Mini-Project -2

Text Similarity Checker

Nowadays by evolution of internet and search engines, the knowledge and data of all the world is easily available for us. During creation of a document, people used to refer or copy multiple documents present online. This copying is sometime legal and sometime illegal. There are many softwares in the market which tell us how much a document is similar to other sources.

In this project we will explore the idea behind this similarity checking.

You will be provided with 3 files containing some text. Perform the following steps on each file:

Suppose the text in a file is:

```
A do run run run, a do run run
```

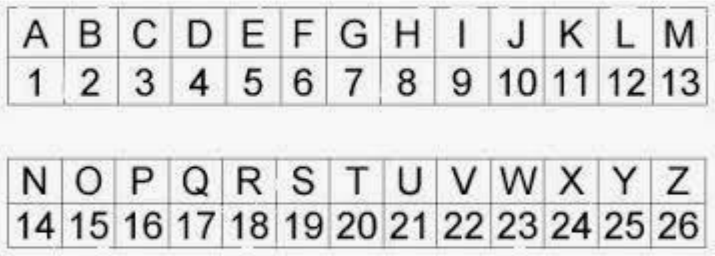
1) Remove all the non-alphanumeric characters from the text. Output will look like this:

```
adorunrunrunadorunrun
```

2) Make substrings of 5 characters each :

```
adoru dorun orunr runru unrun nrunr runru unrun nruna runad unado nador adoru dorun
orunr runru unrun
```

3) Take each substring and using the chart given below, Add the values of each character to come up with a unique number for that substring. Remember to convert all the text to lower or upper case as you like.



A	B	C	D	E	F	G	H	I	J	K	L	M
1	2	3	4	5	6	7	8	9	10	11	12	13

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
14	15	16	17	18	19	20	21	22	23	24	25	26

4) Suppose the output of step 3 is:

```
77 72 42 17 98 50 17 98 8 88 67 39 77 72 42 17 98
```

Each number is representing a substring from step 2. Save these numbers in a file.

5) Create 2x2 matrices from the above data using the same method which we used for creating substrings :

77	72

42	17

72	42
17	98

42	17
98	50

17	98
50	17

98	50
17	98

50	17
98	8

17	98
8	88

And so on. Create all possible matrix till end of data and save these to file.

6) Do the above steps to all text files to generate matrices. Matrices of each text file should be stored in separate files.

Comparison:

Suppose you want to compare file1 and file2.

1) Read one matrix of file1 and one matrix from file2.

2) Take the difference between two matrices of 1st step for example:

Matrix from file1 and file2:

77	72
42	17

18	88
47	70

Difference is:

59	16
5	53

3) Sum all the elements of this difference matrix.

4) If the sum is less than 60 it means we have a match otherwise ignore it.

5) Do the above 4 steps for 10 matrices.

6) Count how many matrix differences were less than 60.

7) Use that count to find the percentage of matches between files.

NOTE:

We have studied three main topics.

1) File Handling
2) Functions
3) 2D arrays

Use all the above topics as much as possible.