cosinesimilarity

September 30, 2024

[ ]:

```
# This Python 3 environment comes with many helpful analytics libraries␣
 ↪installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/
 ↪docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list␣
 ↪all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that␣
 ↪gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved␣
 ↪outside of the current session
```

[2]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# Function to read text from a file
def read_file(filename):
    with open(filename, 'r', encoding='utf-8') as file:
        return file.read()

# Function to compute cosine similarity between two documents
def compute_cosine_similarity(doc1, doc2):
    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform([doc1, doc2])
    return cosine_similarity(tfidf_matrix[0:1], tfidf_matrix[1:2])[0][0]
```

```python
# Input text files
file1 = '/kaggle/input/documentcompare/document1.txt'
file2 = '/kaggle/input/documentcompare/document2.txt'

# Read documents
doc1 = read_file(file1)
doc2 = read_file(file2)

# Compute cosine similarity
similarity = compute_cosine_similarity(doc1, doc2)

# Output the result
print(f"Cosine Similarity: {similarity:.4f}")
```

```
Cosine Similarity: 0.4119
```

[3]:

```python
# Function to read text from a file
def read_file(filename):
    with open(filename, 'r', encoding='utf-8') as file:
        return file.read()


# Read and display documents
doc1 = read_file(file1)
doc2 = read_file(file2)

print("Content of Document 1:\n")
print(doc1)
print("\nContent of Document 2:\n")
print(doc2)
```

```
Content of Document 1:

Artificial intelligence and machine learning are transforming industries. AI is
enabling computers to perform tasks that were once thought to be exclusively
human capabilities, such as understanding language, recognizing patterns, and
making decisions. The future of AI is vast, with possibilities ranging from
healthcare to autonomous vehicles.


Content of Document 2:

Machine learning and artificial intelligence are changing the way industries
operate. AI allows machines to understand patterns, make decisions, and even
process language, which were once human-only abilities. The applications of AI
are extensive, from self-driving cars to advancements in healthcare.
```