Sahil Y. Tike
B212066

1. Tokenization

Tokenization is the process of breaking text into smaller pieces called tokens. These smaller pieces can be sentences, words, or sub-words. For example, the sentence “Iwon” can be tokenized into two word-tokens “I” and “won”.

In [ ]:
```
import nltk
from nltk.tokenize import (TreebankWordTokenizer,
                           word_tokenize,
                           wordpunct_tokenize,
                           TweetTokenizer,
                           MWETokenizer)
```

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

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

Out[8]:
```
True
```

In [ ]:
```
sentence = "Today we would learn about tokenization. Are you all ready?"
```

A. Whitespace tokenization

A WhitespaceTokenizer is a tokenizer that splits on and discards only whitespace characters.

In [ ]:
```

```

In [ ]:
```
print(f'Whitespace tokenization = {sentence.split()}')
```

```
Whitespace tokenization = ['Today', 'we', 'would', 'learn', 'about', 'tokenization.', 'Are', 'you', 'all', 'ready?']
```

B. Punctuation-based tokenization

Punctuation-based tokenization is slightly more advanced than whitespace-based tokenization since it splits on whitespace and punctuations and also retains the punctuations.

In [ ]:
```
print(f'Punctuation-based tokenization = {wordpunct_tokenize(sentence)}')
```

```
Punctuation-based tokenization = ['Today', 'we', 'would', 'learn', 'about',
'tokenization', '.', 'Are', 'you', 'all', 'ready', '?']
```

C. Default/TreebankWordTokenizer

The default tokenization method in NLTK involves tokenization using regular expressions as defined in the Penn Treebank (based on English text). It assumes that the text is already split into sentences.

In [ ]:
```
tokenizer = TreebankWordTokenizer()
print(f'Default/Treebank tokenization = {tokenizer.tokenize(sentence)}')
```

```
Default/Treebank tokenization = ['Today', 'we', 'would', 'learn', 'about', 't
okenization.', 'Are', 'you', 'all', 'ready', '?']
```

D. TweetTokenizer

Special texts, like Twitter tweets, have a characteristic structure and the generic tokenizers mentioned above fail to produce viable tokens when applied to these datasets.

In [ ]:
```
tokenizer = TweetTokenizer()
print(f'Tweet-rules based tokenization = {tokenizer.tokenize(sentence)}')
```

```
Tweet-rules based tokenization = [
                                 'Today', 'we', 'would', 'leanr', 'about',
'tokenization', '.', 'Are', 'you', 'all', 'ready', '?']
```

E. MWETokenizer

The multi-word expression tokenizer is a rule-based, "add-on" tokenizer offered by NLTK. Once the text has been tokenized by a tokenizer of choice, some tokens can be re-groupedinto multi-word expressions.

In [ ]:
```
tokenizer = MWETokenizer()
tokenizer.add_mwe(('Martha', 'Jones'))
print(f'Multi-word expression (MWE) tokenization = {tokenizer.tokenize(word_to
```

```
Multi-word expression (MWE) tokenization = ['Today', 'we', 'would', 'leanr',
'about', 'tokenization', '.                  'all',
                           ', 'Are', 'you',        'ready', '?']
```

2. Stemming and Lemmatization

A. Stemming

Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers. Often when searching text for a certain keyword, it helps if the search returns variations of the word. For instance, searching for "boat" might also return "boats" and "boating". Here, "boat" would be the stem for [boat, boater, boating, boats]. Stemming is a somewhat crude method for cataloging related words; it essentially chops off letters from the end until the stem is reached.

i) Porter Stemmer

In [ ]:
```
# Import the toolkit and the full Porter Stemmer library
import nltk

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
```

ii) Snowball Stemmer

In [ ]:
```
from nltk.stem.snowball import SnowballStemmer

# The Snowball Stemmer requires that you pass a language parameter
s_stemmer = SnowballStemmer(language='english')
words = ['run','runner','running','ran','runs','easily','fairly']
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 this case, the stemmer performed the same as the Porter Stemmer, with the exception that it handled the stem of "fairly" more appropriately with "fair"

B. Lemmatization

In contrast to stemming, lemmatization looks beyond word reduction and considers a language's full vocabulary to apply a morphological analysis to words. The lemma of 'was' is 'be' and the lemma of 'mice' is 'mouse'

Lemmatization is typically seen as much more informative than simple stemming, which is why Spacy has opted to only have Lemmatization available instead of Stemming

I [ ]:

n

```
!pip3 install spacy
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: spacy in /home/dipali/.local/lib/python3.8/sit
e-packages (3.5.0)
Requirement already satisfied: numpy>=1.15.0 in /home/dipali/.local/lib/pytho
n3.8/site-packages (from spacy) (1.24.1)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /home/dipali/.local/l
ib/python3.8/site-packages (from spacy) (3.0.8)
Requirement already satisfied: jinja2 in /home/dipali/.local/lib/python3.8/si
te-packages (from spacy) (3.1.2)
Requirement already satisfied: thinc<8.2.0,>=8.1.0 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (8.1.7)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /home/dipali/.loca
l/lib/python3.8/site-packages (from spacy) (3.3.0)
Requirement already satisfied: packaging>=20.0 in /home/dipali/.local/lib/pyt
hon3.8/site-packages (from spacy) (23.0)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /home/dipali/.loca
l/lib/python3.8/site-packages (from spacy) (6.3.0)
Requirement already satisfied: setuptools in /usr/lib/python3/dist-packages
(from spacy) (45.2.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (4.64.1)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /home/dipali/.l
ocal/lib/python3.8/site-packages (from spacy) (1.0.4)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/lib/python3/di
st-packages (from spacy) (2.22.0)
Requirement already satisfied: pathy>=0.10.0 in /home/dipali/.local/lib/pytho
n3.8/site-packages (from spacy) (0.10.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /home/dipali/.l
ocal/lib/python3.8/site-packages (from spacy) (3.0.12)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /home/dipali/.loca
l/lib/python3.8/site-packages (from spacy) (2.0.8)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (0.7.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (2.0.7)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /home/dipali/.loc
al/lib/python3.8/site-packages (from spacy) (1.0.9)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (1.1.1)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (2.4.5)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in /hom
e/dipali/.local/lib/python3.8/site-packages (from spacy) (1.10.4)
Requirement already satisfied: typing-extensions>=4.2.0 in /home/dipali/.loca
l/lib/python3.8/site-packages (from pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4->sp
acy) (4.4.0)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /home/dipali/.loca
l/lib/python3.8/site-packages (from thinc<8.2.0,>=8.1.0->spacy) (0.0.4)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /home/dipali/.local/lib/
python3.8/site-packages (from thinc<8.2.0,>=8.1.0->spacy) (0.7.9)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /home/dipali/.local/li
b/python3.8/site-packages (from typer<0.8.0,>=0.3.0->spacy) (8.1.3)
```

I [ ]:

In
```
!python -m spacy download en
```

```
/bin/bash: python: command not found
```

In [ ]:
```
# Perform standard imports:
import spacy
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.
```

```
---------------------------------------------------------------------------
OSError                                   Traceback (most recent call last)
Cell In[6], line 3
      1 # Perform standard imports:
      2 import spacy
----> 3 nlp = spacy.load('en_core_web_sm')
      4 def show_lemmas(text):
      5     for token in text:

File ~/.local/lib/python3.8/site-packages/spacy/__init__.py:54, in load(name,
vocab, disable, enable, exclude, config)
     30 def load(
     31     name: Union[str, Path],
     32     *,
   (...)
     37     config: Union[Dict[str, Any], Config] = util.SimpleFrozenDict(),
     38 ) -> Language:
     39     """Load a spaCy model from an installed package or a local path.
     40
     41     name (str): Package name or model path.
   (...)
     52     RETURNS (Language): The loaded nlp object.
     53     """
---> 54     return util.load_model(
     55         name,
     56         vocab=vocab,
     57         disable=disable,
     58         enable=enable,
     59         exclude=exclude,
     60         config=config,
     61     )

File ~/.local/lib/python3.8/site-packages/spacy/util.py:439, in load_model(na
me, vocab, disable, enable, exclude, config)
    437 if name in OLD_MODEL_SHORTCUTS:
    438     raise IOError(Errors.E941.format(name=name, full=OLD_MODEL_SHORTC
UTS[name]))  # type: ignore[index]
--> 439 raise IOError(Errors.E050.format(name=name))

OSError: [E050] Can't find model 'en_core_web_sm'. It doesn't seem to be a Py
thon package or a valid path to a data directory.
```

I [ ]:

n

```
doc = nlp(u"I saw eighteen mice today!")

show_lemmas(doc)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[7], line 1
----> 1 doc = nlp(u"I saw eighteen mice today!")
      3 show_lemmas(doc)

NameError: name 'nlp' is not defined
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: spacy in /home/dipali/.local/lib/python3.8/sit
e-packages (3.5.0)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (0.7.0)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /home/dipali/.loca
l/lib/python3.8/site-packages (from spacy) (6.3.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (4.64.1)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /home/dipali/.loc
al/lib/python3.8/site-packages (from spacy) (1.0.9)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /home/dipali/.l
ocal/lib/python3.8/site-packages (from spacy) (3.0.12)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /home/dipali/.local/l
ib/python3.8/site-packages (from spacy) (3.0.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /home/dipali/.loca
l/lib/python3.8/site-packages (from spacy) (2.0.8)
Requirement already satisfied: pathy>=0.10.0 in /home/dipali/.local/lib/pytho
n3.8/site-packages (from spacy) (0.10.1)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /home/dipali/.loca
l/lib/python3.8/site-packages (from spacy) (3.3.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (2.0.7)
Requirement already satisfied: numpy>=1.15.0 in /home/dipali/.local/lib/pytho
n3.8/site-packages (from spacy) (1.24.1)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /home/dipali/.l
ocal/lib/python3.8/site-packages (from spacy) (1.0.4)
Requirement already satisfied: thinc<8.2.0,>=8.1.0 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (8.1.7)
Requirement already satisfied: setuptools in /usr/lib/python3/dist-packages
(from spacy) (45.2.0)
Requirement already satisfied: jinja2 in /home/dipali/.local/lib/python3.8/si
te-packages (from spacy) (3.1.2)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in /hom
e/dipali/.local/lib/python3.8/site-packages (from spacy) (1.10.4)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (2.4.5)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /home/dipali/.local/li
b/python3.8/site-packages (from spacy) (1.1.1)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/lib/python3/di
st-packages (from spacy) (2.22.0)
Requirement already satisfied: packaging>=20.0 in /home/dipali/.local/lib/pyt
hon3.8/site-packages (from spacy) (23.0)
Requirement already satisfied: typing-extensions>=4.2.0 in /home/dipali/.loca
l/lib/python3.8/site-packages (from pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4->sp
acy) (4.4.0)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /home/dipali/.local/lib/
python3.8/site-packages (from thinc<8.2.0,>=8.1.0->spacy) (0.7.9)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /home/dipali/.loca
l/lib/python3.8/site-packages (from thinc<8.2.0,>=8.1.0->spacy) (0.0.4)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /home/dipali/.local/li
b/python3.8/site-packages (from typer<0.8.0,>=0.3.0->spacy) (8.1.3)
Requirement already satisfied: MarkupSafe>=2.0 in /home/dipali/.local/lib/pyt
hon3.8/site-packages (from jinja2->spacy) (2.1.2)
```