Neeraj Appari T073

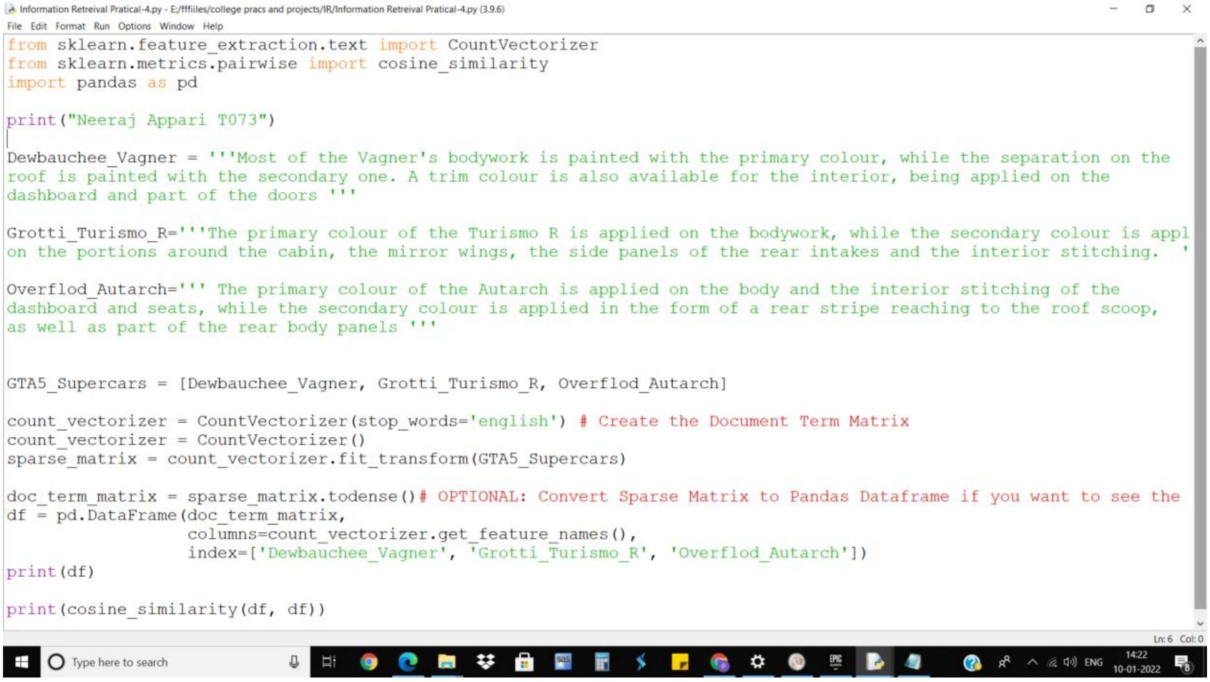
Aim: Write a program to Compute Similarity between two text documents. (Use Cosine Similarity)

Description:

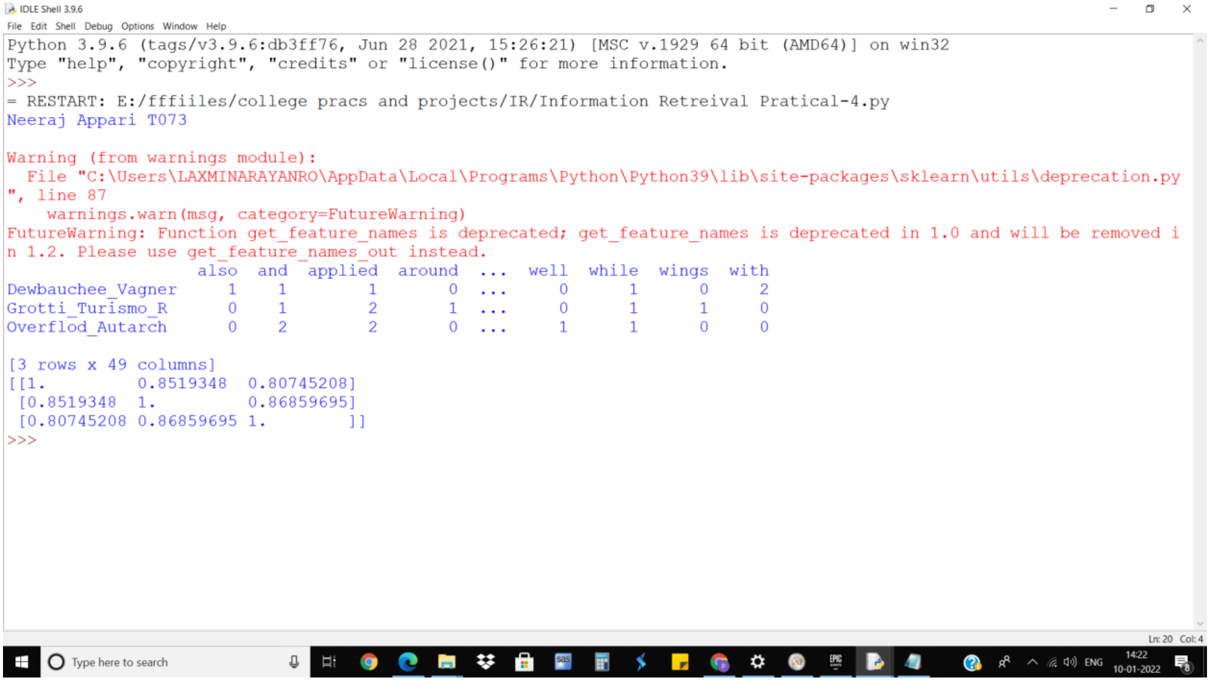
1)Cosine similarity- Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space.

2)Python library (if used)-

The CountVectorizer or the TfidfVectorizer from scikit learn lets user compute Cosine Similarity from sklearn. The output of this comes as a sparse_matrix.


Information Retreival Pratical-4.py - E:/fffiiles/college pracs and projects/IR/Information Retreival Pratical-4.py (3.9.6)
File Edit Format Run Options Window Help
Ln: 6 Col: 0
Type here to search
ENG
14:22
10-01-2022

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import pandas as pd

print("Neeraj Appari T073")

Dewbauchee_Vagner = '''Most of the Vagner's bodywork is painted with the primary colour, while the separation on the
roof is painted with the secondary one. A trim colour is also available for the interior, being applied on the
dashboard and part of the doors '''

Grotti_Turismo_R='''The primary colour of the Turismo R is applied on the bodywork, while the secondary colour is appl
on the portions around the cabin, the mirror wings, the side panels of the rear intakes and the interior stitching.  '

Overflod_Autarch=''' The primary colour of the Autarch is applied on the body and the interior stitching of the
dashboard and seats, while the secondary colour is applied in the form of a rear stripe reaching to the roof scoop,
as well as part of the rear body panels '''


GTA5_Supercars = [Dewbauchee_Vagner, Grotti_Turismo_R, Overflod_Autarch]

count_vectorizer = CountVectorizer(stop_words='english') # Create the Document Term Matrix
count_vectorizer = CountVectorizer()
sparse_matrix = count_vectorizer.fit_transform(GTA5_Supercars)

doc_term_matrix = sparse_matrix.todense()# OPTIONAL: Convert Sparse Matrix to Pandas Dataframe if you want to see the
df = pd.DataFrame(doc_term_matrix,
                  columns=count_vectorizer.get_feature_names(),
                  index=['Dewbauchee_Vagner', 'Grotti_Turismo_R', 'Overflod_Autarch'])
print(df)

print(cosine_similarity(df, df))
```



```
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: E:/fffiiles/college pracs and projects/IR/Information Retreival Pratical-4.py
Neeraj Appari T073

Warning (from warnings module):
  File "C:\Users\LAXMINARAYANRO\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\deprecation.py
", line 87
    warnings.warn(msg, category=FutureWarning)
FutureWarning: Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will be removed i
n 1.2. Please use get_feature_names_out instead.
                   also  and  applied  around  ...  well  while  wings  with
Dewbauchee_Vagner     1    1        1       0  ...     0      1      0     2
Grotti_Turismo_R      0    1        2       1  ...     0      1      1     0
Overflod_Autarch      0    2        2       0  ...     1      1      0     0

[3 rows x 49 columns]
[[1.         0.8519348  0.80745208]
 [0.8519348  1.         0.86859695]
 [0.80745208 0.86859695 1.        ]]
>>>
```