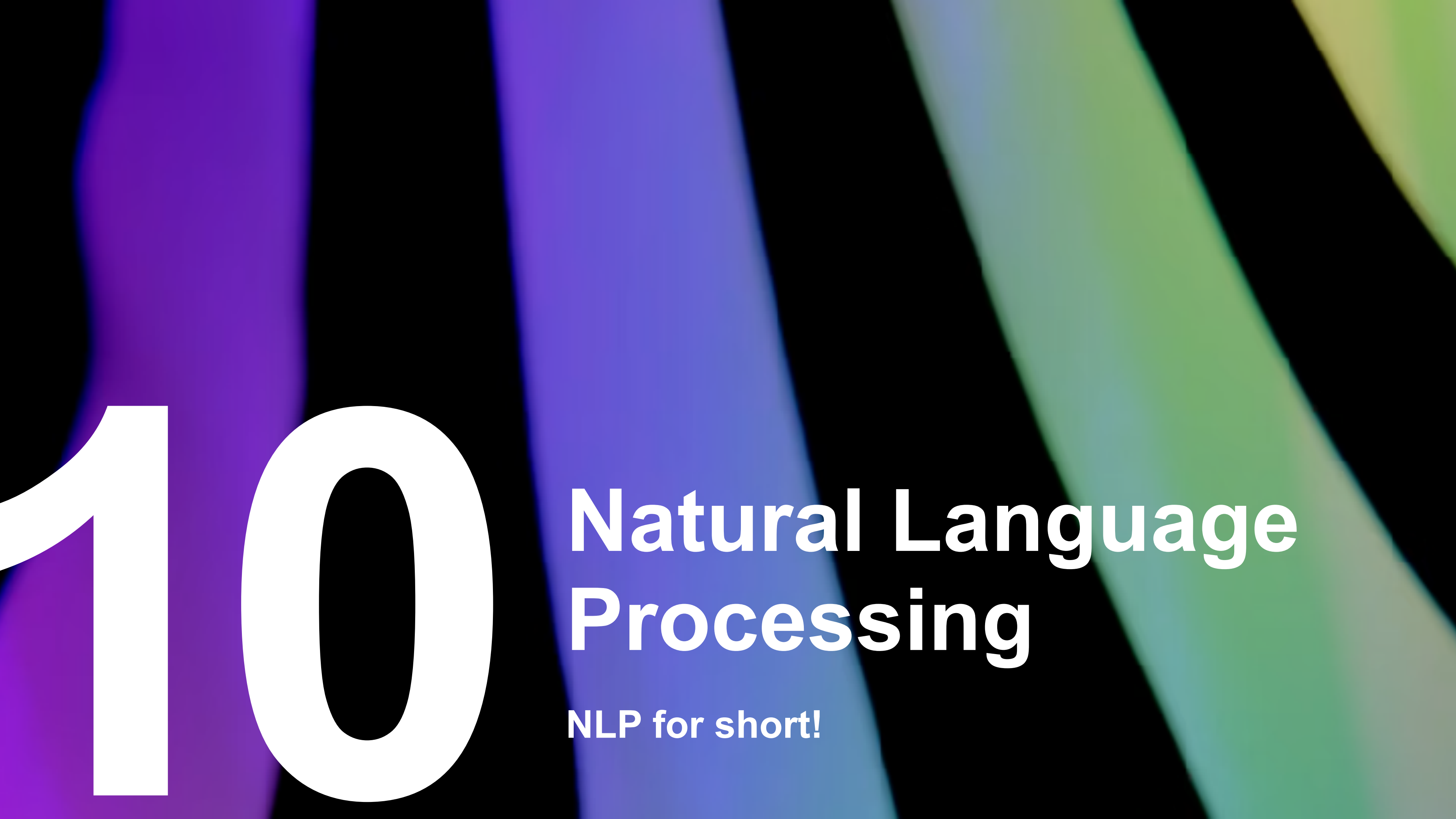

10 Natural Language Processing

NLP for short!

10.1
Understanding **NLP.**

What is NLP - Natural Language Processing

- NLP is a branch of artificial intelligence that deals with the interaction between computers and humans **through the natural languages**
- The **objective** of NLP is to read, decipher, understand, and make sense of human languages in a valuable way
- NLP combines **computational linguistics**—rule-based modeling of human language—with statistical, machine learning, and deep learning models

Uses of NLP:

There are many instances where NLP is utilized to provide value to both companies and individuals. Some of these examples are:

- Classification: Sentiment Analysis, Token classification, gender text alignment
- Sequence Generation: ASR, QA, Fill-Mask, NSP, Translation
- Multiple Choice: Choosing between several candidates

Sentiment Analysis

- Sentiment Analysis is the automated process of **identifying** and **categorizing opinions** expressed in text to determine the writer's attitude **towards a particular topic or product**
- **Input: Text / Audio, Output: Token**
- *Example*: A company uses sentiment analysis to **monitor social media mentions** of their brand, quickly identifying and addressing customer complaints, and leveraging positive feedback for marketing campaigns

Token Classification

- **Token Classification** is the process of **identifying** types of tokens present in a **text**.
- **Input: Text / Token, Output: Token**
- *Example*: **Token Classification** might be introduced in a **e-commerce** site to identify **important notions** in a comment.

QA (Question Answering)

- The task of **Question Answering** can **answer** a **question** given a **context** and sometimes **without context**
- **Input: Text Sequence, Output: Text Sequence**
- *Example*: **QA chatbots** or **search engine models** like **ChatGPT** and **BingAI** to answer general closed domain questions
- Types: **Extractiva QA**, **OpenQA**, **Few Shot QA**

ASR (Automatic Speech Recognition)

- An **ASR** system **recognizes** and **processes** audio to **generate transcriptions** relevant to the audio.
- **Input: Audio Sequence, Output: Text Sequence**
- *Example*: **Speech-to-Text** systems for **ease of machine use** and allows a large domain of people to **utilize computer systems**.

Chatbots Elevating Customer Experience

- **Chatbots**, trained for a particular task, utilize NLP to understand and **respond to human queries effectively**
- *Real-life Example:* Eva - HDFC Bank's AI chatbot **EVA**, which handles over **20,000 conversations daily**, providing customers with instant support and banking services, improving service efficiency and customer satisfaction

Breaking Language Barriers with Translation

- Translations powered by NLP have gone beyond simple text conversion to **understanding context** and **cultural nuances**, which increases accuracy of said translation
- *Real-life Example:* Services like **Google Translate** support real-time translation of conversations and document translations, aiding travelers and international businesses in overcoming language obstacles

NLP in Education

- Language learning apps, automated grading systems, and personalized learning experiences
- *Real-life Example:* **Duolingo**, a language learning app, uses NLP to provide instant feedback on pronunciation and grammar, tailoring lessons to the user's learning pace

10.2
Linguistic Fundamentals.

Understanding Linguistic Fundamentals in NLP

- NLP relies on the understanding of linguistic principles to accurately interpret and generate human language
- Grasping these fundamentals is crucial for creating sophisticated NLP models that can accurately mimic human understanding and production of language

Syntax - The Grammar of Language

- **Syntax** refers to the rules that govern the **structure of sentences**, including word order, punctuation, and grammatical correctness
- **Relevance**: Syntax analysis in NLP helps in parsing sentences and understanding the grammatical relationships between words, which is **vital** for sentence structure analysis and error correction in text

Example of Syntax

In the sentence: *"The quick brown fox jumps over the lazy dog."*

Syntax analysis helps an NLP system to understand that "fox" is the **subject** of the sentence and "jumps" is the **action** being performed **even if the sentence is rearranged** to:

"Over the lazy dog, the quick brown fox jumps."

Semantics - The Meaning behind Words

- **Semantics** involves the interpretation and meaning of words, phrases, and sentences beyond their literal definition
- **Relevance**: Semantic analysis allows NLP systems to **comprehend context**, **ambiguity**, and the **intent** behind language, which is crucial for tasks like sentiment analysis and language translation

Example of Semantics

Consider the word "**bank**."

In "I need to go to the **bank** to withdraw money,"

versus "The river **bank** is flooded,"

Semantics helps NLP differentiate between:
a **financial institution** and the **land alongside a river**

Structure - Organizing Language Systematically

- **Structure** refers to how language is organized at different levels, from sentences to paragraphs to entire texts
- **Relevance**: Understanding structure is essential for **text generation**, **summarization**, and **information extraction**, as it helps maintain coherence and logical flow in language processing tasks

Example of Structure

In a news article, an NLP system uses structural cues to extract the main topic from the headline, summarize content from each paragraph, and identify the overall theme of the article.

A headline is often a sentence.

The topic sentence is the first sentence of a paragraph.

Endings often begins with, “In short” or “In conclusion”

Morphology - The Study of Word Formation

- **Morphology** examines the structure of words and how they are formed from smaller units called morphemes (the smallest grammatical unit in a language)
- **Relevance**: Morphological analysis is used in NLP for **stemming** (reducing words to their base form) and **lemmatization** (finding the lemma of a word based on its intended meaning)

Example of Morphology

The word "unbelievable" can be broken down into:

"un-" (a prefix meaning "not"),
"believe" (the root word),
and "-able" (a suffix meaning "capable of")

10.3
Text **Preprocessing.**

Why NLP Requires Special Preprocessing

Text data is unstructured and often noisy. Special preprocessing is required in NLP to

- Remove irrelevant characters and words that could mislead the analysis.
- Reduce complexity to improve computational efficiency.
- Enhance the model's ability to generalize from the training data.
- Address the intricacies and nuances of human language.

Stopwords and Their Removal

Stopwords are commonly used words (such as 'the', 'is', 'at') that are filtered out before processing since they add noise without informative content

- Benefits include focused analysis and faster processing.
- Caution is advised as some stopwords can change the meaning of a sentence (e.g., 'not').

Punctuation in Text Preprocessing

Punctuation marks are often removed during text preprocessing because:

- They can be irrelevant for understanding the meaning of texts, especially in models focusing on individual words.
- However, in certain contexts like sentiment analysis, exclamation points or question marks can carry sentiment and should be preserved.

Normalization and Lemmatization

Normalization standardizes text, such as converting to lowercase, while lemmatization reduces words to their base or dictionary form

- Helps in reducing the number of unique tokens in the text.
- Lemmatization takes into account the morphological analysis of the words, aiming to remove inflectional endings only and to return the base or dictionary form of a word.

Part of Speech Tagging

Part of Speech (POS) tagging assigns word types to each word (noun, verb, adjective, etc.).

- Essential for understanding the structure of sentences
- Helps in disambiguating words that can represent more than one part of speech (e.g., 'can' as a verb or a noun).

Tokenization

Tokenization is the process of breaking text into individual terms or tokens.

- Can be as simple as splitting by space, or as complex as recognizing words in a sentence.
- Critical for further processing steps like POS tagging or syntactic parsing.
- N-gram utilization for proper tokenization depending on data

Utilizing Preprocessing Techinques

Such preprocessing techniques will not always be effective

- Issues with such preprocessing.
- Understanding when it is required.
- Identifying tasks where it is necessary

END.