In [ ]:
```
# Akshata NLP_01
```

In [ ]:
```
# Perform tokenization (Whitespace, Punctuation-based, Treebank, Tweet, MWE) using NLTK library.
# Use porter stemmer and snowball stemmer for stemming. Use any technique for lemmatization.
```

In [1]:
```
pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\tech bazaar\appdata\roaming\python\python39\site-packages (3.8.1)
Requirement already satisfied: regex>=2021.8.3 in c:\users\tech bazaar\anaconda3\lib\site-packages (from nltk) (2021.8.3)
Requirement already satisfied: joblib in c:\users\tech bazaar\anaconda3\lib\site-packages (from nltk) (1.1.0)
Requirement already satisfied: click in c:\users\tech bazaar\anaconda3\lib\site-packages (from nltk) (8.0.3)
Requirement already satisfied: tqdm in c:\users\tech bazaar\anaconda3\lib\site-packages (from nltk) (4.62.3)
Requirement already satisfied: colorama in c:\users\tech bazaar\anaconda3\lib\site-packages (from click->nltk) (0.4.6)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
import nltk
```

```
C:\Users\Tech Bazaar\anaconda3\lib\site-packages\scipy\__init__.py:146: UserWarning: A NumPy version >=1.16.5 and <1.23.0 is required for this version
of SciPy (detected version 1.26.2
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

In [3]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to C:\Users\Tech
[nltk_data]     Bazaar\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[3]:
```
True
```

In [4]:
```
text = "India is a unique country with diversity. Unity is the main slogan of the country."
print(text)
```

```
India is a unique country with diversity. Unity is the main slogan of the country.
```

In [5]:
```
# Sentence Tokenization
```

In [6]:
```
from nltk.tokenize import sent_tokenize
print(sent_tokenize(text))
```

```
['India is a unique country with diversity.', 'Unity is the main slogan of the country.']
```

In [7]:
```
# Word Tokenization
```

In [8]:
```
from nltk.tokenize import word_tokenize
print(word_tokenize(text))
```

```
['India', 'is', 'a', 'unique', 'country', 'with', 'diversity', '.', 'Unity', 'is', 'the', 'main', 'slogan', 'of', 'the', 'country', '.']
```

In [9]:
```
# Whitespace Tokenization
```

In [10]:
```
print(f'Whitespace Tokenization= {text.split()}')
```

```
Whitespace Tokenization= ['India', 'is', 'a', 'unique', 'country', 'with', 'diversity.', 'Unity', 'is', 'the', 'main', 'slogan', 'of', 'the', 'countr
y.']
```

In [11]:
```
# Punctuation-Based Tokenization
```

In [12]:
```
from nltk.tokenize import wordpunct_tokenize
print(f'Punctuation-Based Tokenization = {wordpunct_tokenize(text)}')
```

```
Punctuation-Based Tokenization = ['India', 'is', 'a', 'unique', 'country', 'with', 'diversity', '.', 'Unity', 'is', 'the', 'main', 'slogan', 'of', 'th
e', 'country', '.']
```

In [13]:
```
# Default / TreebankWordTokenizer
```

In [14]:
```
sentence = "What's your name?"
from nltk.tokenize import TreebankWordTokenizer
tokenizer = TreebankWordTokenizer()
print(f'Default / TreebankWordTokenizer = {tokenizer.tokenize(sentence)}')
```

```
Default / TreebankWordTokenizer = ['What', "'s", 'your', 'name', '?']
```

In [15]:
```
# Tweet Tokenizer
```

In [16]:
```
pip install emoji --upgrade
```

```
Requirement already satisfied: emoji in c:\users\tech bazaar\anaconda3\lib\site-packages (2.9.0)
Note: you may need to restart the kernel to use updated packages.
```

In [17]:
```
import emoji
```

In [18]:
```
print(emoji.emojize('Hi Everyone !! :grinning_face:'))
```

```
Hi Everyone !! 😀
```

In [19]:
```
sentence1 = emoji.emojize('Hi Everyone !! :grinning_face:')
from nltk.tokenize import TweetTokenizer
tokenizer = TweetTokenizer()
print(f'Tweet-Rule-Based Tokenization = {tokenizer.tokenize(sentence1)}')
```

```
Tweet-Rule-Based Tokenization = ['Hi', 'Everyone', '!', '!', '😀']
```

In [20]:
```
# MWET Tokenizer
```

In [21]:
```
sentence2 = "Hope, is the only thing stronger than fear! Hunger Games"
print(word_tokenize(sentence2))
```

```
['Hope', ',', 'is', 'the', 'only', 'thing', 'stronger', 'than', 'fear', '!', 'Hunger', 'Games']
```

In [22]:
```
from nltk.tokenize import MWETokenizer
tokenizer = MWETokenizer()
tokenizer.add_mwe(('Hunger','Games'))
print(f'Multi-Word expression (MWE) Tokenization = {tokenizer.tokenize(word_tokenize(sentence2))}')
```

```
Multi-Word expression (MWE) Tokenization = ['Hope', ',', 'is', 'the', 'only', 'thing', 'stronger', 'than', 'fear', '!', 'Hunger_Games']
```

In [23]:
```
# STEMMING
```

In [24]:
```
# Porter Stemmer
```

In [25]:
```
from nltk.stem.porter import *
p_stemmer = PorterStemmer()
words = ['run','runner','running','ran','runs','easily','fairly']
for word in words:
    print(word+'-->'+p_stemmer.stem(word))
```

```
run-->run
runner-->runner
running-->run
ran-->ran
runs-->run
easily-->easili
fairly-->fairli
```

In [26]:
```
# Snowball Stemmer
```

In [27]:
```
from nltk.stem.snowball import SnowballStemmer
s_stemmer = SnowballStemmer(language='english')
for word in words:
    print(word+'-->'+s_stemmer.stem(word))
```

```
run-->run
runner-->runner
running-->run
ran-->ran
runs-->run
easily-->easili
fairly-->fair
```

In [28]:
```
# Lemmatization
```

In [29]:
```
!pip3 install spacy
```

```
Requirement already satisfied: spacy in c:\users\tech bazaar\anaconda3\lib\site-packages (3.7.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (2.4.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (3.0.9)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (2.0.10)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (4.62.3)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (2.0.8)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (0.9.0)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (1.0.5)
Requirement already satisfied: packaging>=20.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (21.0)
Requirement already satisfied: weasel<0.4.0,>=0.1.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (0.3.4)
Requirement already satisfied: thinc<8.3.0,>=8.1.8 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (8.2.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (3.3.0)
Requirement already satisfied: setuptools in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (58.0.4)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (2.26.0)
Requirement already satisfied: jinja2 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (2.11.3)
Requirement already satisfied: numpy>=1.19.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (1.26.2)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (1.1.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (3.0.12)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (2.5.3)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\users\tech bazaar\anaconda3\lib\site-packages (from spacy) (6.4.0)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\tech bazaar\anaconda3\lib\site-packages (from packaging>=20.0->spacy) (3.0.4)
Requirement already satisfied: typing-extensions>=4.6.1 in c:\users\tech bazaar\anaconda3\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4
->spacy) (4.9.0)
Requirement already satisfied: pydantic-core==2.14.6 in c:\users\tech bazaar\anaconda3\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->s
pacy) (2.14.6)
Requirement already satisfied: annotated-types>=0.4.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->
spacy) (0.6.0)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy) (2.
0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\tech bazaar\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy) (2020.6.2
0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\tech bazaar\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy) (1.26.
7)
Requirement already satisfied: idna<4,>=2.5 in c:\users\tech bazaar\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy) (3.2)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\users\tech bazaar\anaconda3\lib\site-packages (from thinc<8.3.0,>=8.1.8->spacy) (0.1.4)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\users\tech bazaar\anaconda3\lib\site-packages (from thinc<8.3.0,>=8.1.8->spacy) (0.7.11)
Requirement already satisfied: colorama in c:\users\tech bazaar\anaconda3\lib\site-packages (from tqdm<5.0.0,>=4.38.0->spacy) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\tech bazaar\anaconda3\lib\site-packages (from typer<0.10.0,>=0.3.0->spacy) (8.0.3)
Requirement already satisfied: cloudpathlib<0.17.0,>=0.7.0 in c:\users\tech bazaar\anaconda3\lib\site-packages (from weasel<0.4.0,>=0.1.0->spacy) (0.1
6.0)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\tech bazaar\anaconda3\lib\site-packages (from jinja2->spacy) (1.1.1)
```

In [42]:
```
import spacy
spacy.cli.download('en_core_web_sm')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```

In [43]:
```
nlp = spacy.load('en_core_web_sm')
def show_lemmas(text):
    for token in text:
        print(f'{token.text:{12}} {token.pos_:{6}} {token.lemma:<{22}} {token.lemma_}')
```

In [44]:
```
doc = nlp(u"I am a runner running in a race because I love to run  since I ran today. ")
show_lemmas(doc)
```

```
I            PRON   4690420944186131903    I
am           AUX    10382539506755952630   be
a            DET    11901859001352538922   a
runner       NOUN   12640964157389618806   runner
running      VERB   12767647472892411841   run
in           ADP    3002984154512732771    in
a            DET    11901859001352538922   a
race         NOUN   8048469955494714898    race
because      SCONJ  16950148841647037698   because
I            PRON   4690420944186131903    I
love         VERB   3702023516439754181    love
to           PART   3791531372978436496    to
run          VERB   12767647472892411841   run
             SPACE  8532415787641010193
since        SCONJ  10066841407251338481   since
I            PRON   4690420944186131903    I
ran          VERB   12767647472892411841   run
today        NOUN   11042482332948150395   today
.            PUNCT  12646065887601541794   .
```