Natural Language Processing (NLP)

Why NLP?

Think about your daily life - NLP is everywhere! Here are some examples:

- **Google Search:** When you search "best pizza near me", Google understands your intent and shows relevant results
- **Google Translate:** "Hello" → "Hola" (Spanish) - the system understands one language and converts it to another

Without NLP vs With NLP

Without NLP	With NLP
Computers understand only code (0s and 1s)	Computers understand human language
You must click buttons	You can ask in plain English
No voice assistants	Siri, Alexa, Google Assistant work
No auto-translation	Instant translation in 100+ languages

Where NLP Fits in AI

NLP sits within the broader AI ecosystem:

- **AI** → Machine Learning (ML) → Deep Learning (DL) → Multi-layered Neural Networks
- **NLP Pipeline:** Dataset → Text → Model → Applications

NLP Applications

- Text Summarization
- Chatbots
- Code Suggestions

- Language Translation

Key Challenges

- Human communication is fundamentally different from machine understanding
- **Major Challenge:** Understanding sarcasm and context

NLP Roadmap (Study Guide)

Libraries & Frameworks

ML Libraries	DL Frameworks
NLTK	TensorFlow
SpaCy	PyTorch
TextBlob	-

Key Architectures

- BERT (Bidirectional Encoder Representations)
- Transformers
- Bidirectional LSTM, Encoders, Decoders
- RNN, LSTM, GRU, RNN → Deep Learning models

Text Preprocessing Pipeline

Dataset → Text Preprocessing → Text Processing → Word Vectors

Step 1: Text Preprocessing

1. Tokenization

Converting sentences into words (tokens)
Example: "You won 1000000$" → ["You", "won", "1000000", "$"]

2. Stop Words Removal

Removing non-meaningful words (the, is, a, an, etc.)

Example: "Hey buddy I want to go to your house" → Remove "Hey", "I", "to", "your"

3. Stemming

Reducing words to their base/root form (may not be a valid word)
Example: "Historical", "History" → "histori"

- **Advantage:** Fast processing
- **Disadvantage:** May lose word meaning

4. Lemmatization

Reducing words to their dictionary form (always a valid word)
Example: "History", "Historical" → "history" | "Finally", "Final", "Finalized" → "final"

- **Advantage:** Produces meaningful words
- **Disadvantage:** Slower than stemming

When to Use What?

Stemming Use Cases	Lemmatization Use Cases
Spam Classification	Text Summarization
Review Classification	Language Translation
-	Chatbots

Step 2: Text Processing (Words → Vectors)

1. One Hot Encoding

Each word is represented as a binary vector
Example Corpus: D1: "A man eat food" | D2: "Cat eat food" | D3: "People watch Krish"

Vocabulary: [A, man, eat, food, cat, people, watch, krish]
D1 encoding: [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]

2. Bag of Words (BoW)

Count frequency of each word in vocabulary
Process: Remove stop words → Lowercase all words → Count frequencies

Vocabulary	Frequency
good	3
boy	2
girl	2

BoW: Pros & Cons

Advantages ✓	Disadvantages ✗
Simple to implement	Sparse Matrix (lots of zeros)
Intuitive understanding	Out of vocabulary problem
-	Can't capture semantic meaning
-	Word ordering is lost

3. TF-IDF (Term Frequency - Inverse Document Frequency)

Weighs words by importance: common words get lower scores, rare words get higher scores
Captures semantic meaning between words better than simple frequency counts

4. Word2Vec

Creates dense vector representations that capture semantic relationships
Captures semantic information using N-grams

Similarity Measures

Cosine Similarity

Measures the angle between two vectors

- cos(90°) = 0 → Completely different (perpendicular)
- cos(0°) = 1 → Identical (same direction)

Application: Recommendation Systems (e.g., Avengers and Iron Man are similar movies)

Euclidean Distance

Measures the straight-line distance between two points in vector space

Basic NLP Terminology

Term	Definition
Corpus	Paragraph / All the data points
Document	Sentence / Each data instance
Vocabulary	Unique words in the dataset
Words/Tokens	Separate word units

Practical Example: Email Spam Classifier

Email Body	Email Subject	Output
You won 100000$	Billionaire	**SPAM**
Hey, how are you	Hello	**HAM**
Credit card worth	Winner	**SPAM**

Pipeline: Tokenization → Stop Words Removal → Stemming → Lemmatization → Classification