In [2]:

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from nltk.tokenize import sent_tokenize
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
```

In [3]:

```
stopWords = stopwords.words('english')
print(stopWords)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're
e", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves',
'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'i
t', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselve
s', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'tho
se', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has',
'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'bu
t', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for',
'with', 'about', 'against', 'between', 'into', 'through', 'during', 'befor
e', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'o
n', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'the
re', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'mo
re', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'sa
me', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "d
on't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y',
'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "d
oesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "is
n't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 's
han', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "were
n't", 'won', "won't", 'wouldn', "wouldn't"]
```

In [4]:

```
text1 = "Fifty people were killed and 50 others wounded in a terror attack on two mosques i
```

In [5]:

```
text2 = "The number of people killed in Friday’s massacre in Christchurch rose to 50 when a
```

In [6]:

```
wordnet_lemmatizer = WordNetLemmatizer()
def normalize(sentence):
    l1=list()
    sentence_words = nltk.word_tokenize(sentence)
    #print(sentence_words)
    for word in sentence_words:
        #print(word)
        #print(wordnet_lemmatizer.lemmatize(word))
        l1.append(wordnet_lemmatizer.lemmatize(word, pos="v"))
        print(l1)
    return l1
```

In [7]:

```
vectorizer = TfidfVectorizer(tokenizer=normalize)
def cosine_sim(text1, text2):
    tfidf = vectorizer.fit_transform([text1, text2])
    return ((tfidf * tfidf.T).A)[0,1]
```

In [8]:

```
cosine_sim(text1,text2)
```

```
['fifty']
['fifty', 'people']
['fifty', 'people', 'be']
['fifty', 'people', 'be', 'kill']
['fifty', 'people', 'be', 'kill', 'and']
['fifty', 'people', 'be', 'kill', 'and', '50']
['fifty', 'people', 'be', 'kill', 'and', '50', 'others']
['fifty', 'people', 'be', 'kill', 'and', '50', 'others', 'wound']
['fifty', 'people', 'be', 'kill', 'and', '50', 'others', 'wound', 'in']
['fifty', 'people', 'be', 'kill', 'and', '50', 'others', 'wound', 'in',
'a']
['fifty', 'people', 'be', 'kill', 'and', '50', 'others', 'wound', 'in',
'a', 'terror']
['fifty', 'people', 'be', 'kill', 'and', '50', 'others', 'wound', 'in',
'a', 'terror', 'attack']
['fifty', 'people', 'be', 'kill', 'and', '50', 'others', 'wound', 'in',
'a', 'terror', 'attack', 'on']
['fifty', 'people', 'be', 'kill', 'and', '50', 'others', 'wound', 'in',
'a', 'terror', 'attack', 'on', 'two']
```

In [9]:

```
textt1 = "This Agreement is governed by the laws of the State of Missouri without reference
```

In [10]:

```
textt2 = "The laws of the State of Missouri shall apply to this Agreement."
```

In [11]:

```
cosine_sim(textt1,textt2)
```

```
['this']
['this', 'agreement']
['this', 'agreement', 'be']
['this', 'agreement', 'be', 'govern']
['this', 'agreement', 'be', 'govern', 'by']
['this', 'agreement', 'be', 'govern', 'by', 'the']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without', 'reference']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without', 'reference', 'to']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without', 'reference', 'to', 'its']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without', 'reference', 'to', 'its', 'conflict']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without', 'reference', 'to', 'its', 'conflict', 'o
f']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without', 'reference', 'to', 'its', 'conflict', 'o
f', 'law']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without', 'reference', 'to', 'its', 'conflict', 'o
f', 'law', 'principles']
['this', 'agreement', 'be', 'govern', 'by', 'the', 'laws', 'of', 'the', 'sta
te', 'of', 'missouri', 'without', 'reference', 'to', 'its', 'conflict', 'o
f', 'law', 'principles', '.']
['the']
['the', 'laws']
['the', 'laws', 'of']
['the', 'laws', 'of', 'the']
['the', 'laws', 'of', 'the', 'state']
['the', 'laws', 'of', 'the', 'state', 'of']
['the', 'laws', 'of', 'the', 'state', 'of', 'missouri']
['the', 'laws', 'of', 'the', 'state', 'of', 'missouri', 'shall']
['the', 'laws', 'of', 'the', 'state', 'of', 'missouri', 'shall', 'apply']
['the', 'laws', 'of', 'the', 'state', 'of', 'missouri', 'shall', 'apply', 't
o']
['the', 'laws', 'of', 'the', 'state', 'of', 'missouri', 'shall', 'apply', 't
o', 'this']
['the', 'laws', 'of', 'the', 'state', 'of', 'missouri', 'shall', 'apply', 't
o', 'this', 'agreement']
['the', 'laws', 'of', 'the', 'state', 'of', 'missouri', 'shall', 'apply', 't
o', 'this', 'agreement', '.']
```

Out[11]:

```
0.6353562124318741
```

In [12]:

```
testttt1 = "I ate an apple yesterday."
testttt2 = "I am eating an apple right now."
```

In [15]:

```
cosine_sim(testttt1,testttt2)
```

```
['i']
['i', 'eat']
['i', 'eat', 'an']
['i', 'eat', 'an', 'apple']
['i', 'eat', 'an', 'apple', 'yesterday']
['i', 'eat', 'an', 'apple', 'yesterday', '.']
['i']
['i', 'be']
['i', 'be', 'eat']
['i', 'be', 'eat', 'an']
['i', 'be', 'eat', 'an', 'apple']
['i', 'be', 'eat', 'an', 'apple', 'right']
['i', 'be', 'eat', 'an', 'apple', 'right', 'now']
['i', 'be', 'eat', 'an', 'apple', 'right', 'now', '.']
```

Out[15]:

```
0.5727393584196199
```

In [ ]: