Day 69

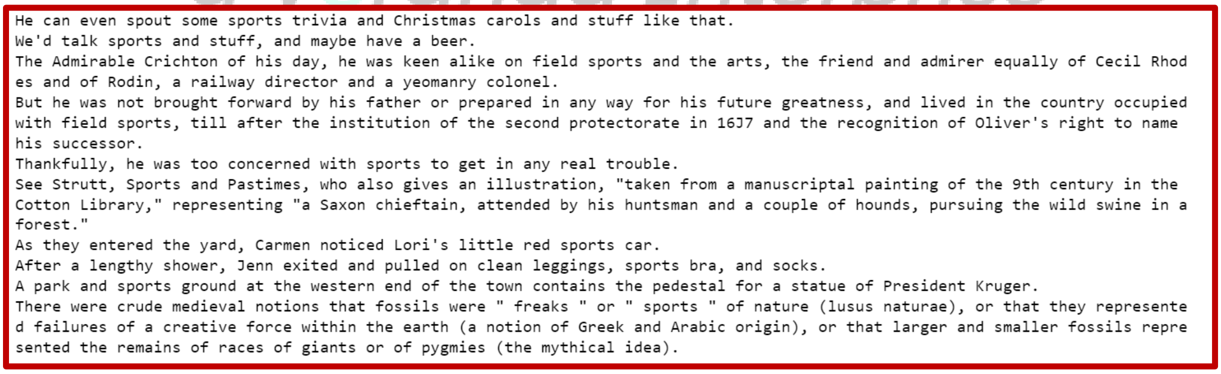
DIY

Q1. Problem Statement: Feature Extraction

Write a Python program that reads the *demotext2.txt* text file (provided on LMS), and then the following are the tasks that are to be taken into consideration while constructing the solution.

1. Load the *demotext2.txt* text file into a variable and then close the file
2. Do sentence wise tokenization and list out generated tokens
3. Transform each token into a lower case
4. Do vectorization using TFID Vectorizer
5. Generate vector-matrix

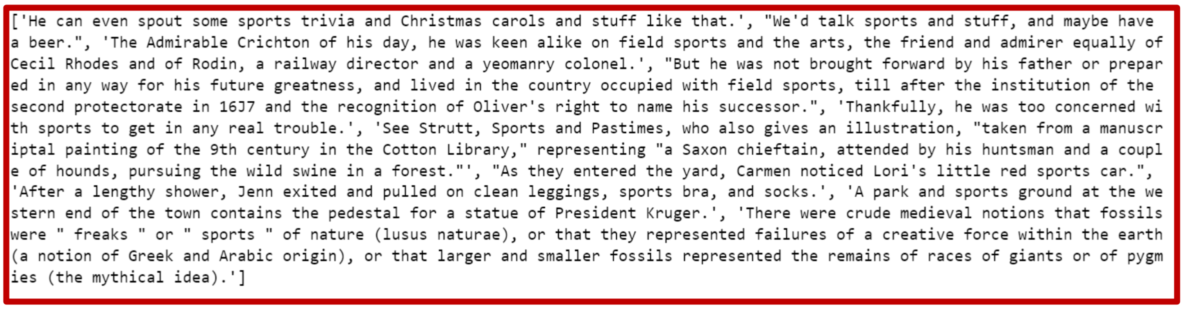
Input Format:



```
He can even spout some sports trivia and Christmas carols and stuff like that.
We'd talk sports and stuff, and maybe have a beer.
The Admirable Crichton of his day, he was keen alike on field sports and the arts, the friend and admirer equally of Cecil Rhod
es and of Rodin, a railway director and a yeomanry colonel.
But he was not brought forward by his father or prepared in any way for his future greatness, and lived in the country occupied
with field sports, till after the institution of the second protectorate in 16J7 and the recognition of Oliver's right to name
his successor.
Thankfully, he was too concerned with sports to get in any real trouble.
See Strutt, Sports and Pastimes, who also gives an illustration, "taken from a manuscriptal painting of the 9th century in the
Cotton Library," representing "a Saxon chieftain, attended by his huntsman and a couple of hounds, pursuing the wild swine in a
forest."
As they entered the yard, Carmen noticed Lori's little red sports car.
After a lengthy shower, Jenn exited and pulled on clean leggings, sports bra, and socks.
A park and sports ground at the western end of the town contains the pedestal for a statue of President Kruger.
There were crude medieval notions that fossils were " freaks " or " sports " of nature (lusus naturae), or that they represente
d failures of a creative force within the earth (a notion of Greek and Arabic origin), or that larger and smaller fossils repre
sented the remains of races of giants or of pygmies (the mythical idea).
```

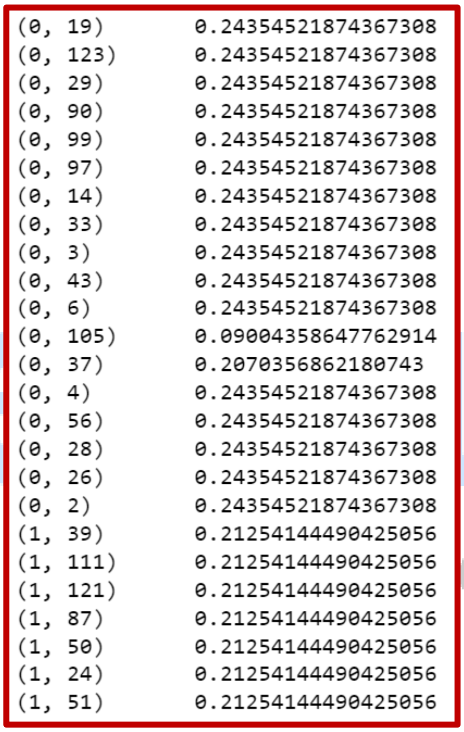
Sample Output:

2. Do sentence wise tokenization and list out generated tokens



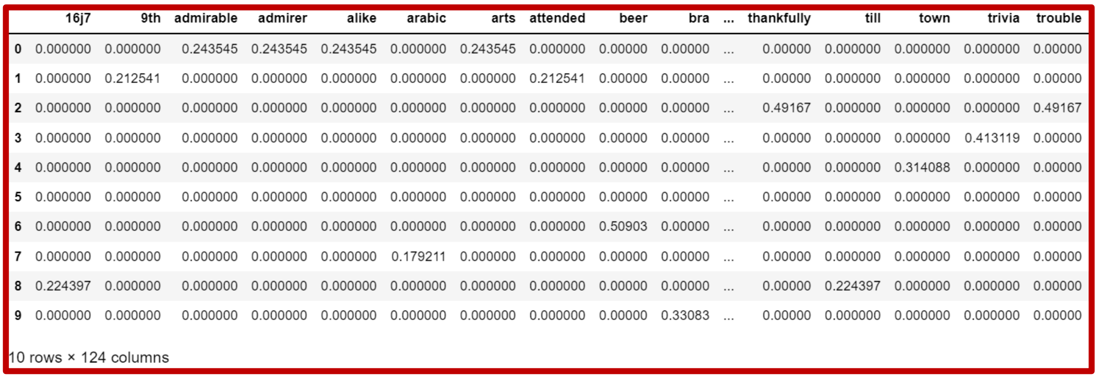
```
['He can even spout some sports trivia and Christmas carols and stuff like that.', "We'd talk sports and stuff, and maybe have
a beer.", 'The Admirable Crichton of his day, he was keen alike on field sports and the arts, the friend and admirer equally of
Cecil Rhodes and of Rodin, a railway director and a yeomanry colonel.', "But he was not brought forward by his father or prepar
ed in any way for his future greatness, and lived in the country occupied with field sports, till after the institution of the
second protectorate in 16J7 and the recognition of Oliver's right to name his successor.", 'Thankfully, he was too concerned wi
th sports to get in any real trouble.', 'See Strutt, Sports and Pastimes, who also gives an illustration, "taken from a manuscr
iptal painting of the 9th century in the Cotton Library," representing "a Saxon chieftain, attended by his huntsman and a coupl
e of hounds, pursuing the wild swine in a forest."', "As they entered the yard, Carmen noticed Lori's little red sports car.",
'After a lengthy shower, Jenn exited and pulled on clean leggings, sports bra, and socks.', 'A park and sports ground at the we
stern end of the town contains the pedestal for a statue of President Kruger.', 'There were crude medieval notions that fossils
were " freaks " or " sports " of nature (lusus naturae), or that they represented failures of a creative force within the earth
(a notion of Greek and Arabic origin), or that larger and smaller fossils represented the remains of races of giants or of pygm
ies (the mythical idea).']
```

4. Do vectorization using TFID Vectorizer



```
  (0, 19)	0.24354521874367308
  (0, 123)	0.24354521874367308
  (0, 29)	0.24354521874367308
  (0, 90)	0.24354521874367308
  (0, 99)	0.24354521874367308
  (0, 97)	0.24354521874367308
  (0, 14)	0.24354521874367308
  (0, 33)	0.24354521874367308
  (0, 3)	0.24354521874367308
  (0, 43)	0.24354521874367308
  (0, 6)	0.24354521874367308
  (0, 105)	0.09004358647762914
  (0, 37)	0.2070356862180743
  (0, 4)	0.24354521874367308
  (0, 56)	0.24354521874367308
  (0, 28)	0.24354521874367308
  (0, 26)	0.24354521874367308
  (0, 2)	0.24354521874367308
  (1, 39)	0.21254144490425056
  (1, 111)	0.21254144490425056
  (1, 121)	0.21254144490425056
  (1, 87)	0.21254144490425056
  (1, 50)	0.21254144490425056
  (1, 24)	0.21254144490425056
  (1, 51)	0.21254144490425056
```

5. Generate vector-matrix



	16j7	9th	admirable	admirer	alike	arabic	arts	attended	beer	bra	...	thankfully	till	town	trivia	trouble
0	0.000000	0.000000	0.243545	0.243545	0.243545	0.000000	0.243545	0.000000	0.00000	0.00000	...	0.00000	0.000000	0.000000	0.000000	0.00000
1	0.000000	0.212541	0.000000	0.000000	0.000000	0.000000	0.000000	0.212541	0.00000	0.00000	...	0.00000	0.000000	0.000000	0.000000	0.00000
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	...	0.49167	0.000000	0.000000	0.000000	0.49167
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	...	0.00000	0.000000	0.000000	0.413119	0.00000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	...	0.00000	0.000000	0.314088	0.000000	0.00000
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	...	0.00000	0.000000	0.000000	0.000000	0.00000
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.50903	0.00000	...	0.00000	0.000000	0.000000	0.000000	0.00000
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.179211	0.000000	0.000000	0.00000	0.00000	...	0.00000	0.000000	0.000000	0.000000	0.00000
8	0.224397	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	...	0.00000	0.224397	0.000000	0.000000	0.00000
9	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.33083	...	0.00000	0.000000	0.000000	0.000000	0.00000

10 rows × 124 columns