NAME : SAI CHARAN.P

ROLL:NO: 2403a52343

LAB : 03

Aim: Understand POS tagging challenges in informal, noisy text.

Step 1:Install and import Libraries

```
!pip install nltk
import nltk
from nltk.tokenize import TweetTokenizer
from nltk.corpus import twitter_samples
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.12/dist-packages (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.12/dist-packages (from nltk) (8.3.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.12/dist-packages (from nltk) (1.5.3)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.12/dist-packages (from nltk) (2025.11.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-packages (from nltk) (4.67.1)
```

Step 2:Download Required NLTK Resources

```
nltk.download('twitter_samples')
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger_eng')
```

```
[nltk_data] Downloading package twitter_samples to /root/nltk_data...
[nltk_data]   Unzipping corpora/twitter_samples.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger_eng.zip.
True
```

Step 3: Load Tweets Dataset

```
tweets = twitter_samples.strings('positive_tweets.json')

for i in range(3):
  print("Tweet",i+1)
  print(tweets[i])
  print()
```

```
Tweet 1
#FollowFriday @France_Inte @PKuchly57 @Milipol_Paris for being top engaged members in my community this week :)

Tweet 2
@Lamb2ja Hey James! How odd :/ Please call our Contact Centre on 02392441234 and we will be able to assist you :) Many thank

Tweet 3
@DespiteOfficial we had a listen last night :) As You Bleed is an amazing track. When are you in Scotland?!
```

Step 4: Tokenization using TweetTokenizer

```
tokenizer = TweetTokenizer(
    preserve_case=False,
    strip_handles=True,
    reduce_len=True
)
tokenized_tweets = [tokenizer.tokenize(tweet) for tweet in tweets[:5]]

for i,tokens in enumerate(tokenized_tweets):
  print("Tweet",i+1,"Tokens:")
  print(tokens)
  print()
```

```
Tweet 1 Tokens:
['#followfriday', 'for', 'being', 'top', 'engaged', 'members', 'in', 'my', 'community', 'this', 'week', ':)']

Tweet 2 Tokens:
['hey', 'james', '!', 'how', 'odd', ':/', 'please', 'call', 'our', 'contact', 'centre', 'on', '02392441234', 'and', 'we', 'w
```

```
Tweet 3 Tokens:
['we', 'had', 'a', 'listen', 'last', 'night', ':)', 'as', 'you', 'bleed', 'is', 'an', 'amazing', 'track', '.', 'when', 'are'

Tweet 4 Tokens:
['congrats', ':)']

Tweet 5 Tokens:
['yeaaah', 'yipppy', '!', '!', '!', 'my', 'accnt', 'verified', 'rqst', 'has', 'succeed', 'got', 'a', 'blue', 'tick', 'mark',
```

Step 5: POS tagging Using NLTK

```
text = "Arjun loves playing cricket 😍 in his free time...  And he also feels sad when he doesn't play cricket😔."

tokens = tokenizer.tokenize(text)

tags = nltk.pos_tag(tokens)

print("Original Text:", text)
print("Tokenized Text:", tokens)
print("POS Tags:",tags)
```

```
Original Text: Arjun loves playing cricket 😍 in his free time...  And he also feels sad when he doesn't play cricket😔.
Tokenized Text: ['Arjun', 'loves', 'playing', 'cricket', '😍', 'in', 'his', 'free', 'time', '...', 'And', 'he', 'also', 'fe
POS Tags: [('Arjun', 'NNP'), ('loves', 'VBZ'), ('playing', 'VBG'), ('cricket', 'NN'), ('😍', 'NN'), ('in', 'IN'), ('his', '
```

Step 6: Extract nouns and verbs

```
nouns = []
verbs = []

for word,tag in tags:
  if tag.startswith('NN'):
    nouns.append(word)
  elif tag.startswith('VB'):
    verbs.append(word)

print("Nouns:",nouns)
print("Verbs:",verbs)
```

```
Nouns: ['Arjun', 'cricket', '😍', 'time', 'cricket', '😔']
Verbs: ['loves', 'playing', 'feels', "doesn't", 'play']
```