CHP 1 :INTRODUCTION TO NLP

Introduction to Natural Language Processing (NLP)

Define: Natural Language Processing (NLP) is about teaching computers to understand and process human language. Unlike structured data (like spreadsheets), human language is unstructured and complex, making it challenging for computers to interpret accurately. NLP aims to bridge this gap by making sense of textual and spoken data.

Key Steps in NLP

1. **Sentence Segmentation**
    - **What It Is**: Breaking text into individual sentences.
    - **Example**:
        - Input: "San Pedro is a town... It is the second-largest town..."
        - Output: "San Pedro is a town... It is the second-largest town..."
2. **Word Tokenization**
    - **What It Is**: Splitting sentences into words (tokens).
    - **Example**:
        - Input: "San Pedro is a town..."
        - Output: ['San', 'Pedro', 'is', 'a', 'town', '...']
3. **Predicting Parts of Speech (POS)**
    - **What It Is**: Identifying if each word is a noun, verb, adjective, etc.
    - **Example**:
        - Input: "Town is large."
        - Output: 'Town' – Noun, 'is' – Verb, 'large' – Adjective
4. **Lemmatization**
    - **What It Is**: Reducing words to their root forms.
    - **Example**:
        - Input: "Buffaloes grazing..."
        - Output: 'Buffalo' (root form)

5. **Identifying Stop Words**
    - **What It Is**: Filtering out common words like 'the', 'is', 'and' that don't add much meaning.
    - **Example**: Removing 'the', 'is' from "The cat is on the mat."
6. **Dependency Parsing**
    - **What It Is**: Analyzing the relationships between words in a sentence.
    - **Example**:
        - Input: "San Pedro is a town..."
        - Output: Parse tree with 'is' as the root, showing relationships.
7. **Finding Noun Phrases**
    - **What It Is**: Grouping related words that represent the same idea.
    - **Example**:
        - Input: "Second-largest town in the Belize District"
        - Output: Group 'second-largest town' as a noun phrase.
8. **Named Entity Recognition (NER)**
    - **What It Is**: Identifying and categorizing real-world entities like people, places, and organizations.
    - **Example**:
        - Input: "San Pedro is a town in Belize."
        - Output: 'San Pedro' – Geographic Entity, 'Belize' – Geographic Entity
9. **Coreference Resolution**
    - **What It Is**: Determining which words refer to the same entity.
    - **Example**:
        - Input: "San Pedro is a town. It is in Belize."
        - Output: 'It' refers to 'San Pedro'

Summary

NLP involves several steps to make sense of human language:

- Segmenting text into sentences and words.
- Classifying parts of speech.

- Reducing words to their root forms.
- Filtering out common words.
- Understanding word relationships.
- Grouping related words.
- Identifying real-world entities.
- Resolving references to the same entity.

Text Pre-Processing in NLP

Text pre-processing is a crucial step in preparing raw text for analysis or machine learning models. It involves cleaning and transforming the text to make it more manageable and suitable for processing. Here's a detailed explanation of some key text pre-processing techniques:

1. Regular Expressions (Regex)

What It Is: Regular expressions (regex) are sequences of characters that define a search pattern. They are used for pattern matching within text.

How It Works: Regex allows you to search for specific patterns in text, such as dates, phone numbers, or any other structured data. It can also be used to clean text by removing or replacing unwanted characters.

Example:

- **Pattern**: \d+
  - **Meaning**: Matches one or more digits.
  - **Use**: To find all numbers in a text.
- **Pattern**: \b\w+\b
  - **Meaning**: Matches any word boundary.
  - **Use**: To tokenize text by identifying words.

Example Usage:

- **Text**: "The price is $45.67."
- **Regex Pattern**: \d+(\.\d{2})?
  - **Matches**: 45.67
  - **Use**: Extract prices from text.

2. Tokenization

What It Is: Tokenization is the process of splitting text into smaller units called tokens. These tokens can be words, phrases, or symbols.

How It Works:

- **Word Tokenization**: Splits a text into individual words.
    - **Example**: "I love NLP" → ['I', 'love', 'NLP']
- **Sentence Tokenization**: Splits a text into sentences.
    - **Example**: "I love NLP. It is fascinating." → ['I love NLP.', 'It is fascinating.']

Why It's Important: Tokenization simplifies text into manageable pieces, allowing further analysis like counting word frequencies or applying algorithms.

3. Stemming

What It Is: Stemming is the process of reducing words to their base or root form. It removes prefixes and suffixes to get to the core meaning of a word.

How It Works:

- **Algorithm**: Uses rules to strip affixes from words.
- **Goal**: To standardize different forms of a word.

Example:

- **Word**: "running"
    - **Stemmed Form**: "run"
- **Word**: "fishing"
    - **Stemmed Form**: "fish"

Why It's Important: Stemming helps in reducing the complexity of text data by consolidating different forms of a word into one, improving the effectiveness of text analysis and search.

4. Minimum Edit Distance

What It Is: Minimum edit distance (also known as Levenshtein distance) measures the number of edits (insertions, deletions, or substitutions) needed to transform one string into another.

How It Works:

- **Algorithm**: Calculates the smallest number of operations required to convert one string into another.
- **Example Calculation**:
    - **String 1**: "kitten"

- **String 2**: "sitting"
- **Operations**:
  - Substitute 'k' with 's'
  - Substitute 'e' with 'i'
  - Insert 't' at the end
  - Insert 'i' at the end
  - Insert 'g' at the end
  - Insert 's' at the end
- **Minimum Edit Distance**: 3

Why It's Important: Edit distance helps in tasks like spell-checking, correcting typos, and matching similar strings by quantifying how close or different two strings are.

Summary

- **Regular Expressions**: Pattern matching for extracting or cleaning specific text elements.
- **Tokenization**: Breaking text into smaller units for easier analysis.
- **Stemming**: Reducing words to their root form to standardize text.
- **Minimum Edit Distance**: Measuring the similarity between strings by counting edits needed to transform one into another.

These pre-processing steps are fundamental in preparing text data for more advanced analyses and model training in NLP.

FOR MINIMUM EDIT DISTANCE NUMERICAL YT REFEFENCE: