Assessment: Natural Language Processing (NLP) on Databricks using NLTK or spaCy

Duration:

4 Hours (to be delivered by 14th Feb 2025, 4:00 PM). No deadline will be extended.

Support available only during session hours.

Objective:

Assess the understanding and practical implementation of NLP techniques using NLTK or spaCy on Databricks.

Instructions:

- Choose either NLTK or spaCy for completing the tasks.

- Implement the code in a Databricks notebook and submit the results.

- Provide explanations for each step in markdown cells.

- Ensure clean, well-commented code.

Assessment Tasks:

Task 1: Text Preprocessing (20 Marks)

1. Load the following text in Databricks:

"Databricks provides a unified analytics platform to accelerate data science workflows. NLP is a powerful tool for extracting insights from text."

2. Perform the following preprocessing steps:

- Tokenization

- Stopword removal

- Lemmatization (or stemming)

Deliverable: A DataFrame with original text and processed text columns.

Task 2: Named Entity Recognition (NER) (20 Marks)

1. Use spaCy or NLTK to extract Named Entities from the following text:

"Databricks was founded by the creators of Apache Spark in 2013. It is headquartered in San Francisco and serves global enterprises."

2. Display the recognized entities and their categories.

Deliverable: A table with Named Entities and their categories.

Task 3: Sentiment Analysis (20 Marks)

1. Load the following review dataset as a list:

reviews = ["Databricks makes AI development seamless!", "The NLP processing speed is slow and needs improvement.", "Excellent platform with great support!"]

2. Perform sentiment analysis and categorize each review as positive, negative, or neutral.

Deliverable: A table showing each review with its sentiment category.

Task 4: Part-of-Speech (POS) Tagging (20 Marks)

1. Perform POS tagging on the sentence:

"Natural language processing enables machines to understand human text."

2. Display the results as a DataFrame with words and their respective POS tags.

Deliverable: A table showing words and their POS tags.

Task 5: Custom NLP Pipeline (20 Marks)

1. Define a custom NLP pipeline in NLTK or spaCy that:

- Tokenizes the text

- Removes stopwords

- Performs lemmatization

- Identifies Named Entities

2. Test the pipeline on the following text:

"Azure and Databricks are key players in cloud-based AI solutions."

Deliverable: A function implementing the pipeline and displaying results.

Grading Criteria:

- Code Execution (40 Marks): Proper implementation and correct outputs.

- Concept Understanding (30 Marks): Clear explanations of each step.

- Code Readability (20 Marks): Well-structured, commented code.

- Innovation (10 Marks): Use of additional NLP techniques beyond the given tasks.