Plagirism Checker

File Edit View Insert Runtime Tools Help Last edited on June 28

Comment Share S

+ Code + Text Connect

```
import os
from numpy import vectorize
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
```

```
[ ] %%writefile test1.txt
    "The theory of relativity revolutionized our understanding of space and time."
```

```
Writing test1.txt
```

```
[ ] %%writefile test2.txt
    "The Mona Lisa is one of Leonardo da Vinci's most famous paintings."
```

```
Writing test2.txt
```

```
[ ] %%writefile test3.txt
    "Shakespeare's plays are considered masterpieces of English literature."
```

```
Writing test3.txt
```

```
[ ] %%writefile test4.txt
     "The discovery of penicillin revolutionized the field of medicine."
```

```
Writing test4.txt
```

```
[ ] sample_files = [doc for doc in os.listdir() if doc.endswith('.txt')]
    sample_contents = [open(File).read() for File in sample_files]
```

```
[ ] vectorize = lambda Text: TfidfVectorizer().fit_transform(Text).toarray()
```

```
[ ]  similarity = lambda doc1, doc2: cosine_similarity([doc1, doc2])
```

```
[ ]  vectors = vectorize(sample_contents)
     s_vectors = list(zip(sample_files, vectors))
```

```
[ ]  def check_plagiarism():
         results = set()
         global s_vectors
         for sample_a, text_vector_a in s_vectors:
             new_vectors = s_vectors.copy()
             current_index = new_vectors.index((sample_a, text_vector_a))
             del new_vectors[current_index] #remove compared sample,vector
             for sample_b, text_vector_b in new_vectors:      #compare rest sample,vector
                 sim_score = similarity(text_vector_a, text_vector_b)[0][1]
                 sample_pair = sorted((sample_a, sample_b))
                 score = sample_pair[0], sample_pair[1], sim_score
                 results.add(score)
         return results
```

```
[ ]  for sample in sample_files,sample_contents:
           print(sample)
           print("**********************************************************")
```

```
['test2.txt', 'test3.txt', 'test1.txt', 'test4.txt']
**********************************************************
['"The Mona Lisa is one of Leonardo da Vinci\'s most famous paintings."\n', '"Shakespeare\'s plays are considered masterpieces of English literature."\n', '"The theory of relativity re
**********************************************************
```

```
for sample in sample_files,sample_contents:
      print(sample)
      print("**********************************************************")
```

```
['test2.txt', 'test3.txt', 'test1.txt', 'test4.txt']
**********************************************************
['"The Mona Lisa is one of Leonardo da Vinci\'s most famous paintings."\n', '"Shakespeare\'s plays are considered masterpieces of English literature."\n', '"The theory of relativity re
**********************************************************
```

```
for data in check_plagiarism():
    print(data)
```

```
('test2.txt', 'test4.txt', 0.15354035649944353)
('test2.txt', 'test3.txt', 0.030900153805878076)
('test1.txt', 'test2.txt', 0.09647661856692721)
('test3.txt', 'test4.txt', 0.07454314056274658)
('test1.txt', 'test4.txt', 0.3087167251078705)
('test1.txt', 'test3.txt', 0.066882786139585)
```