

Name - Ch. Jagan Subudhi

Assessment

Reg no - 2201020364

10 MARKS

CRANES Reg no - CL20250106018923105

Group - AI & DS (5)

① Write a python function using scikit-learn a dataset into training and testing sets with a test size of 25%, use 'train-test-split'.

Ans. From sklearn.model_selection import train_test_split.

def Split-dataset(X, Y):

X_train, X_test, y_train, y_test = train_test_split

(X, y, test_size = 0.25, random_state = 42)

return X_train, X_test, y_train, y_test

② Write python code using NLTK or Spacy to tokenize the following sentence:

'Natural Language Processing enables Computers to understand human language.'

Ans - NLTK

~~Import~~ import

import nltk

nltk.download('punkt')

from nltk.tokenize import word_tokenize

sentence = 'Natural Language Processing enables Computers to understand human language.'

tokens = word_tokenize(sentence)

print(tokens)

Spacy

import spacy

nlp = spacy.load('en-core-web-sm')

doc = nlp('Natural Language Processing enables Computers to understand human language')

tokens = [token.text for token in doc]

print(tokens)

③ Write a code snippet to fit a linear regression model using Scikit-learn on sample data and print the model coefficients.

Ans.

```
from sklearn.linear_model import LinearRegression  
import numpy as np
```

```
# Sample data
```

```
X = np.array([[1], [2], [3], [4], [5]])
```

```
Y = np.array([2, 4, 6, 8, 10])
```

```
model = LinearRegression()
```

```
model.fit(X, Y)
```

```
print("Coefficients:", model.coef_)
```

```
print("Intercept:", model.intercept_)
```

④ Explain the difference between Stemming and Lemmatization in Natural Language processing.

Ans. Stemming reduces a word to its root form by chopping off suffixes, often without considering the context, which may lead to incorrect roots. Lemmatization considers the word's meaning and part of speech to return a correct base

⑤ What is overfitting in machine learning? How can it be prevented?

Ans. overfitting occurs when a machine learning model learns the training data too well, including its noise and outliers, which reduces its ability to generalize to new data. It can be prevented using techniques like cross-validation, pruning, regularization, reducing model complexity or using more training data.