localhost:8888/notebooks/Desktop/PR7.ipynb

jupyter PR7 Last Checkpoint: 5 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

In [1]:
```
import nltk
import string
import re
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk import pos_tag
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [2]:
```
# Download necessary resources
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping taggers\averaged_perceptron_tagger.zip.
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\PC\AppData\Roaming\nltk_data...
```

Out[2]: True

In [3]:
```
document = "Natural Language Processing (NLP) is a sub-field of Artificial Intelligence that focuses on the interaction between
```

In [4]:
```
#tokenization
tokens = word_tokenize(document)
print("Tokens:", tokens)
```

```
Tokens: ['Natural', 'Language', 'Processing', '(', 'NLP', ')', 'is', 'a', 'sub-field', 'of', 'Artificial', 'Intelligence', 'tha
t', 'focuses', 'on', 'the', 'interaction', 'between', 'computers', 'and', 'humans', 'using', 'natural', 'language', '.']
```

localhost:8888/notebooks/Desktop/PR7.ipynb

jupyter PR7 Last Checkpoint: 6 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

In [5]:
```
# POS Tagging
pos_tags = pos_tag(tokens)
print("POS Tags:", pos_tags)
```

```
POS Tags: [('Natural', 'JJ'), ('Language', 'NNP'), ('Processing', 'NNP'), ('(', '('), ('NLP', 'NNP'), (')', ')'), ('is', 'VB
Z'), ('a', 'DT'), ('sub-field', 'NN'), ('of', 'IN'), ('Artificial', 'NNP'), ('Intelligence', 'NNP'), ('that', 'WDT'), ('focuse
s', 'VBZ'), ('on', 'IN'), ('the', 'DT'), ('interaction', 'NN'), ('between', 'IN'), ('computers', 'NNS'), ('and', 'CC'), ('human
s', 'NNS'), ('using', 'VBG'), ('natural', 'JJ'), ('language', 'NN'), ('.', '.')]
```

In [6]:
```
# Stop Words Removal
stop_words = set(stopwords.words('english'))
tokens_no_stopwords = [word for word in tokens if word.lower() not in stop_words]
print("After Stopword Removal:", tokens_no_stopwords)
```

```
After Stopword Removal: ['Natural', 'Language', 'Processing', '(', 'NLP', ')', 'sub-field', 'Artificial', 'Intelligence', 'focu
ses', 'interaction', 'computers', 'humans', 'using', 'natural', 'language', '.']
```

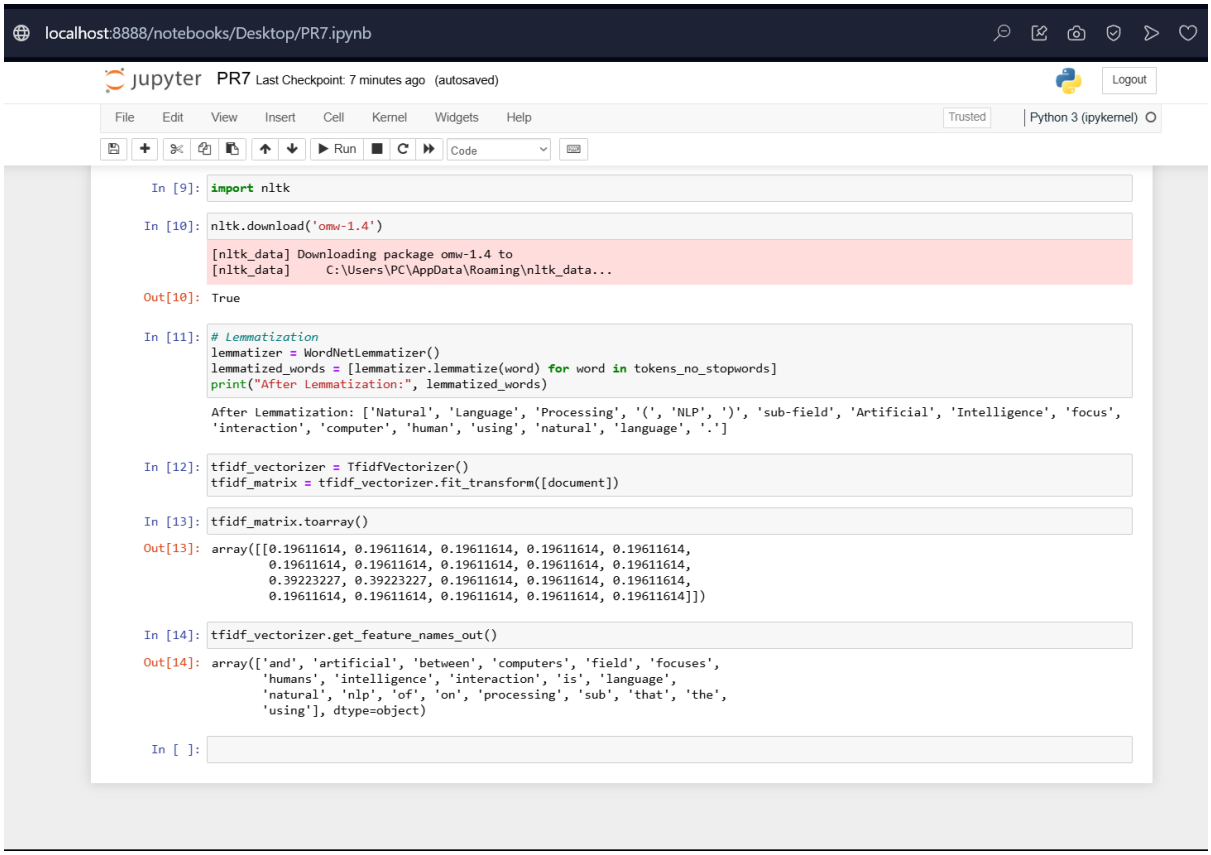
In [7]:
```
# Stemming
stemmer = PorterStemmer()
stemmed_words = [stemmer.stem(word) for word in tokens_no_stopwords]
print("After Stemming:", stemmed_words)
```

```
After Stemming: ['natur', 'languag', 'process', '(', 'nlp', ')', 'sub-field', 'artifici', 'intellig', 'focus', 'interact', 'com
put', 'human', 'use', 'natur', 'languag', '.']
```

In [8]:
```
# Lemmatization
lemmatizer = WordNetLemmatizer()
lemmatized_words = [lemmatizer.lemmatize(word) for word in tokens_no_stopwords]
print("After Lemmatization:", lemmatized_words)
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
C:\ProgramData\Anaconda3\lib\site-packages\nltk\corpus\util.py in __load(self)
     83                 try:
---> 84                     root = nltk.data.find(f"{self.subdir}/{zip_name}")
     85                 except LookupError:

C:\ProgramData\Anaconda3\lib\site-packages\nltk\data.py in find(resource_name, paths)
    582     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 583     raise LookupError(resource_not_found)
    584
```


localhost:8888/notebooks/Desktop/PR7.ipynb
jupyter PR7 Last Checkpoint: 7 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [9]: import nltk
In [10]: nltk.download('omw-1.4')
[nltk_data] Downloading package omw-1.4 to
[nltk_data] C:\Users\PC\AppData\Roaming\nltk_data...
Out[10]: True
In [11]: # Lemmatization
lemmatizer = WordNetLemmatizer()
lemmatized_words = [lemmatizer.lemmatize(word) for word in tokens_no_stopwords]
print("After Lemmatization:", lemmatized_words)
After Lemmatization: ['Natural', 'Language', 'Processing', '(', 'NLP', ')', 'sub-field', 'Artificial', 'Intelligence', 'focus',
'interaction', 'computer', 'human', 'using', 'natural', 'language', '.']
In [12]: tfidf_vectorizer = TfidfVectorizer()
tfidf_matrix = tfidf_vectorizer.fit_transform([document])
In [13]: tfidf_matrix.toarray()
Out[13]: array([[0.19611614, 0.19611614, 0.19611614, 0.19611614, 0.19611614,
0.19611614, 0.19611614, 0.19611614, 0.19611614, 0.19611614,
0.39223227, 0.39223227, 0.19611614, 0.19611614, 0.19611614,
0.19611614, 0.19611614, 0.19611614, 0.19611614, 0.19611614]])
In [14]: tfidf_vectorizer.get_feature_names_out()
Out[14]: array(['and', 'artificial', 'between', 'computers', 'field', 'focuses',
'humans', 'intelligence', 'interaction', 'is', 'language',
'natural', 'nlp', 'of', 'on', 'processing', 'sub', 'that', 'the',
'using'], dtype=object)
In [ ]: