

Name : Nithin

Roll No : 2403A52355

Lab : 3

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] Package twitter_samples is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data] /root/nltk_data...
[nltk_data] Package averaged_perceptron_tagger_eng is already up-to-
[nltk_data] date!
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 and Tweet Tokenizer

```
tokenizer = TweetTokenizer(
    preserve_case=False,
    strip_handles=True,
    reduce_len=True
)

tokenized_tweets = [tokenizer.tokenize(tweet) for tweet in tweets[:3]]

for i, tokens in enumerate(tokenized_tweets):
    print(f"Original Tweet {i+1}:")

    print(tweets[i])

    print(f"Tokenized Tweet {i+1} (processed with advanced tokenizer):")

    print(tokens)
    print()
```

```

Original Tweet 1:
#FollowFriday @France_Inte @PKuchly57 @Milipol_Paris for being top engaged members in my community this week :)
Tokenized Tweet 1 (processed with advanced tokenizer):
['#followfriday', 'for', 'being', 'top', 'engaged', 'members', 'in', 'my', 'community', 'this', 'week', ':)']

Original Tweet 2:
@Lamb2ja Hey James! How odd :/ Please call our Contact Centre on 02392441234 and we will be able to assist you :) Many thank
Tokenized Tweet 2 (processed with advanced tokenizer):
['hey', 'james', '!', 'how', 'odd', ':/', 'please', 'call', 'our', 'contact', 'centre', 'on', '02392441234', 'and', 'we', 'w

Original Tweet 3:
@DespiteOfficial we had a listen last night :) As You Bleed is an amazing track. When are you in Scotland?!
Tokenized Tweet 3 (processed with advanced tokenizer):
['we', 'had', 'a', 'listen', 'last', 'night', ':)', 'as', 'you', 'bleed', 'is', 'an', 'amazing', 'track', '.', 'when', 'are'

```

#### step 5 : POS tagging using NLTK

```

text = "Akhil bhAAi 🤔 is playing with bunt 🐾 "

tokenized_text = tokenization.tokenize(text)

tagset = nltk.pos_tag(tokenized_text)

print("original text: ",text)
print("Tokenized text: ",tokenized_text)
print("Pos tags: ",tagset)

```

```

original text: Akhil bhAAi 🤔 is playing with bunt 🐾
Tokenized text: ['Akhil', 'bhAAi', '🤔', 'is', 'playing', 'with', 'bunt', '🐾']
Pos tags: [('Akhil', 'NNP'), ('bhAAi', 'NN'), ('🤔', 'NN'), ('is', 'VBZ'), ('playing', 'VBG'), ('with', 'IN'), ('bunt', 'N'), ('🐾', 'NN')]

```

#### Step - 6 : Extract nouns and verbs

```

nouns = []
verbs = []

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

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

Nouns: ['Akhil', 'bhAAi', '🤔', 'bunt', '🐾']
Verbs: ['is', 'playing']

```