Assignment: NLP Analysis of Customer Support Chat Data

Scenario

You are working as a data scientist for "QuickHelp," a customer support company that handles thousands of support queries across various products. Your task is to analyze a sample of customer support chat data to better understand common customer concerns, streamline support services, and provide insights for improving chatbot responses.

Task Overview

Using a provided chat log dataset, preprocess and analyze the text data to extract key insights. Apply NLP techniques like tokenization, stemming, lemmatization, POS tagging, and Named Entity Recognition (NER) to identify patterns and trends in customer queries.

Tasks

1. **Data Preprocessing**
    - **Tokenization**: Tokenize the chat data, splitting sentences and words into tokens.
    - **Stemming and Lemmatization**: Perform stemming and lemmatization on the tokenized words, and compare their results on extracted tokens. Explain when each method is appropriate.
2. **POS Tagging**
    - Use POS tagging to categorize tokens, identifying their grammatical roles. Discuss any observed patterns in word usage, and identify commonly used nouns, verbs, and adjectives in customer queries.
3. **Named Entity Recognition (NER)**
    - Implement NER to extract entities like product names, locations, dates, and customer names. Group queries by the identified product or location to see if certain regions or products require more support.
4. **Analysis & Insights**
    - **Common Issues**: Identify frequently occurring words or phrases related to complaints or common questions.
    - **Topic Discovery**: Apply basic topic analysis to find out if there are recurring themes in customer queries.
    - **Sentiment Insight**: Based on the lemmatized and stemmed words, provide a brief sentiment analysis by classifying queries as positive, neutral, or negative using simple heuristic methods.
5. **Reporting**
    - Write a report summarizing the preprocessing steps and insights derived from the data. Include sample code and analysis to support your findings.

Dataset

Each student will receive a subset of customer chat logs. The dataset contains columns like:

- `customer_id`
- `chat_date`
- `message_text`

Deliverables

- **Jupyter Notebook** with code, explanations, and output for each step.
- **Summary Report** explaining the preprocessing methods, analysis steps, and insights.

Chat Log Dataset

customer_id	chat_date	message_text
101	2024-01-10 09:00:00	"I'm having trouble logging into my account, please help me."
102	2024-02-04 09:15:00	"How do I change my password? I forgot the old one."
103	2024-04-10 09:30:00	"Can I return a product that I bought last week? It's defective."
104	2024-05-20 09:45:00	"When will my order be shipped? I haven't received any updates."
105	2024-06-08 10:00:00	"I need to update my shipping address for my recent order."
106	2024-06-20 10:15:00	"Is it possible to get a refund on a defective item I bought a month ago?"
107	2024-07-15 10:30:00	"My credit card was charged incorrectly. Can you assist with that?"
108	2024-08-18 10:45:00	"I was charged twice for the same order. Please check it."
109	2024-10-05 11:00:00	"Do you have any new deals or discounts for the upcoming holiday season?"
110	2024-11-10 11:15:00	"Can you explain the warranty policy on your electronics products?"