In [7]:

```
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.corpus import stopwords
import nltk
nltk.download('stopwords')

# Bring in standard stopwords
stopWords = stopwords.words('english')

print ("\nCalculating WhatsApp Group similarity scores ->")

# Open and read a bunch of files
f = open('CSE.txt')
doc1 = str(f.read())

f = open('GGB.txt')
doc2 = str(f.read())

# Create a string to use to test the similarity scoring
train_string = 'The'

# Construct the training set as a list
train_set = [ doc1, doc2]

# Set up the vectoriser, passing in the stop words
tfidf_vectorizer = TfidfVectorizer(stop_words=stopWords)

# Apply the vectoriser to the training set
tfidf_matrix_train = tfidf_vectorizer.fit_transform(train_set)

# Print the score
print ("\n The Similarity Score is -> [*] ",cosine_similarity(tfidf_matrix_train
[0:1], tfidf_matrix_train))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/rushilmehtani/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!

Calculating WhatsApp Group similarity scores ->

 The Similarity Score is -> [*]  [[1.         0.87343953]]
```

In [ ]: