Handout: of NLP

1. Introduction to NLP

Natural Language Processing (NLP) is a field of artificial intelligence that enables computers to understand, interpret, and generate human language. It combines computational linguistics with machine learning to process and analyze large amounts of natural language data. NLP is essential for various applications such as chatbots, translation services, sentiment analysis, and information retrieval.

NLP focuses on understanding both the structure and meaning of human language, enabling machines to derive meaningful information from texts, voice, or other forms of human communication.

Key Applications of NLP:

- **Text Classification**: Automatically categorizing documents (e.g., spam detection, sentiment analysis).
- **Named Entity Recognition (NER)**: Identifying entities like names, places, dates in a text.
- **Language Translation**: Translating text between different languages.
- **Speech Recognition**: Converting spoken language into text.

2. Basic NLP Terms

a. Tokenization

Tokenization is the process of breaking down a text into smaller pieces called "tokens." A token can be a word, sentence, or punctuation mark. Tokenization is the first step in most NLP tasks as it helps convert raw text into a structured format.

b. Stemming

Stemming involves reducing a word to its base or root form by stripping suffixes (e.g., "running" becomes "run"). This process is less accurate as it often produces non-lexical forms (words that may not exist in the dictionary).

c. Lemmatization

Lemmatization is a more sophisticated process that reduces words to their base form, called a "lemma," based on their meaning and context. It ensures the word is valid (e.g., "running" becomes "run", "better" becomes "good").

d. Named Entity Recognition (NER)

NER is the process of identifying and classifying named entities in a text, such as people, organizations, locations, dates, etc. This is useful for extracting structured information from unstructured text.

3. Example Code Using SpaCy

SpaCy is one of the most popular NLP libraries in Python. It provides efficient tools for tokenization, lemmatization, part-of-speech tagging, NER, and more. Below is a complete example that demonstrates these tasks using dummy inputs.

Installation:

You need to install spacy and download a language model (e.g., en_core_web_sm):

```
pip install spacy
python -m spacy download en_core_web_sm
```

Example with 3x Inputs

```
import spacy

# Load the small English NLP model
nlp = spacy.load("en_core_web_sm")

# Dummy inputs
texts = [
    "Apple is looking at buying U.K. startup for $1 billion.",
    "John Doe works at OpenAI in San Francisco.",
    "The new Tesla model was released in 2021."
]

# Process each text using SpaCy's NLP pipeline
for text in texts:
    doc = nlp(text)

    # Tokenization
    print(f"\nText: {text}")
    print("Tokens:")
    for token in doc:
        print(token.text, end=" | ")

    # Lemmatization and Part-of-Speech (POS) tagging
    print("\n\nLemmas and POS tags:")
    for token in doc:
        print(f"{token.text} -> Lemma: {token.lemma_}, POS: {token.pos_}")

    # Named Entity Recognition (NER)
    print("\nNamed Entities:")
    for ent in doc.ents:
        print(f"{ent.text} -> {ent.label_}")

    print("\n" + "="*40 + "\n")
```

Output:

The output would look like this for the three inputs:

```
Text: Apple is looking at buying U.K. startup for $1 billion.
Tokens:
Apple | is | looking | at | buying | U.K. | startup | for | $ | 1 | billion | . |

Lemmas and POS tags:
Apple -> Lemma: Apple, POS: PROPN
is -> Lemma: be, POS: AUX
looking -> Lemma: look, POS: VERB
at -> Lemma: at, POS: ADP
buying -> Lemma: buy, POS: VERB
U.K. -> Lemma: U.K., POS: PROPN
startup -> Lemma: startup, POS: NOUN
for -> Lemma: for, POS: ADP
$ -> Lemma: $, POS: SYM
1 -> Lemma: 1, POS: NUM
billion -> Lemma: billion, POS: NUM
. -> Lemma: ., POS: PUNCT

Named Entities:
Apple -> ORG
U.K. -> GPE
$1 billion -> MONEY

========================================

Text: John Doe works at OpenAI in San Francisco.
Tokens:
John | Doe | works | at | OpenAI | in | San | Francisco | . |

Lemmas and POS tags:
John -> Lemma: John, POS: PROPN
Doe -> Lemma: Doe, POS: PROPN
works -> Lemma: work, POS: VERB
at -> Lemma: at, POS: ADP
OpenAI -> Lemma: OpenAI, POS: PROPN
in -> Lemma: in, POS: ADP
San -> Lemma: San, POS: PROPN
Francisco -> Lemma: Francisco, POS: PROPN
. -> Lemma: ., POS: PUNCT

Named Entities:
John Doe -> PERSON
OpenAI -> ORG
San Francisco -> GPE

========================================

Text: The new Tesla model was released in 2021.
Tokens:
The | new | Tesla | model | was | released | in | 2021 | . |

Lemmas and POS tags:
The -> Lemma: the, POS: DET
```

```
new -> Lemma: new, POS: ADJ
Tesla -> Lemma: Tesla, POS: PROPN
model -> Lemma: model, POS: NOUN
was -> Lemma: be, POS: AUX
released -> Lemma: release, POS: VERB
in -> Lemma: in, POS: ADP
2021 -> Lemma: 2021, POS: NUM
. -> Lemma: ., POS: PUNCT

Named Entities:
Tesla -> ORG
2021 -> DATE

========================================
```