STA 3180 Statistical Modelling: Natural Language Processing

# Natural Language Processing (NLP) - STA 3180 Statistical Modelling

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) that deals with the analysis, understanding, and generation of natural language. It is used to develop applications such as text-to-speech systems, automatic summarization, machine translation, question answering, and natural language understanding. NLP is a rapidly growing field with many applications in various industries.

## Key Concepts

* **Text Preprocessing**: Text preprocessing is the process of cleaning and preparing text data for further analysis. It involves removing punctuation, stop words, and other irrelevant words, as well as tokenizing and lemmatizing the text.

* **Tokenization**: Tokenization is the process of breaking a sentence into individual words or phrases.

* **Lemmatization**: Lemmatization is the process of reducing a word to its base form. For example, the lemma of the word “running” is “run”.

* **Part-of-Speech Tagging**: Part-of-speech tagging is the process of assigning a part-of-speech tag to each word in a sentence. The most common tags are nouns, verbs, adjectives, adverbs, and pronouns.

* **Named Entity Recognition**: Named entity recognition is the process of identifying and classifying named entities in a text, such as people, organizations, locations, and dates.

* **Sentiment Analysis**: Sentiment analysis is the process of determining the sentiment of a text, such as whether it is positive, negative, or neutral.

## Coding Examples

### Text Preprocessing

```
Start of Code

import nltk
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

text = "This is an example of text preprocessing."

# Tokenize the text
tokens = word_tokenize(text)

# Remove punctuation
words = [word for word in tokens if word.isalpha()]
```

```
# Lemmatize the words
lemmatizer = WordNetLemmatizer()
lemmas = [lemmatizer.lemmatize(word) for word in words]
print(lemmas)
End of Code
```

### Part-of-Speech Tagging

```
Start of Code
import nltk
from nltk.tokenize import word_tokenize
text = "This is an example of part-of-speech tagging."
# Tokenize the text
tokens = word_tokenize(text)
# Tag the tokens
tagged_tokens = nltk.pos_tag(tokens)
print(tagged_tokens)
End of Code
```

## Practice Multiple Choice Questions

Q1. What is the process of breaking a sentence into individual words or phrases called?

A. Tokenization

B. Lemmatization

C. Part-of-speech tagging

D. Named entity recognition

Answer: A. Tokenization