```
!pip install nltk pandas numpy
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.12/dist-packages (3.9.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (2.2.2)
Requirement already satisfied: numpy in /usr/local/lib/python3.12/dist-packages (2.0.2)
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)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/dist-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.3)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2->pandas) (1.
```

```
import pandas as pd
import numpy as np
import re
import nltk
from collections import defaultdict, Counter

nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
```

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

```
df = pd.read_csv('/content/Twitter_Data.csv')  # change filename if needed
df = df[['clean_text']]  # change column name if required
df.dropna(inplace=True)

print(df.head())
```

```
                                          clean_text
0  when modi promised “minimum government maximum...
1  talk all the nonsense and continue all the dra...
2  what did just say vote for modi  welcome bjp t...
3  asking his supporters prefix chowkidar their n...
4  answer who among these the most powerful world...
```

```
def clean_tweet(text):
    text = re.sub(r'http\S+', '', text)       # remove URLs
    text = re.sub(r'@\w+', '', text)           # remove mentions
    text = re.sub(r'[^A-Za-z\s]', '', text)    # remove special chars
    text = text.lower().strip()
    return text

df['category'] = df['clean_text'].apply(clean_tweet)
print(df['category'].head())
```

```
0    when modi promised minimum government maximum ...
1    talk all the nonsense and continue all the dra...
2    what did just say vote for modi  welcome bjp t...
3    asking his supporters prefix chowkidar their n...
4    answer who among these the most powerful world...
Name: category, dtype: object
```

```
nltk.download('punkt_tab')
df['tokens'] = df['category'].apply(nltk.word_tokenize)
print(df['tokens'].head())
```

```
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
0    [when, modi, promised, minimum, government, ma...
1    [talk, all, the, nonsense, and, continue, all,...
2    [what, did, just, say, vote, for, modi, welcom...
3    [asking, his, supporters, prefix, chowkidar, t...
4    [answer, who, among, these, the, most, powerfu...
Name: tokens, dtype: object
```

```
nltk.download('averaged_perceptron_tagger_eng')
df['pos_tags'] = df['tokens'].apply(nltk.pos_tag)
print(df['pos_tags'].head())
```

```
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger_eng.zip.
0    [(when, WRB), (modi, NN), (promised, VBD), (mi...
1    [(talk, NN), (all, PDT), (the, DT), (nonsense,...
2    [(what, WP), (did, VBD), (just, RB), (say, VB)...
3    [(asking, VBG), (his, PRP$), (supporters, NNS)...
4    [(answer, NN), (who, WP), (among, IN), (these,...
Name: pos_tags, dtype: object
```

```
tagged_sentences = df['pos_tags'].tolist()

print(tagged_sentences[0])
```

```
[('when', 'WRB'), ('modi', 'NN'), ('promised', 'VBD'), ('minimum', 'JJ'), ('government', 'NN'), ('maximum', 'JJ'), ('governa
```

```
transition_counts['NN'].most_common(5)
```

```
[('NN', 323492), ('IN', 84326), ('VBD', 54182), ('NNS', 53731), ('JJ', 49954)]
```

```
word_freq = Counter([word for sent in df['tokens'] for word in sent])
rare_words = [w for w, c in word_freq.items() if c == 1]

print("Number of rare words:", len(rare_words))
```

```
Number of rare words: 62190
```

```
test_tweet = "love this movie"
tokens = nltk.word_tokenize(test_tweet)
print(tokens)
```

```
['love', 'this', 'movie']
```