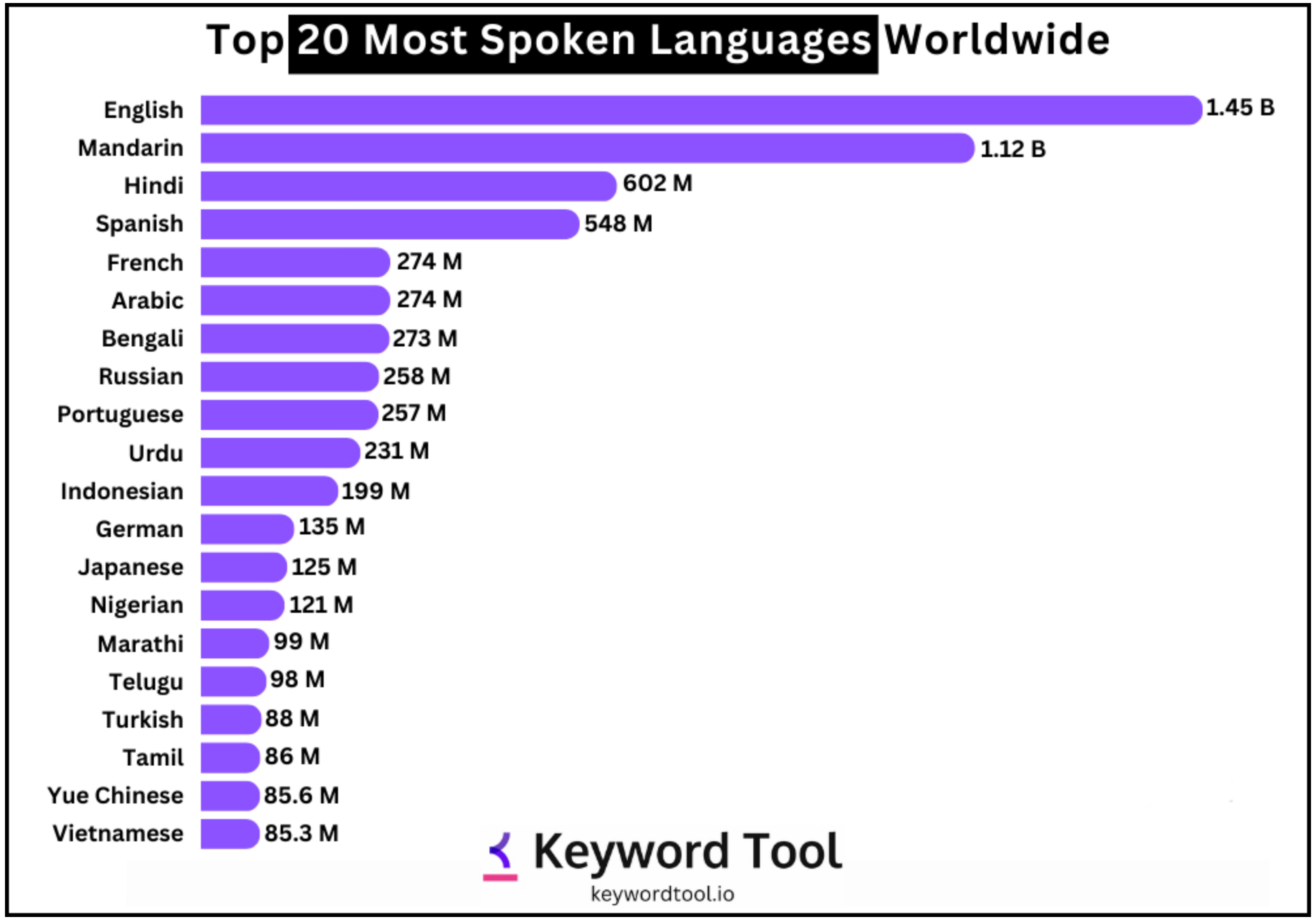
  Hindi

# Present Scenario with Human Language





# Worldwide Spoken languages



**Fundamental goal:**

- Deep understanding of broad language

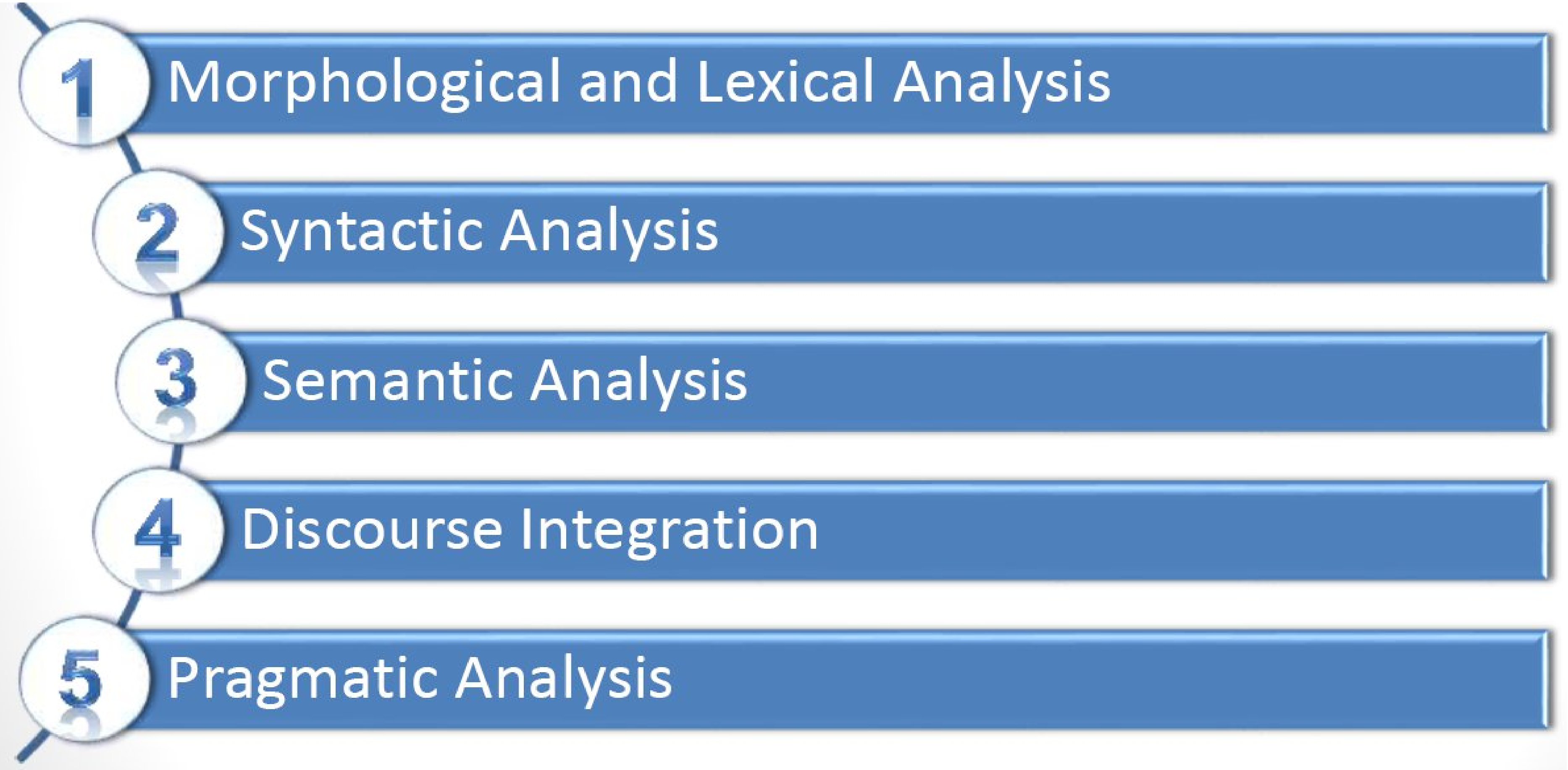
**Engineering goal:**

- Design, implement, and test systems that process natural languages for practical applications

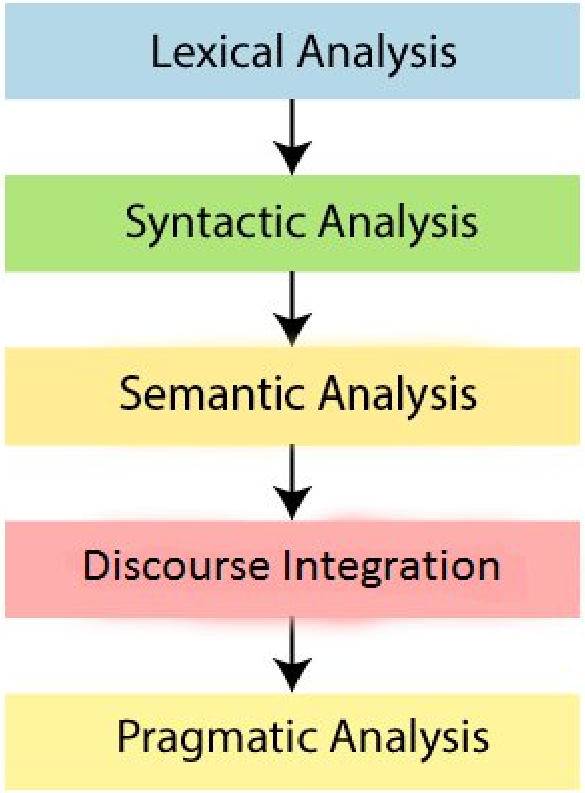
# NLP Tasks

- Word tokenization
- Sentence boundary detection
- Part-of-speech (POS) tagging
    - To identify the part-of-speech (e.g. noun, verb) of each word
- Named Entity (NE) recognition
    - To identify proper nouns (e.g. names of person, location, organization; domain terminologies)
- Parsing
    - To identify the syntactic structure of a sentence
- Semantic analysis
    - To derive the meaning of a sentence

# Phases of NLP



# Five phases of NLP



## Lexical Analysis

- The first phase of NLP is the 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.

## 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
- E.g. “The girl the go to the school”. This word definitely be rejected by the English syntactic analyzer
- E.g. “Ram is a good guy”. This word definitely be accepted by the English syntactic analyzer

## Semantic Analysis

- Semantic concern the (literal) meaning of words, phrases, and sentences
- This abstract the dictionary meaning or the exact meaning from context
- The structures which are created by the syntactic analyzer are assigned meaning
- For example, analyze the sentence “Ram is great.” In this sentence, the speaker is talking either about Lord Ram or about a person whose name is Ram. That is why the job, to get the proper meaning of the sentence, of semantic analyzer is important.
- E.g.. “colorless blue idea” .This would be rejected by the analyzer as colorless blue do not make any sense together

## Discourse Integration

- The meaning of any single sentence depends upon the sentences that precedes it and also invokes the meaning of the sentences that follow it
- Discourse integration takes into account the context of the text. It also considers the meaning of the following sentence. For example - "This is not true". The word "This" in this sentence depends upon the context of the word "This" in the previous sentences.
- E.g. the word “it” in the sentence “she wanted it” depends upon the prior discourse context

## Pragmatic Analysis

- It deals with overall communication and interpretation of language.
- 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?, Open the door"” should have been interpreted as a request rather than an order
- E.g. “I heart you!". Pragmatically, "heart" in this sentence means "love"- hearts are commonly used as a symbol for love, and to "heart" someone has come to mean that you love someone

# Why NLP is difficult?

- NLP is difficult because ambiguity and uncertainty exist in the language.

## Ambiguity

- There are following ambiguities exist in the language -

### 1. Lexical Ambiguity

- Lexical Ambiguity exists in the presence of two or more possible meanings of the sentence within a single word.

**Example:**

- **Manya is looking for a match.**
- In the above example, the word match refers to that either Manya is looking for a partner or Manya is looking for a match. (Cricket or other match)

# Ambiguity

## 2. Syntactic Ambiguity

Syntactic Ambiguity exists in the presence of two or more possible meanings within the sentence.

**Example:**

**I saw the girl with the binocular.**

- In the above example, did I have the binoculars? Or did the girl have the binoculars?

# Ambiguity

## 3. Referential Ambiguity

➢Referential Ambiguity exists when you are referring to something using the pronoun.

**Example:**

**Kiran went to Sunita. She said, "I am hungry."**

- In the above sentence, you do not know that who is hungry, either Kiran or Sunita.

# NLP APIs

- Natural Language Processing APIs allow developers to integrate human-to-machine communications and complete several useful tasks such as speech recognition, chatbots, spelling correction, sentiment analysis, etc.

**A list of NLP APIs is given below:**

- IBM Watson
- Chatbot
- Speech to text
- Sentiment Analysis
- Translation API by SYSTRAN
- Text Analysis API by AYLIEN
- Cloud NLP
- Google Cloud Natural Language

# NLP Libraries

- **Scikit-learn:** It provides a wide range of algorithms for building machine learning models in Python.
- **Natural language Toolkit (NLTK):** NLTK is a complete toolkit for all NLP techniques.
- **Pattern**: It is a web mining module for NLP and machine learning.
- **TextBlob**: It provides an easy interface to learn basic NLP tasks like sentiment analysis, noun phrase extraction, or pos-tagging.
- **Quepy:** Quepy is used to transform natural language questions into queries in a database query language.
- **SpaCy**: SpaCy is an open-source NLP library which is used for Data Extraction, Data Analysis, Sentiment Analysis, and Text Summarization.
- **Gensim:** Gensim works with large datasets and processes data streams.

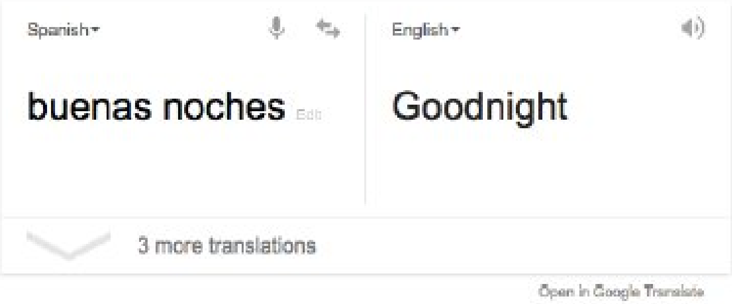
# Applications of NLP

1. Machine Translation
2. Information Retrieval
3. Question Answering
4. Chatbot
5. Information Extraction
6. Summarization
7. Sentiment Analysis
8. Auto Completion
9. Spam Detection
10. ……and many more

# Use Case - Translation

buenas noches

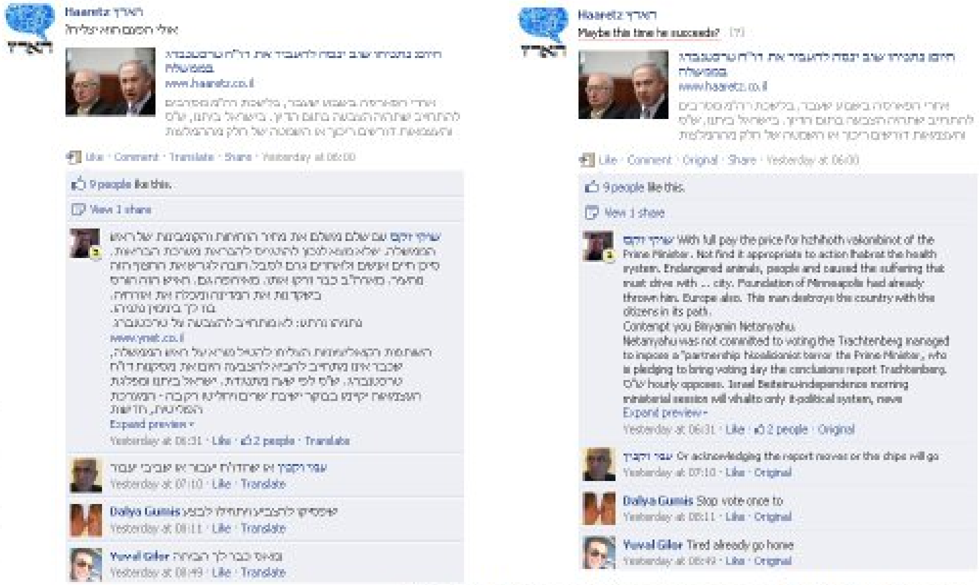
About 20,800,000 results (0.54 seconds)



buenas noches

Goodnight

3 more translations



Haaretz הארץ

אולי הפעם הוא יצליח?

[illegible]

www.haaretz.co.il

[illegible]

Like · Comment · Translate · Share · Yesterday at 06:00

9 people like this.

View 1 share

[illegible]

www.ynet.co.il

[illegible]

Expand preview

Yesterday at 06:31 · Like · 2 people · Translate

[illegible]

Yesterday at 07:10 · Like · Translate

Dalya Gumis [illegible]

Yesterday at 08:11 · Like · Translate

Yuval Giler [illegible]

Yesterday at 08:49 · Like · Translate

Haaretz הארץ

Maybe this time he succeeds? [?]

[illegible]

www.haaretz.co.il

[illegible]

Like · Comment · Original · Share · Yesterday at 06:00

9 people like this.

View 1 share

[illegible] With full pay the price for hzhhoth vakonbinot of the Prime Minister. Not find it appropriate to action habrat the health system. Endangered animals, people and caused the suffering that must drive with ... city. Foundation of Minneapolis had already thrown him. Europe also. This man destroys the country with the citizens in its path.
Contempt you Binyamin Netanyahu.
Netanyahu was not committed to voting the Trachtenberg managed to impose a "partnership hicoalisionist terror the Prime Minister, who is pledging to bring voting day the conclusions report Trachtenberg. O"W hourly opposes. Israel Beiteinu-independence morning ministerial session will vhalto only it-political system, neve
Expand preview

Yesterday at 06:31 · Like · 2 people · Original

[illegible] Or acknowledging the report moves or the chips will go

Yesterday at 07:10 · Like · Original

Dalya Gumis Stop vote once to

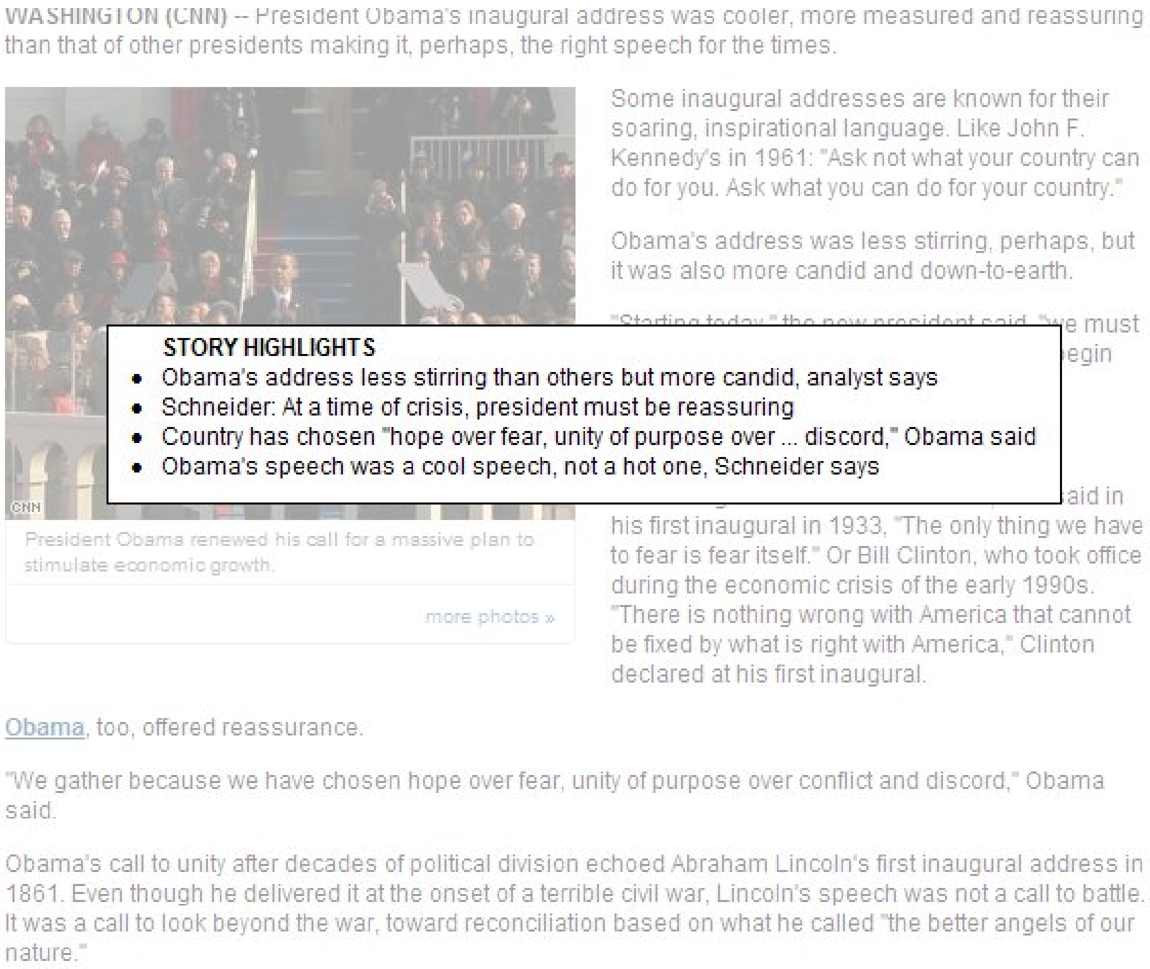
Yesterday at 08:11 · Like · Original

Yuval Giler Tired already go home

Yesterday at 08:49 · Like · Original

Facebook translation, image credit: Meedan.org

# Use Case- Summarization



**WASHINGTON (CNN)** -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.

President Obama renewed his call for a massive plan to stimulate economic growth.

more photos »

Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must

egin

aid in his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

**STORY HIGHLIGHTS**

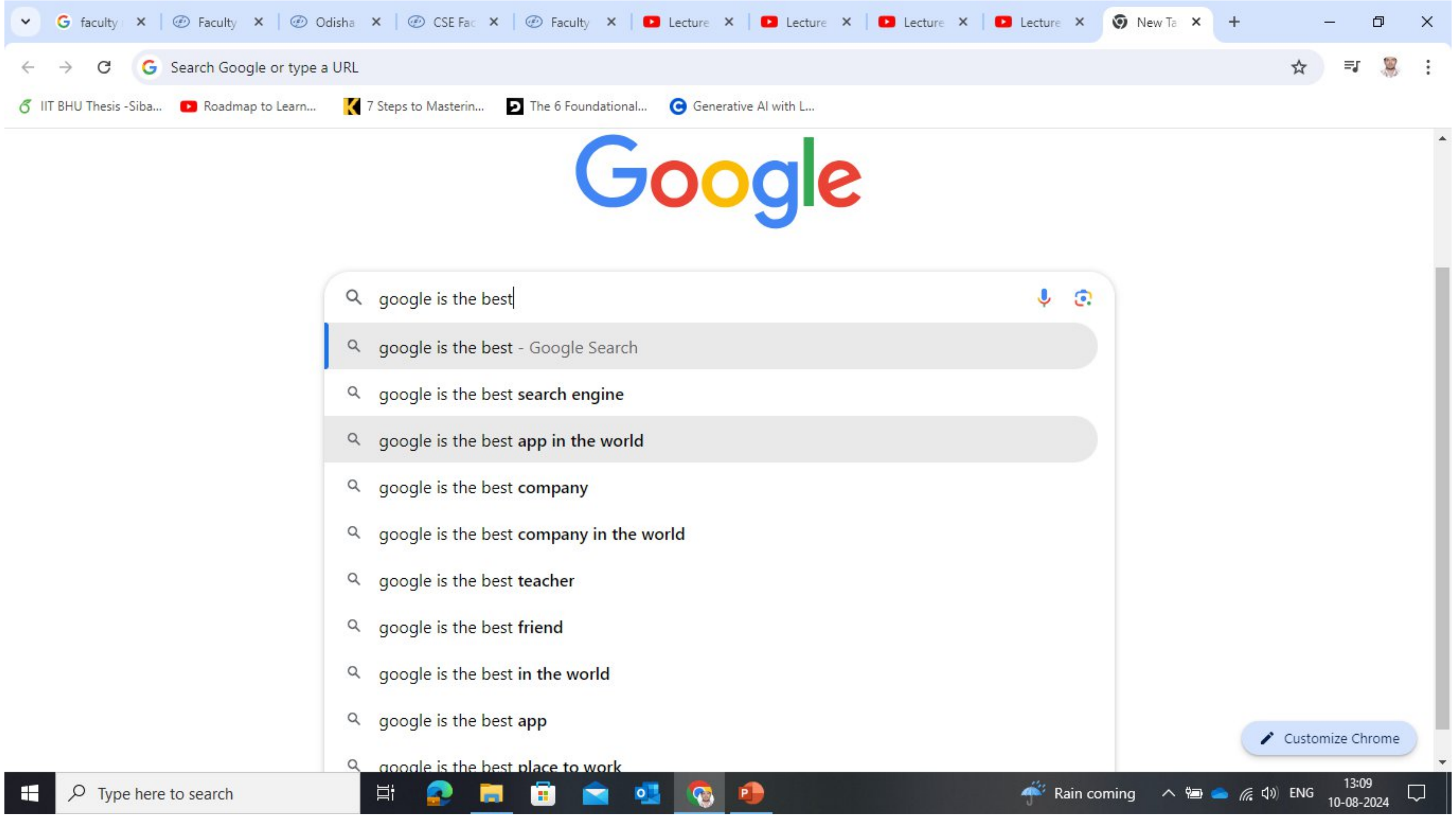
- Obama's address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen "hope over fear, unity of purpose over ... discord," Obama said
- Obama's speech was a cool speech, not a hot one, Schneider says

Obama, too, offered reassurance.

"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our nature."

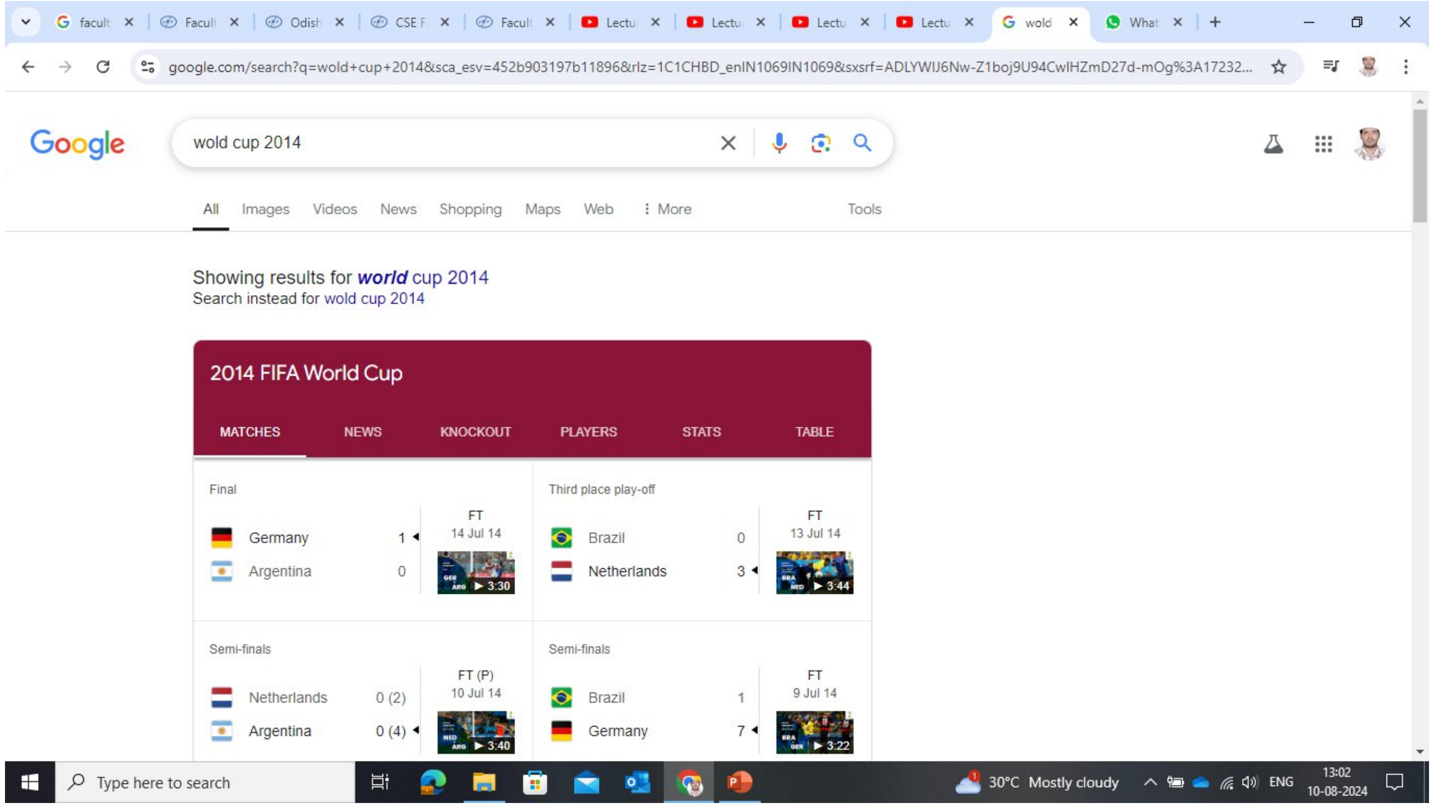
# Use Case – Auto Completion



google is the best

- google is the best - Google Search
- google is the best **search engine**
- google is the best **app in the world**
- google is the best **company**
- google is the best **company in the world**
- google is the best **teacher**
- google is the best **friend**
- google is the best **in the world**
- google is the best **app**
- google is the best **place to work**

# Use Case – Suggestion



google.com/search?q=wold+cup+2014&sca_esv=452b903197b11896&rlz=1C1CHBD_enIN1069IN1069&sxsrf=ADLYWIJ6Nw-Z1boj9U94CwIHZmD27d-mOg%3A17232...
Google
wold cup 2014
All
Images
Videos
News
Shopping
Maps
Web
More
Tools

Showing results for ***world*** cup 2014
Search instead for wold cup 2014

2014 FIFA World Cup

Final
Germany 1
Argentina 0
FT
14 Jul 14
3:30

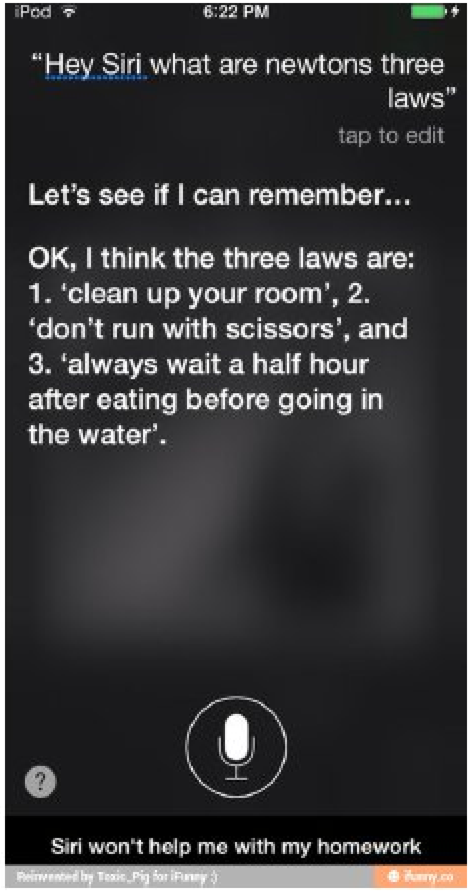
Third place play-off
Brazil 0
Netherlands 3
FT
13 Jul 14
3:44

Semi-finals
Netherlands 0 (2)
Argentina 0 (4)
FT (P)
10 Jul 14
3:40

Semi-finals
Brazil 1
Germany 7
FT
9 Jul 14
3:22

# Use Case – Question Answering

'Watson' computer wins at 'Jeopardy'



"Hey Siri what are newtons three laws"

Let's see if I can remember...

OK, I think the three laws are: 1. 'clean up your room', 2. 'don't run with scissors', and 3. 'always wait a half hour after eating before going in the water'.

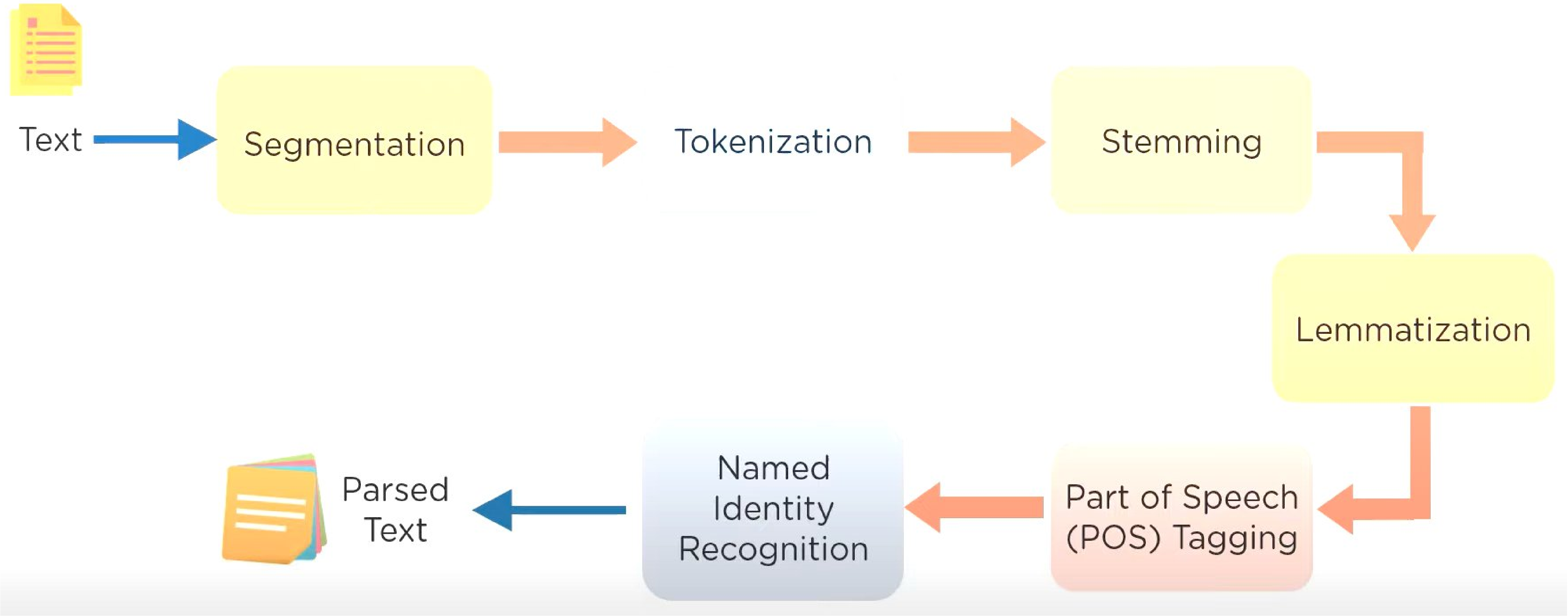
Siri won't help me with my homework

# Text Preprocessing Steps:

1. Regular Expression
2. Tokenization
3. Lemmatization
4. Stemming

*We will understand these steps in next class in a detailed manner.*

# NLP Pipeline



Text

Segmentation

Tokenization

Stemming

Lemmatization

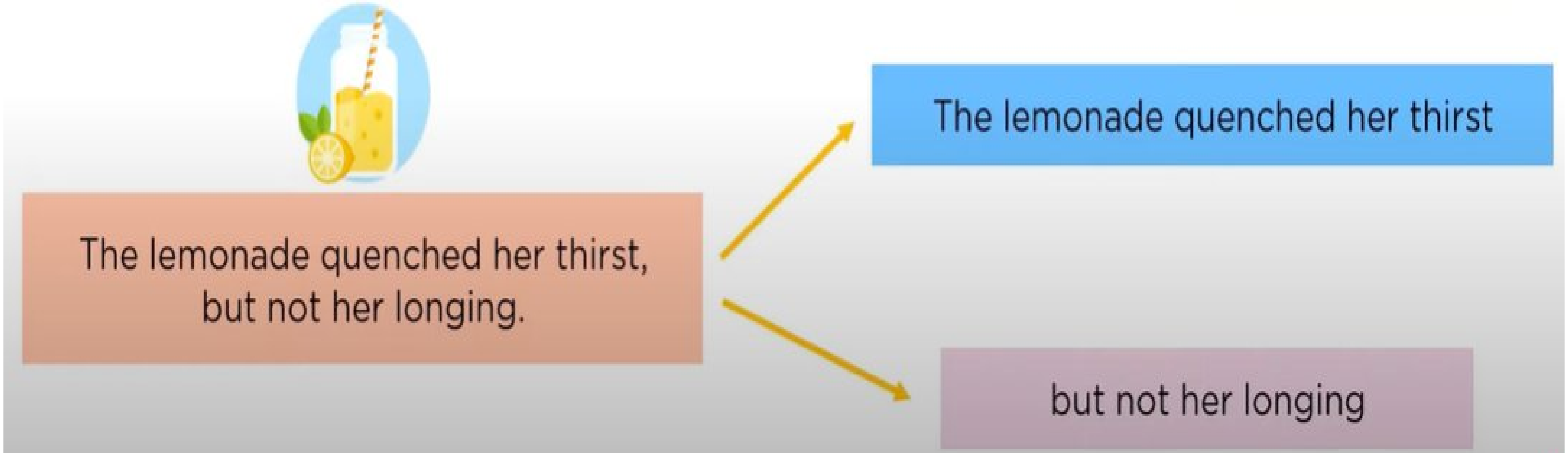
Part of Speech (POS) Tagging

Named Identity Recognition

Parsed Text

# NLP Pipeline – Segmentation

- Segmentation: The process of dividing a sentence into its component sentences, usually along punctuation marks.



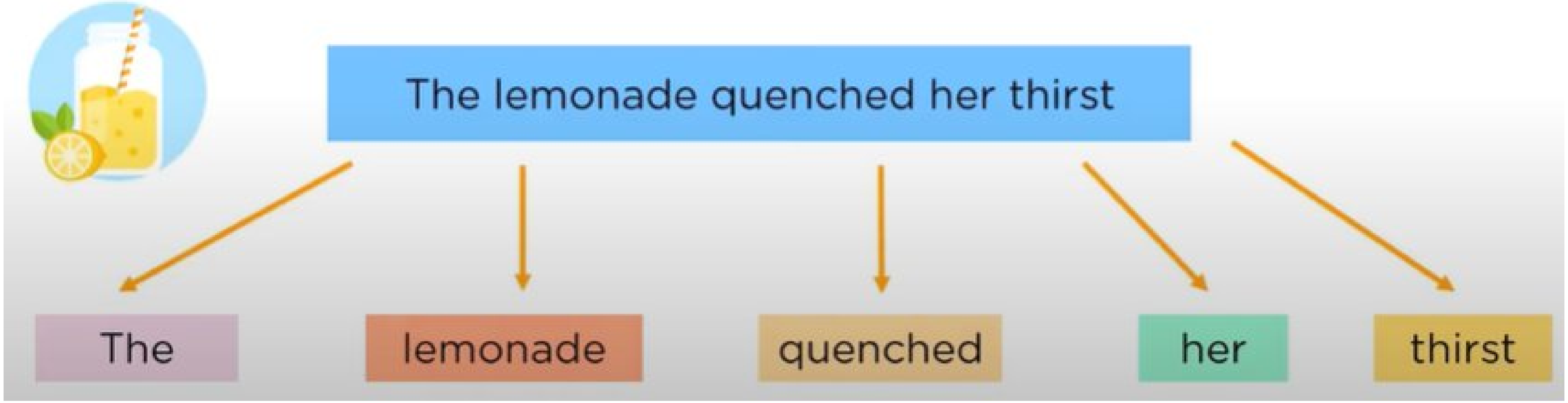
The lemonade quenched her thirst, but not her longing.

The lemonade quenched her thirst

but not her longing

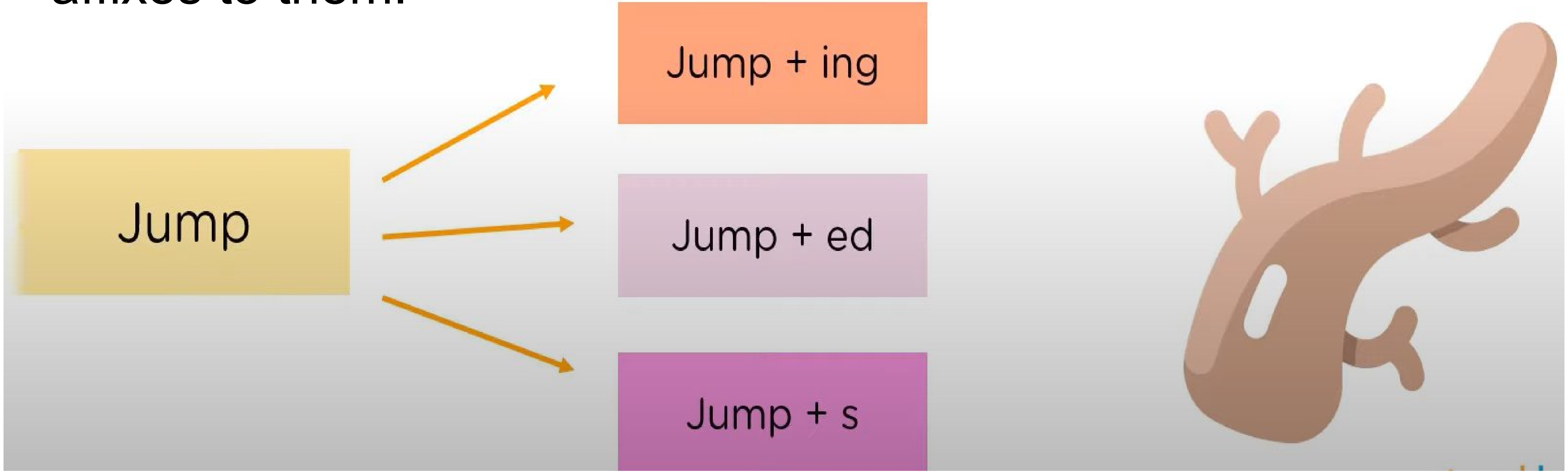
# NLP Pipeline - Tokenization

- Tokenization: The process of splitting sentences into their constituent words.



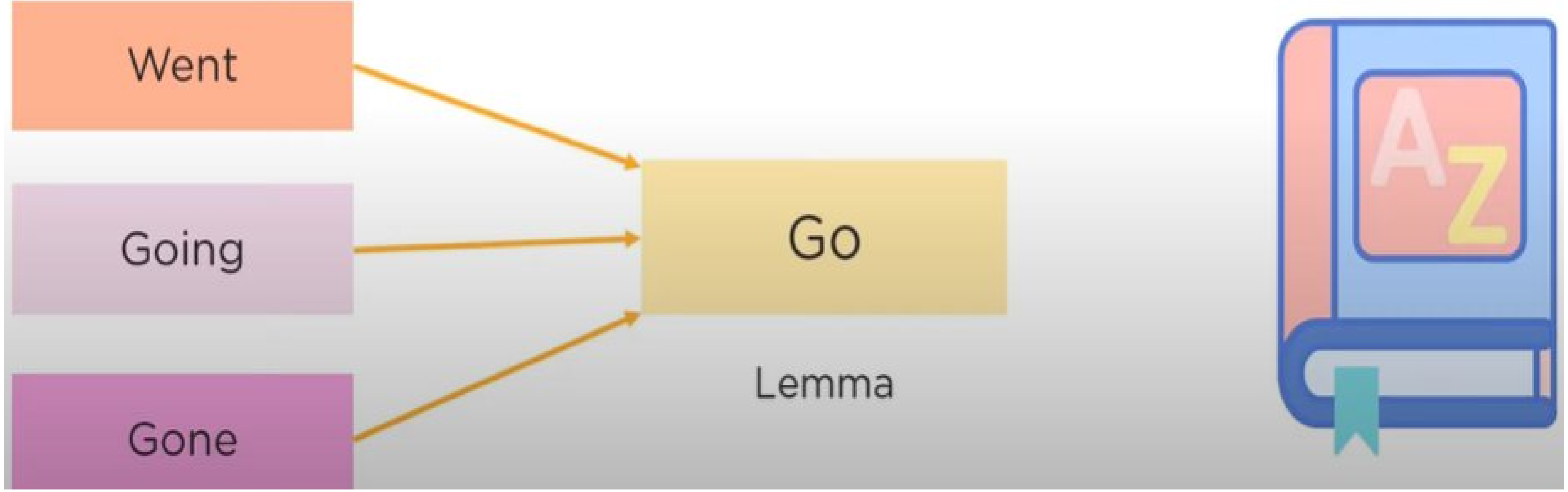
# NLP Pipeline - Stemming

- Stemming: The process of obtaining the Word Stem of a word. Word Stem give new words upon adding affixes to them.



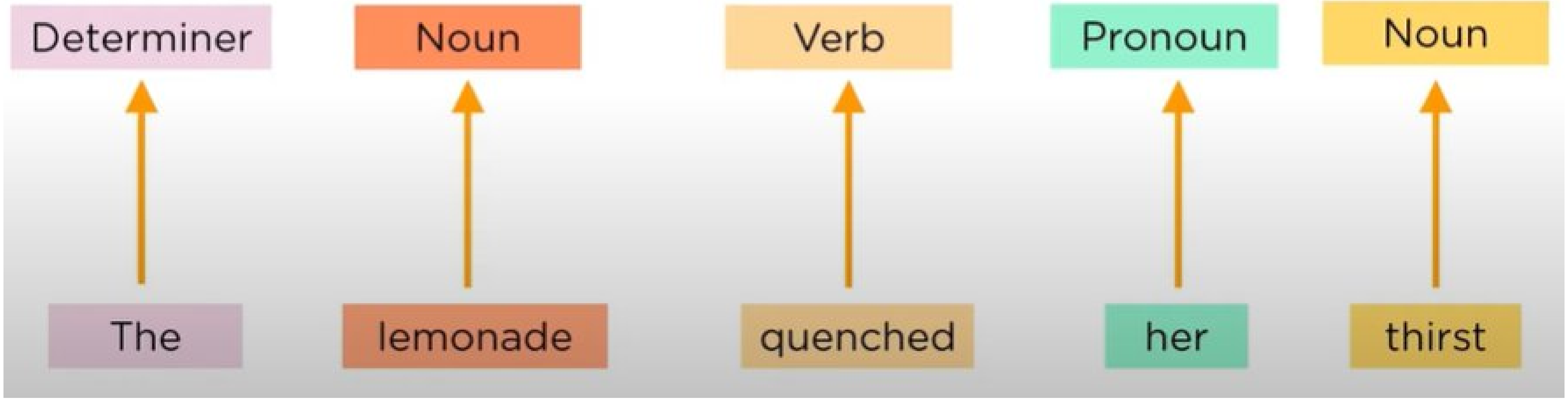
# NLP Pipeline - Lemmatization

- Lemmatization: The process of obtaining the Root Stem of a word. Root Stem give new base form of a word.



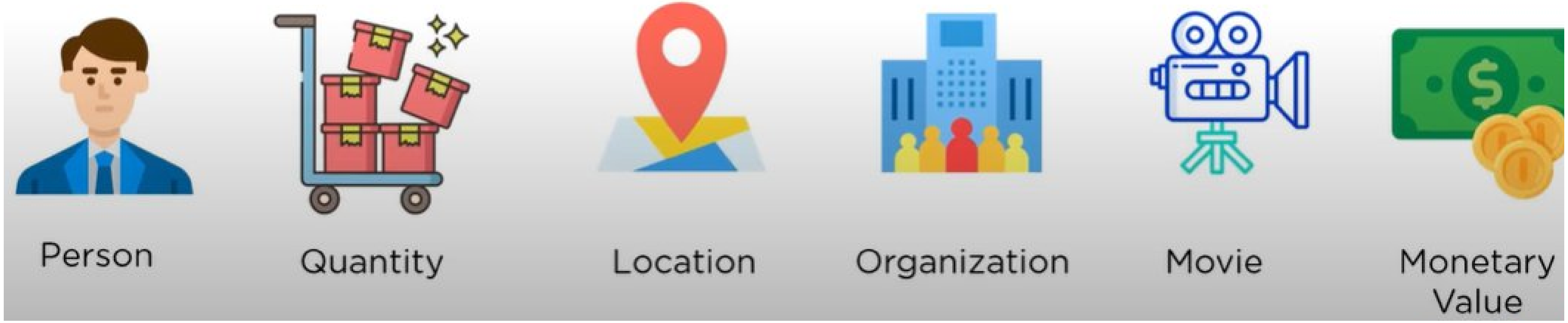
# NLP Pipeline – Parts of Speech Tagging

- Parts of Speech Tagging: Identifies which part of speech a word belongs to. It tags a word as a verb, noun, pronoun etc.



# NLP Pipeline – Named Entity Recognition

- Named Entity Recognition: Classifying the words into subcategories. The subcategories are:



# Thank you