```
D1 = "I enjoy studying data science"
D2 = "I enjoy studying artificial intelligence"
D3 = "I love playing football"
D4 = "I love playing basketball"
D5 = "Data science involves statistics"
D6 = "Artificial intelligence involves neural networks"
D7 = "Football is an outdoor sport"
D8 = "Basketball is an indoor sport"
```

```
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd
documents = [D1, D2, D3, D4, D4, D5, D6, D7, D8]
vectorizer = CountVectorizer(max_df=0.95, min_df=1, stop_words="english")
bow = vectorizer.fit_transform(documents)
feature_names = vectorizer.get_feature_names_out()
df_bow = pd.DataFrame(bow.toarray(), columns=feature_names)
print("Bag of Words (TF-IDF) Representation:")
print(df_bow)
```

```
Bag of Words (TF-IDF) Representation:
   artificial  basketball  data  enjoy  football  indoor  intelligence  \
0           0           0     1      1         0       0             0
1           1           0     0      1         0       0             1
2           0           0     0      0         1       0             0
3           0           1     0      0         0       0             0
4           0           1     0      0         0       0             0
5           0           0     1      0         0       0             0
6           1           0     0      0         0       0             1
7           0           0     0      0         1       0             0
8           0           1     0      0         0       1             0

   involves  love  networks  neural  outdoor  playing  science  sport  \
0         0     0         0       0        0        0        1      0
1         0     0         0       0        0        0        0      0
2         0     1         0       0        0        1        0      0
3         0     1         0       0        0        1        0      0
4         0     1         0       0        0        1        0      0
5         1     0         0       0        0        0        1      0
6         1     0         1       1        0        0        0      0
7         0     0         0       0        1        0        0      1
8         0     0         0       0        0        0        0      1

   statistics  studying
0           0         1
1           0         1
2           0         0
3           0         0
4           0         0
5           1         0
6           0         0
7           0         0
8           0         0
```

```
from sklearn.metrics.pairwise import cosine_similarity
import pandas as pd
cosine_sim_matrix = cosine_similarity(df_bow)
df_cosine_sim = pd.DataFrame(cosine_sim_matrix, index=documents, columns=documents)
print("Cosine Similarity Matrix:")
print(df_cosine_sim)
```

```
Cosine Similarity Matrix:
                                                    I enjoy studying data science  \
I enjoy studying data science                                                 1.0
I enjoy studying artificial intelligence                                      0.5
I love playing football                                                       0.0
I love playing basketball                                                     0.0
I love playing basketball                                                     0.0
Data science involves statistics                                              0.5
Artificial intelligence involves neural networks                              0.0
Football is an outdoor sport                                                  0.0
Basketball is an indoor sport                                                 0.0

                                                    I enjoy studying artificial intelligence  \
I enjoy studying data science                                                       0.500000
I enjoy studying artificial intelligence                                            1.000000
I love playing football                                                             0.000000
I love playing basketball                                                           0.000000
I love playing basketball                                                           0.000000
Data science involves statistics                                                    0.000000
Artificial intelligence involves neural networks                                    0.447214
Football is an outdoor sport                                                        0.000000
Basketball is an indoor sport                                                       0.000000

                                                    I love playing football  \
```

```
I enjoy studying data science                                0.000000
I enjoy studying artificial intelligence                     0.000000
I love playing football                                      1.000000
I love playing basketball                                    0.666667
I love playing basketball                                    0.666667
Data science involves statistics                             0.000000
Artificial intelligence involves neural networks             0.000000
Football is an outdoor sport                                 0.333333
Basketball is an indoor sport                                0.000000

                                                    I love playing basketball  \
I enjoy studying data science                                        0.000000
I enjoy studying artificial intelligence                             0.000000
I love playing football                                              0.666667
I love playing basketball                                            1.000000
I love playing basketball                                            1.000000
Data science involves statistics                                     0.000000
Artificial intelligence involves neural networks                     0.000000
Football is an outdoor sport                                         0.000000
Basketball is an indoor sport                                        0.333333

                                                    I love playing basketball  \
I enjoy studying data science                                        0.000000
I enjoy studying artificial intelligence                             0.000000
I love playing football                                              0.666667
I love playing basketball                                            1.000000
I love playing basketball                                            1.000000
Data science involves statistics                                     0.000000
Artificial intelligence involves neural networks                     0.000000
Football is an outdoor sport                                         0.000000
Basketball is an indoor sport                                        0.333333

                                                    Data science involves statistics  \
I enjoy studying data science                                                0.500000
```

```
from sklearn.metrics import jaccard_score
import numpy as np
import pandas as pd
binary_bow = (df_bow > 0).astype(int)
num_documents = binary_bow.shape[0]
jaccard_sim_matrix = np.zeros((num_documents, num_documents))
for i in range(num_documents):
    for j in range(num_documents):
        vec_i = binary_bow.iloc[i]
        vec_j = binary_bow.iloc[j]
        intersection = np.sum(np.logical_and(vec_i, vec_j))
        union = np.sum(np.logical_or(vec_i, vec_j))
        if union == 0:
            jaccard_sim_matrix[i, j] = 0.0
        else:
            jaccard_sim_matrix[i, j] = intersection / union
df_jaccard_sim = pd.DataFrame(jaccard_sim_matrix, index=documents, columns=documents)
print("Jaccard Similarity Matrix:")
print(df_jaccard_sim)
```

```
Jaccard Similarity Matrix:
                                                    I enjoy studying data science  \
I enjoy studying data science                                        1.000000
I enjoy studying artificial intelligence                             0.333333
I love playing football                                              0.000000
I love playing basketball                                            0.000000
I love playing basketball                                            0.000000
Data science involves statistics                                     0.333333
Artificial intelligence involves neural networks                     0.000000
Football is an outdoor sport                                         0.000000
Basketball is an indoor sport                                        0.000000

                                                    I enjoy studying artificial intelligence  \
I enjoy studying data science                                                       0.333333
I enjoy studying artificial intelligence                                            1.000000
I love playing football                                                             0.000000
I love playing basketball                                                           0.000000
I love playing basketball                                                           0.000000
Data science involves statistics                                                    0.000000
Artificial intelligence involves neural networks                                    0.285714
Football is an outdoor sport                                                        0.000000
Basketball is an indoor sport                                                       0.000000

                                                    I love playing football  \
I enjoy studying data science                                           0.0
I enjoy studying artificial intelligence                                0.0
I love playing football                                                 1.0
I love playing basketball                                               0.5
I love playing basketball                                               0.5
Data science involves statistics                                        0.0
Artificial intelligence involves neural networks                        0.0
Football is an outdoor sport                                            0.2
Basketball is an indoor sport                                           0.0
```

```
                                                I love playing basketball  \
I enjoy studying data science                                         0.0
I enjoy studying artificial intelligence                              0.0
I love playing football                                               0.5
I love playing basketball                                             1.0
I love playing basketball                                             1.0
Data science involves statistics                                      0.0
Artificial intelligence involves neural networks                      0.0
Football is an outdoor sport                                          0.0
Basketball is an indoor sport                                         0.2

                                                I love playing basketball  \
I enjoy studying data science                                         0.0
I enjoy studying artificial intelligence                              0.0
I love playing football                                               0.5
I love playing basketball                                             1.0
I love playing basketball                                             1.0
Data science involves statistics                                      0.0
Artificial intelligence involves neural networks                      0.0
Football is an outdoor sport                                          0.0
Basketball is an indoor sport                                         0.2

                                                Data science involves statistics  \
```

```
import nltk
from nltk.corpus import wordnet
from nltk.tokenize import word_tokenize
import numpy as np
import pandas as pd
try:
    nltk.data.find('corpora/wordnet')
except LookupError:
    nltk.download('wordnet')
try:
    nltk.data.find('tokenizers/punkt_tab')
except LookupError:
    nltk.download('punkt_tab')
def document_wordnet_similarity(doc1, doc2):
    tokens1 = word_tokenize(doc1.lower())
    tokens2 = word_tokenize(doc2.lower())
    synsets1 = [s for token in tokens1 for s in wordnet.synsets(token)]
    synsets2 = [s for token in tokens2 for s in wordnet.synsets(token)]
    if not synsets1 or not synsets2:
        return 0.0
    max_similarities = []
    for s1 in synsets1:
        max_sim_for_s1 = 0.0
        for s2 in synsets2:
            sim = s1.path_similarity(s2)
            if sim is not None and sim > max_sim_for_s1:
                max_sim_for_s1 = sim
        max_similarities.append(max_sim_for_s1)
    if max_similarities:
        return np.mean(max_similarities)
    else:
        return 0.0
documents = [D1, D2, D3, D4, D5, D6, D7, D8]
num_documents = len(documents)
wordnet_sim_matrix = np.zeros((num_documents, num_documents))
for i in range(num_documents):
    for j in range(num_documents):
        if i == j:
            wordnet_sim_matrix[i, j] = 1.0
        else:
            wordnet_sim_matrix[i, j] = document_wordnet_similarity(documents[i], documents[j])
df_wordnet_sim = pd.DataFrame(wordnet_sim_matrix, index=documents, columns=documents)
print("WordNet Similarity Matrix (Path Similarity):")
print(df_wordnet_sim)
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
WordNet Similarity Matrix (Path Similarity):
                                                I enjoy studying data science  \
I enjoy studying data science                                       1.000000
I enjoy studying artificial intelligence                            0.742014
I love playing football                                             0.284466
I love playing basketball                                           0.284466
Data science involves statistics                                    0.484707
Artificial intelligence involves neural networks                    0.208868
Football is an outdoor sport                                        0.219211
Basketball is an indoor sport                                       0.219211

                                                I enjoy studying artificial intelligence  \
```

```
I enjoy studying data science                                       0.843810
I enjoy studying artificial intelligence                            1.000000
I love playing football                                             0.284193
I love playing basketball                                           0.284193
Data science involves statistics                                    0.207234
Artificial intelligence involves neural networks                    0.478071
Football is an outdoor sport                                        0.227366
Basketball is an indoor sport                                       0.227366

                                                    I love playing football  \
I enjoy studying data science                                      0.435556
I enjoy studying artificial intelligence                           0.404977
I love playing football                                            1.000000
I love playing basketball                                          0.972222
Data science involves statistics                                   0.199542
Artificial intelligence involves neural networks                   0.214044
Football is an outdoor sport                                       0.324458
Basketball is an indoor sport                                      0.268902

                                                    I love playing basketball  \
I enjoy studying data science                                        0.435556
I enjoy studying artificial intelligence                             0.404977
I love playing football                                              0.972222
I love playing basketball                                            1.000000
Data science involves statistics                                     0.199542
Artificial intelligence involves neural networks                     0.214044
Football is an outdoor sport                                         0.265816
Basketball is an indoor sport                                        0.321371

                                                    Data science involves statistics  \
I enjoy studying data science                                             0.394129
I enjoy studying artificial intelligence                                  0.237121
I love playing football                                                   0.204256
I love playing basketball                                                 0.204256
Data science involves statistics                                          1.000000
Artificial intelligence involves neural networks                          0.445773
Football is an outdoor sport                                              0.221549
Basketball is an indoor sport                                             0.221549

                                                    Artificial intelligence involves neural networks  \
I enjoy studying data science                                                                 0.238393
I enjoy studying artificial intelligence                                                      0.495486
```

```
Start coding or generate with AI.
```