Natural Language Processing:

What is NLP?

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that enables machines to understand, interpret, and generate human language. It combines computational linguistics with machine learning and deep learning.

What is NLTK?

NLTK (Natural Language Toolkit) is one of the most popular Python libraries for working with text data in Natural Language Processing (NLP). It provides tools for text preprocessing, tokenization, stemming, lemmatization, and more.

Components of NLP

Text Preprocessing:

- **Tokenization**: Splitting text into words or sentences.
- **Stopword Removal**: Removing common words (e.g., "is", "the") that don't add much meaning.
- **Stemming**: Reducing words to their root form (e.g., "running" → "run").
- **Lemmatization**: Converting words to their dictionary form (e.g., "better" → "good").

Text Representation:

- **Bag of Words (BoW):** Represents text as a collection of word counts.
- **TF-IDF**: Weighs words based on importance (Term Frequency-Inverse Document Frequency).
- **Word Embeddings**: Dense vector representations of words (e.g., Word2Vec, GloVe).

Parsing and Understanding:

- **Part-of-Speech Tagging**: Identifies grammatical roles (noun, verb, etc.).
- **Named Entity Recognition (NER)**: Detects entities like names, dates, locations.
- **Dependency Parsing**: Analyzes sentence structure to understand relationships between words.

What is Garbage Data in NLP?

In **Natural Language Processing (NLP)**, **garbage data** refers to textual data that is irrelevant, noisy, or poorly formatted, which can significantly impact the quality of NLP models.

Impact of Garbage Data on NLP

- Decreased Model Performance
- Increased Noise
- Inefficient Training
- Bias Introduction

How to Handle Garbage Data in NLP

- **Data Cleaning**:
- **Text Filtering**:
- **Handling Redundancy**:
- **Preprocessing**:
  - Tokenization
  - Stemming/Lemmatization
  - Sentence Segmentation
- **Outlier Detection in Text**

Tokenization: Tokenization is the process of splitting text into smaller units, called tokens, such as words, sentences, or phrases. These tokens are the basic building blocks used in natural language processing (NLP).

Tokenization Types

Sentence Tokenizer

Word Tokenizer

Word Punctuation Tokenizer

Treebank Word Tokenizer

Sentence Tokenizer :

A sentence tokenizer is a tool in natural language processing (NLP) that splits text into individual sentences. This is useful for analyzing or processing text at the sentence level.

- **Word Tokenizer:**
  A word tokenizer is a tool in Natural Language Processing (NLP) that breaks a piece of text into individual words or tokens.

- **Word Punctuation Tokenizer :**
  A Word Punctuation Tokenizer is a tool used in natural language processing to split text into tokens (words, punctuation marks) while keeping punctuation as separate tokens.

- **Treebank Word Tokenizer :**
  The Treebank Word Tokenizer is a tokenizer in NLP that splits text into words, handling punctuation and contractions according to the Penn Treebank tokenization rules.

- **Stemming:**
  Converting the words into their root format, it may not have meaning.

  Stemming is the process of removing the last few characters of a given word, to obtain a shorter form, even if that form doesn’t have any meaning.

- **Porter Stemmer:**

  The Porter Stemmer is a tool in natural language processing (NLP) that reduces words to their base or root form by removing common suffixes.

- **Regex Stemmer (Regular Expression Stemmer)**

  A regex stemmer simplifies a word to its root form using a set of regular expression rules. It's a basic, rule-based approach to stemming, often faster and simpler than algorithmic stemmers like Porter or Snowball.

- **Snowball Stemmer:**

  The Snowball Stemmer is a tool used in natural language processing (NLP) to reduce words to their root form. It works by removing suffixes and simplifying words while keeping their core meaning.

- **Lemmatization:**
  Converting the words into their root format, with meaningful word.

Lemmatization takes more time as compared to stemmimg because it finds meaningful word/ representation.
Stemming just needs to get a base word and therefore takes less time.

POS Tags:

Part-of-Speech (POS) tags are labels assigned to words in a sentence to indicate their grammatical role.

Common POS Tags:

1. **NOUN (NN)** - Names of people, places, things, or concepts (e.g., cat, book, happiness).
    - Singular: NN (dog), Plural: NNS (dogs), Proper: NNP (London), Proper Plural: NNPS (Americans).
2. **PRONOUN (PRP)** - Words that replace nouns (e.g., he, she, it).
    - Personal: PRP (he, they), Possessive: PRP$ (his, their).
3. **VERB (VB)** - Action or state (e.g., run, is).
    - Base: VB (run), Past: VBD (ran), Gerund: VBG (running), Past Participle: VBN (run), Singular Present: VBZ (runs).
4. **ADJECTIVE (JJ)** - Describes a noun (e.g., big, beautiful).
    - Comparative: JJR (bigger), Superlative: JJS (biggest).
5. **ADVERB (RB)** - Modifies verbs, adjectives, or other adverbs (e.g., quickly, very).
    - Comparative: RBR (faster), Superlative: RBS (fastest).
6. **PREPOSITION (IN)** - Links nouns to other words (e.g., in, on, at).
7. **CONJUNCTION** - Joins words, phrases, or clauses.
    - Coordinating: CC (and, but), Subordinating: IN (because, although).
8. **DETERMINER (DT)** - Introduces nouns (e.g., the, a, this).
    - Articles: DT (the, a), Quantifiers: CD (five).
9. **INTERJECTION (UH)** - Expresses emotion (e.g., wow, oh).
10. **PARTICLE (RP)** - Small words tied to verbs (e.g., look *up*, run *out*).
11. **NUMERAL (CD)** - Numbers (e.g., one, 100).

12. **PUNCTUATION (.)** - Symbols that structure text (e.g., ., ?, !).

Text Preprocessing Stopwords:

Stopwords in NLP are common words (e.g., "is," "the," "in," "on") that usually carry little meaningful information and are often removed during text preprocessing to focus on the key content of the text.

Why Remove Stopwords?

1. **Focus on Meaningful Words**: Improves the quality of text analysis by emphasizing content-rich words.
2. **Reduce Noise**: Prevents common words from dominating algorithms like text classification or topic modeling.
3. **Improve Efficiency**: Reduces data size and computational load.

1. What is NLP?

NLP is a field of AI that enables machines to understand, interpret, and generate human language.

---

2. What are the main components of NLP?

- **Syntax**: Sentence structure analysis (e.g., POS tagging, parsing).
- **Semantics**: Meaning interpretation (e.g., named entity recognition, word sense disambiguation).

---

3. What is tokenization?

Tokenization is the process of breaking text into smaller units (tokens) such as words, sentences, or subwords.

---

4. What is stemming vs lemmatization?

- **Stemming**: Reduces words to their root form (e.g., "running" → "run").
- **Lemmatization**: Converts words to their dictionary form (e.g., "better" → "good").

---

5. What are stop words?

Common words (e.g., "the," "is") often removed in text processing to focus on meaningful words.

---

6. What is the Bag of Words (BoW) model?

A text representation model where each document is represented as a vector of word counts or frequencies, ignoring grammar and order.

---

7. What is TF-IDF?

- **Term Frequency-Inverse Document Frequency**: A weighting technique to measure how important a word is to a document in a collection.

---

8. What is word embedding?

A dense vector representation of words in a continuous space, capturing semantic meaning (e.g., Word2Vec, GloVe).

---

9. What are common NLP tasks?

- Text classification
- Sentiment analysis
- Machine translation
- Named entity recognition (NER)
- Summarization

---

10. What is the difference between NER and POS tagging?

- **NER**: Identifies entities like names, dates, locations.
- **POS Tagging**: Labels words with grammatical tags (e.g., noun, verb).

---

11. What is the difference between rule-based and ML-based NLP?

- **Rule-based**: Uses predefined linguistic rules.
- **ML-based**: Learns patterns from data using algorithms like SVM, neural networks.

---

12. What is sequence-to-sequence modeling?

An architecture where input sequences (e.g., sentences) are mapped to output sequences (e.g., translations) using models like RNNs, LSTMs, or Transformers.

---

13. What are Transformers in NLP?

A deep learning model architecture that uses self-attention mechanisms, popularized by models like BERT, GPT.

---

14. What is BERT?

- **Bidirectional Encoder Representations from Transformers**: A pre-trained language model for context-aware word embeddings.

---

15. What is attention in NLP?

A mechanism that allows the model to focus on relevant parts of the input sequence when making predictions.

---

16. What is sentiment analysis?

The process of determining the sentiment (positive, negative, neutral) of a text.

---

17. What is language modeling?

Predicting the next word in a sequence (e.g., GPT) or the probability

of a given sequence of words (e.g., n-gram models, neural language models).

---

18. What is text preprocessing in NLP?

Preparing raw text data for analysis by:

- Removing punctuation
- Lowercasing
- Tokenizing
- Removing stop words
- Stemming or lemmatizing

---

19. What is an n-gram?

A contiguous sequence of nnn items (words or characters) from a given text. Examples:

- Unigram: "I"
- Bigram: "I am"
- Trigram: "I am learning"

---

20. What is cosine similarity?

A metric used to measure the similarity between two vectors by computing the cosine of the angle between them, commonly used in document comparison.

---

21. What are common evaluation metrics in NLP?

- **Accuracy**: For classification tasks.
- **BLEU**: For machine translation.
- **ROUGE**: For summarization.
- **F1-Score**: For imbalanced datasets.

---

22. What is transfer learning in NLP?

Using pre-trained models (e.g., BERT, GPT) on a large dataset and fine-tuning them on a specific task.

---

23. What are common challenges in NLP?

- Ambiguity (e.g., polysemy, homonyms)
- Sarcasm and irony
- Context understanding
- Low-resource languages
- Handling idioms and metaphors

---

24. What are some popular NLP libraries?

- **NLTK**: Natural Language Toolkit
- **spaCy**: Fast NLP processing
- **Hugging Face Transformers**: Pre-trained models
- **Gensim**: Topic modeling and word embeddings

---

25. What is conversational AI?

AI systems designed to interact with humans via natural language (e.g., chatbots, virtual assistants).