

# Day62\_AI\_GenAI\_AgenticAI\_Intro\_to\_NLP

August 12, 2025

## From Today: Starting AI, Generative AI, and Agentic AI

From today, we begin our journey into **Artificial Intelligence (AI)** — the science of making machines think and act like humans.

### What we will explore:

1. **AI (Artificial Intelligence)** – Machines performing tasks that usually require human intelligence.
2. **Generative AI (Gen AI)** – AI models that can **create** new content such as text, images, music, and code.  
*Example:* ChatGPT, Midjourney, DALL · E.
3. **Agentic AI** – AI systems that can **autonomously take actions** to achieve goals, often combining multiple AI tools, memory, and planning.  
*Example:* An AI travel planner that books tickets, suggests itineraries, and adjusts plans automatically.

## First Stop: NLP (Natural Language Processing)

Before AI can think or act for us, it needs to **understand our language** — and that's where **NLP** comes in.

**NLP** = The field of AI that helps computers understand, interpret, and respond to human language.

Today we'll start with:

- How NLP mimics human language understanding.
- Libraries: **NLTK**, **spaCy**, **Gensim**, **Stanford NLP**.
- Core NLP concepts: **NLU** (Natural Language Understanding) & **NLG** (Natural Language Generation).
- A practical demo: Tokenizing text into words, sentences, and paragraphs.

## 1 Introduction to NLP (Natural Language Processing)

### What is NLP?

Natural Language Processing (**NLP**) is a branch of **Artificial Intelligence** that helps computers understand, interpret, and respond to human language.

It acts like a bridge between **human communication** and **computer understanding**.

### How it works:

NLP tries to mimic how the **human brain processes language** by:

1. Breaking down language into smaller parts (words, sentences, paragraphs).
2. Understanding meaning using grammar, syntax, and semantics.
3. Using context to respond or take action.

**Common NLP Applications** - Voice assistants (Alexa, Siri, Google Assistant) - Spell checkers (Google Docs, MS Word) - Chatbots (customer service, bookings) - Translation (Google Translate, DeepL) - Information extraction (search engines) - Keyword search (Google, Amazon search) - Making appointments, buying items, etc.

## Two Main Parts of NLP

1. **NLU** – Natural Language Understanding
  - The computer reads text/speech and understands its meaning.
  - Examples: Tokenization, POS tagging, Named Entity Recognition.
2. **NLG** – Natural Language Generation
  - The computer generates natural-sounding text or speech.
  - Examples: Chatbot replies, text summarization, machine translation.

## Libraries we will use:

- **NLTK** (Natural Language Toolkit) → main library for basic NLP tasks in Python.
- **spaCy** → fast, industrial NLP processing.
- **Gensim** → topic modeling, word embeddings.
- **Stanford NLP (Stanza)** → advanced linguistic analysis.

## Hierarchy of Text in NLP

1. **Token (word)** → smallest unit of text.
2. **Sentence** → collection of tokens that form a complete thought.
3. **Paragraph** → collection of sentences about the same topic.
4. **Document** → collection of paragraphs (article, book chapter, etc.).
5. **Corpus** → large collection of documents used for analysis or training.

## 2 Let's See NLP in Action

Before diving deep into theory, let's see a **practical demo** of how NLP starts working.

We'll:

1. Import the **NLTK** library (main tool for basic NLP in Python).
2. Download the built-in language processing resources.
3. Load a sample **text document** about AI.
4. Use NLTK to break it down into smaller parts (tokens).

This will help us understand what's really happening inside an NLP pipeline before going into detailed concepts.

## 2.1 Step 1 Explanation:

- We import nltk for NLP tasks.
- nltk.download() opens a GUI to install required resources like tokenizers.
- The AI variable stores our example text.

```
[1]: # Step 1: Import NLTK and download resources
import os
import nltk
```

```
[ ]: # Opens the NLTK downloader window (you can download punkt, stopwords, etc.)
nltk.download()
```

```
[2]: # Example AI text (Document)
AI = '''Artificial Intelligence refers to the intelligence of machines. This is
↳in contrast to the natural intelligence of
humans and animals. With Artificial Intelligence, machines perform functions
↳such as learning, planning, reasoning and
problem-solving. Most noteworthy, Artificial Intelligence is the simulation of
↳human intelligence by machines.
It is probably the fastest-growing development in the World of technology and
↳innovation. Furthermore, many experts believe
AI could solve major challenges and crisis situations.'''
```

## Tokenization – The First Step in NLP

Before a machine can understand text, it must **break it into parts**:

- **Word Tokenization** → Splitting text into words (tokens).
- **Sentence Tokenization** → Splitting text into sentences.

These tokens are the building blocks for every other NLP task.

## 2.2 Step 2 Explanation:

- word\_tokenize() splits text into words and punctuation.
- Each piece is called a token.
- Tokens include words like “Artificial”, “Intelligence” and punctuation like “.”.

```
[3]: # Step 2: Word Tokenization
from nltk.tokenize import word_tokenize

AI_tokens = word_tokenize(AI)
print("Tokens:", AI_tokens)
print("Number of tokens:", len(AI_tokens))
```

```
Tokens: ['Artificial', 'Intelligence', 'refers', 'to', 'the', 'intelligence',
'of', 'machines', '.', 'This', 'is', 'in', 'contrast', 'to', 'the', 'natural',
'intelligence', 'of', 'humans', 'and', 'animals', '.', 'With', 'Artificial',
'Intelligence', ',', 'machines', 'perform', 'functions', 'such', 'as',
'learning', ',', 'planning', ',', 'reasoning', 'and', 'problem-solving', '.']
```

```
'Most', 'noteworthy', ',', 'Artificial', 'Intelligence', 'is', 'the',  
'simulation', 'of', 'human', 'intelligence', 'by', 'machines', '.', 'It', 'is',  
'probably', 'the', 'fastest-growing', 'development', 'in', 'the', 'World', 'of',  
'technology', 'and', 'innovation', '.', 'Furthermore', ',', 'many', 'experts',  
'believe', 'AI', 'could', 'solve', 'major', 'challenges', 'and', 'crisis',  
'situations', '.']
```

Number of tokens: 81

### 2.3 Step 3 Explanation:

- `sent_tokenize()` splits text into sentences.
- It uses punctuation and grammar rules to decide sentence boundaries.
- Now the machine can identify complete sentences.

```
[5]: # Step 3: Sentence Tokenization  
from nltk.tokenize import sent_tokenize  
  
AI_sent = sent_tokenize(AI)  
print("Sentences:", AI_sent)  
print("Number of sentences:", len(AI_sent))
```

```
Sentences: ['Artificial Intelligence refers to the intelligence of machines.',  
'This is in contrast to the natural intelligence of\nhumans and animals.', 'With  
Artificial Intelligence, machines perform functions such as learning, planning,  
reasoning and\nproblem-solving.', 'Most noteworthy, Artificial Intelligence is  
the simulation of human intelligence by machines.', 'It is probably the fastest-  
growing development in the World of technology and innovation.', 'Furthermore,  
many experts believe\nAI could solve major challenges and crisis situations.']
```

Number of sentences: 6

### 2.4 Step 4 Explanation:

- `blankline_tokenize()` splits text into paragraphs based on blank lines.
- This is helpful when working with larger documents.

```
[4]: # Step 4: Paragraph Tokenization  
from nltk.tokenize import blankline_tokenize  
  
AI_blank = blankline_tokenize(AI)  
print("Paragraphs:", AI_blank)  
print("Number of paragraphs:", len(AI_blank))
```

```
Paragraphs: ['Artificial Intelligence refers to the intelligence of machines.  
This is in contrast to the natural intelligence of\nhumans and animals. With  
Artificial Intelligence, machines perform functions such as learning, planning,  
reasoning and\nproblem-solving. Most noteworthy, Artificial Intelligence is the  
simulation of human intelligence by machines.\nIt is probably the fastest-  
growing development in the World of technology and innovation. Furthermore, many  
experts believe\nAI could solve major challenges and crisis situations.']
```

Number of paragraphs: 1

### Summary

- NLP = making computers understand & work with human language.
- We learned about **NLU** and **NLG**.
- First NLP step is **Tokenization**:
  - Word Tokenization → machine knows individual words.
  - Sentence Tokenization → machine knows sentence boundaries.
  - Paragraph Tokenization → machine knows text structure.