Resources:NLP

Porter Stemmer

A stemmer reduces words to their word stem (e.g., "running" → "run").

```
from nltk.stem import PorterStemmer


stemmer = PorterStemmer()

words = ["running", "jumps", "easily", "flying"]

stems = [stemmer.stem(word) for word in words]

print(stems)
```

Lemmatizer

Lemmatization returns the base or dictionary form of a word.

```
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

words = ["running", "better", "cats"]

lemmas = [lemmatizer.lemmatize(word, pos='v') for word in words]

print(lemmas)
```

Penn Treebank POS Tagging

Penn Treebank defines a tag set for Parts of Speech (POS).

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

sentence = "The quick brown fox jumps over the lazy dog"
tokens = nltk.word_tokenize(sentence)
pos_tags = nltk.pos_tag(tokens)
print(pos_tags)
```

Brill's Tagger

Brill Tagger is a rule-based POS tagger (less commonly used now, but still educational).

```
from nltk.tag import brill, brill_trainer

from nltk.tag import UnigramTagger, DefaultTagger

from nltk.corpus import treebank

nltk.download('treebank')


train_data = treebank.tagged_sents()[:3000]

default_tagger = DefaultTagger('NN')

unigram_tagger = UnigramTagger(train_data, backoff=default_tagger)
```

```
templates = brill.fntbl37()

trainer = brill_trainer.BrillTaggerTrainer(initial_tagger=unigram_tagger, templates=templates, trace=3)

brill_tagger = trainer.train(train_data[:100])

print(brill_tagger.tag(['This', 'is', 'a', 'test']))
```

WordNet

```
from nltk.corpus import wordnet

nltk.download('wordnet')


word = "bank"

synsets = wordnet.synsets(word)

for s in synsets:

    print(f"Synset: {s.name()}, Definition: {s.definition()}, Example: {s.examples()}")
```

PropBank (Proposition Bank)

```
from nltk.corpus import propbank

nltk.download('propbank')

instances = propbank.instances()[:3]

for i in instances:

    print(f"Roleset ID: {i.roleset}, Sentence ID:
{i.fileid}")
```

FrameNet : A rich lexical database that links words to semantic frames.

```
import nltk
nltk.download('framenet_v17')
from nltk.corpus import framenet as fn
```

A rich lexical database that links words to semantic frames.

```
frame = fn.frame_by_name('Commerce_buy')
print(f"Frame name: {frame.name}")
print(f"Definition: {frame.definition}")
print(f"Lexical Units: {[lu.name for lu in frame.lexUnit.values()]}")
```

Brown Corpus : Categorized texts from a wide variety of sources.

```
from nltk.corpus import brown

nltk.download('brown')



print("Categories:", brown.categories())

print("Sample Words (News):",
brown.words(categories='news')[:20])
```

British National Corpus (BNC)

Not bundled with NLTK due to its licensing. You can access parts via BNC XML or BNC Baby, or use the bnc Python package (if locally available).

Accessing and using the British National Corpus (BNC) in Python requires downloading it separately (because it's not bundled with NLTK due to licensing). However, you can still analyze BNC if you have access to either:

- BNC XML Edition (downloaded from http://www.natcorp.ox.ac.uk/)
- BNC Baby – a smaller version for educational use
- Or you can use BNC via the bnc package if you have a local installation or database.