NAME : PAREENITA A.SHIRSATH PRN : 221101062 ROLL.NO : 57

B.E.A.I.&D.S.

NLP EXPERIMENT NO : 06

```
import pandas as pd
import re
from collections import defaultdict

# Load dataset
df = pd.read_csv("ngram_dataset.csv")
sentences = df["sentence"].tolist()

def tokenize_text(text):
    text = re.sub(r'[^\\w\\s]', '', text)  # Remove punctuation
    return text.split()

def build_ngram_model(sentences, n):
    model = defaultdict(lambda: defaultdict(int))
    for sentence in sentences:
        tokens = tokenize_text(sentence)
        for i in range(len(tokens) - n + 1):
            context = tuple(tokens[i:i + n - 1])
            next_word = tokens[i + n - 1]
            model[context][next_word] += 1
    return model

def predict_next_word(model, context):
    if context not in model:
        return "unknown"
    next_words = model[context]
    return max(next_words, key=next_words.get)

# Build bigram model
n = 2
model = build_ngram_model(sentences, n)

# Test predictions
contexts = [
    ("The",),
    ("The", "sun"),
    ("The", "river"),
    ("The", "cat")
]

for context in contexts:
    print(f"Context: {context}, Predicted next word: {predict_next_
```

```
Context: ('The',), Predicted next word: unknown
Context: ('The', 'sun'), Predicted next word: unknown
Context: ('The', 'river'), Predicted next word: unknown
Context: ('The', 'cat'), Predicted next word: unknown
```

ngram_dataset.csv

1 to 10 of 10 entries Filter

sentence
The sun rises in the east
The moon shines at night
The stars twinkle in the sky
The earth revolves around the sun
The river flows towards the sea
Birds fly in the blue sky
The train runs on the railway track
The children play in the park
The cat sleeps on the mat
The dog barks at strangers

Show 10 per page