SNIST

Sreenidhi Institute of Science and Technology

Because, Life is all about taking the right decisions

DUPLI TRACKER

Presenters-

Ashwin Kumar Uppala- 19311A1901

Aakanksha R Rangdal- 19311A19E7

Mandepudi Rani Chowdary- 19311A19E9

Category- Prototype

Sreenidhi Institute of Science and Technology

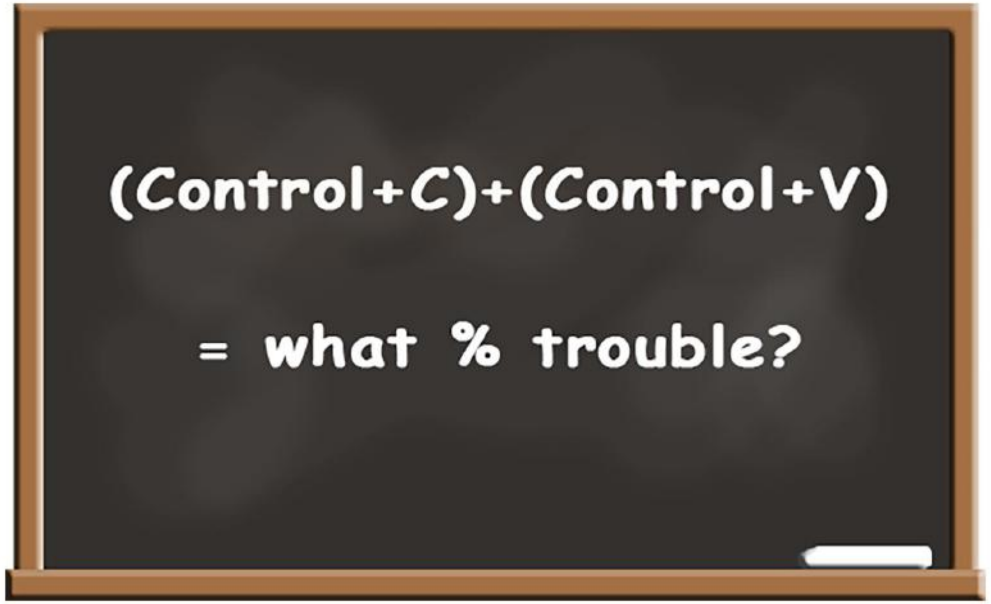
ECM Department

Mentor- Dr Manu Gupta

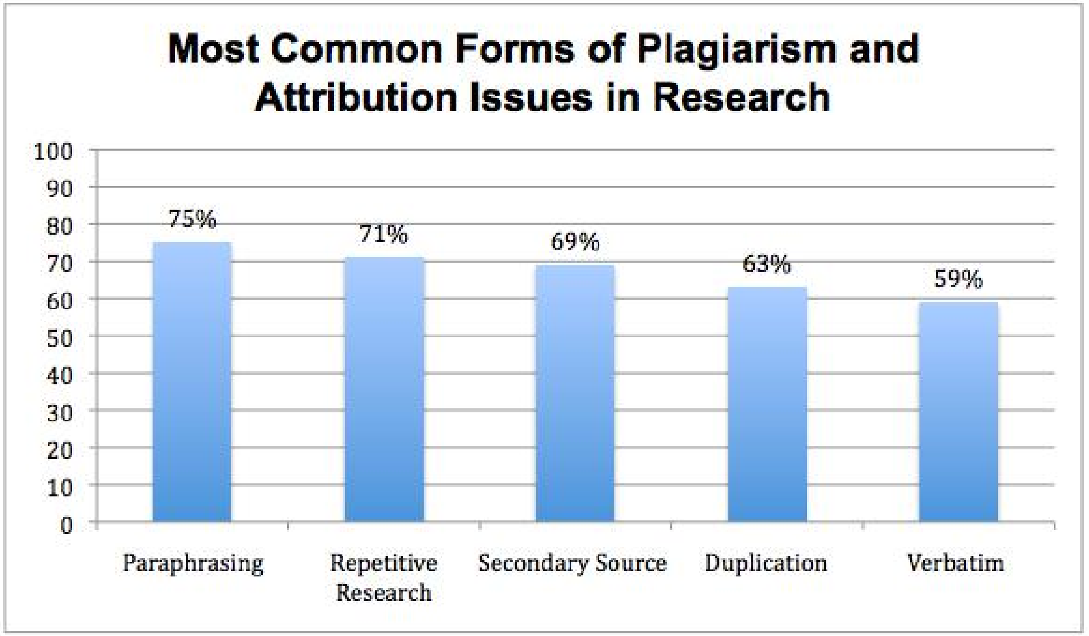
WHAT IS PLAGIARISM?

The practice of taking someone else's work or ideas and passing them off as one's own.

WHY AVOID IT?

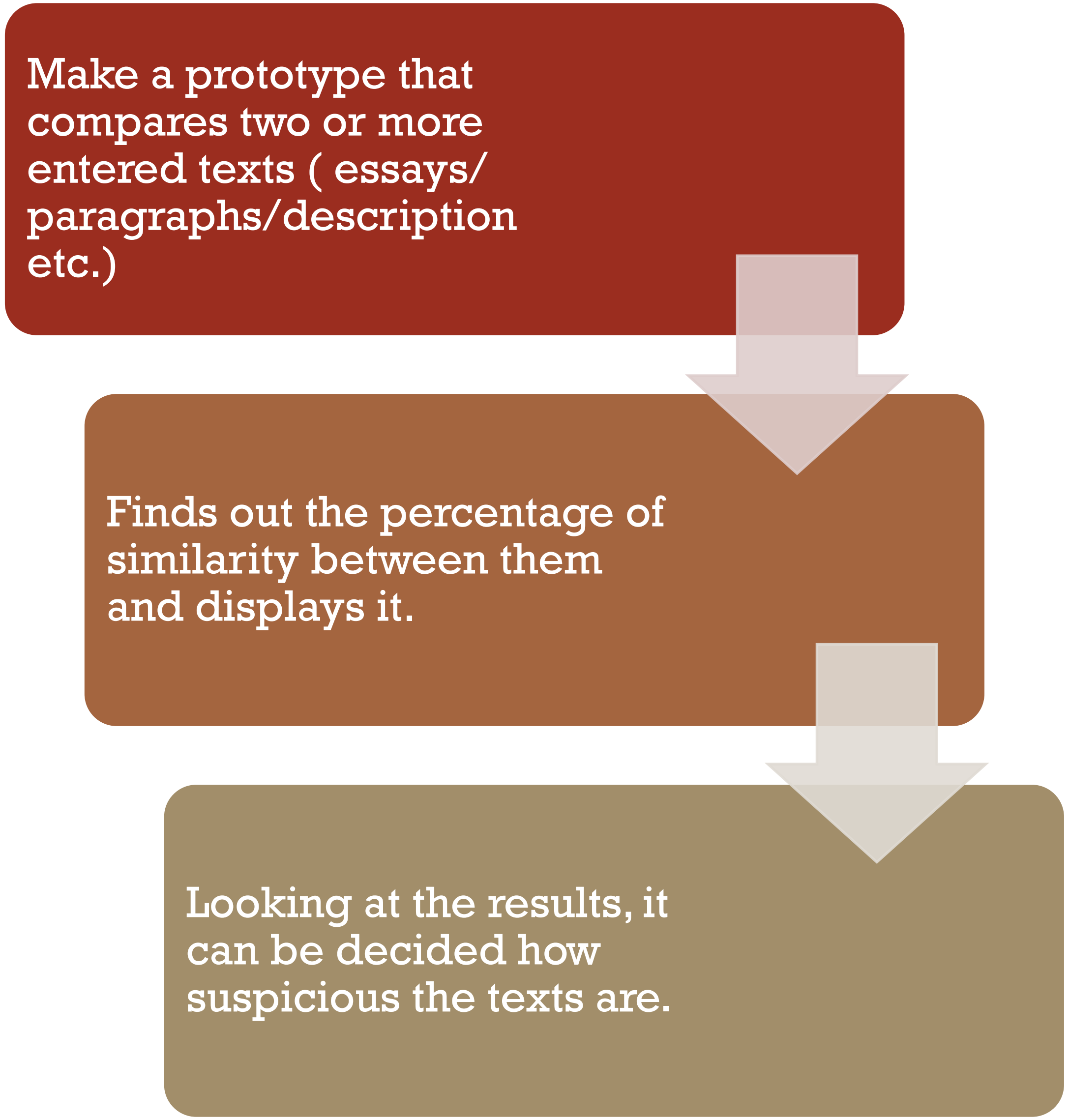


- **Plagiarism** is unethical because it is a form of theft.
- By taking the **ideas and words of others and pretending they are your own**, you are **stealing** someone else's intellectual property.
- **This can** get **you** expelled from your course, college and/or university.
- **It can** result in your work being destroyed.
- **Plagiarism can** result in legal action, fines, penalties and imprisonment etc.



TYPES of plagiarism over decades.

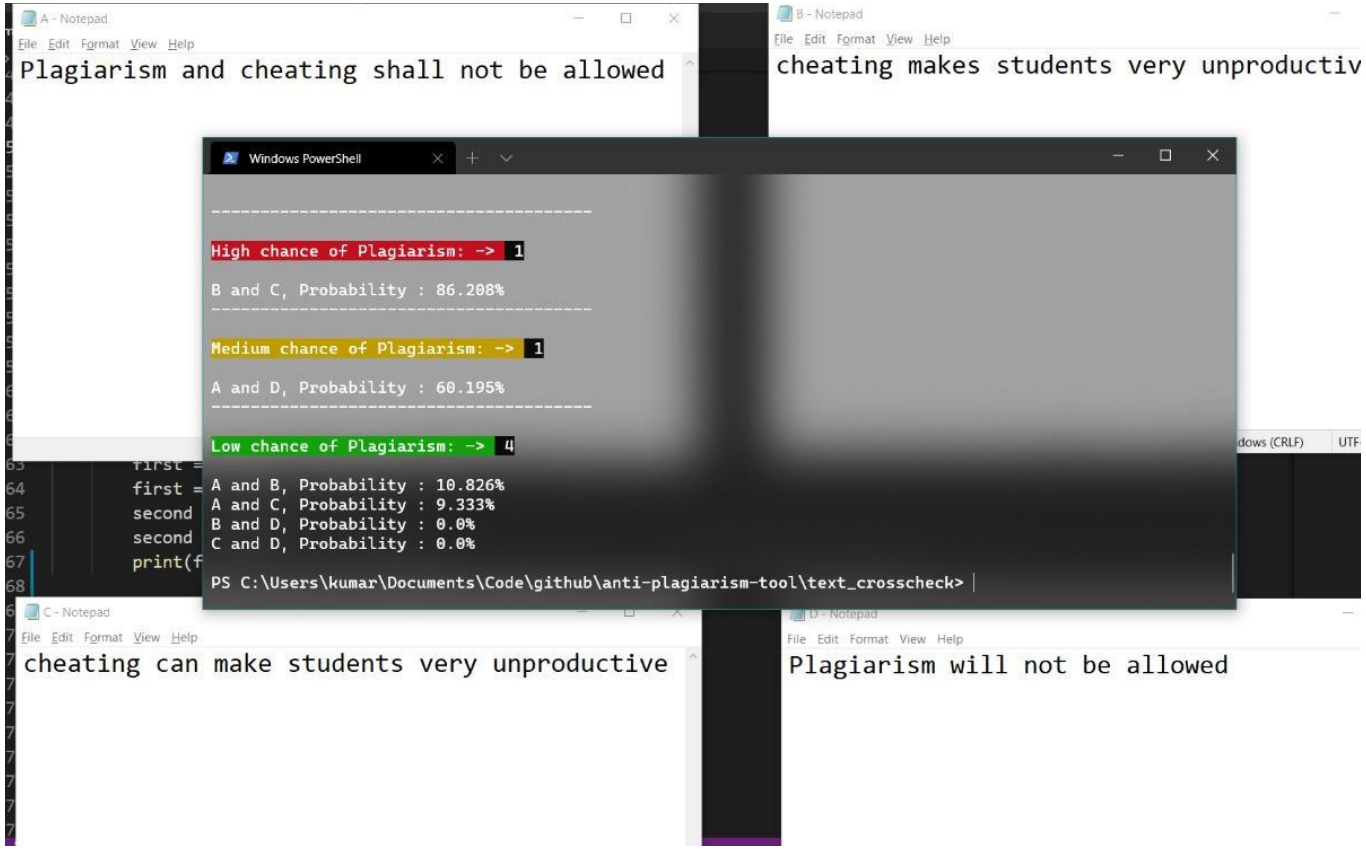
WHAT IS OUR IDEA TO AVOID IT?



Make a prototype that compares two or more entered texts (essays/ paragraphs/description etc.)

Finds out the percentage of similarity between them and displays it.

Looking at the results, it can be decided how suspicious the texts are.



```
Plagiarism and cheating shall not be allowed
```

```
cheating makes students very unproductiv
```

```
----------------------------------------

High chance of Plagiarism: -> 1

B and C, Probability : 86.208%
----------------------------------------

Medium chance of Plagiarism: -> 1

A and D, Probability : 60.195%
----------------------------------------

Low chance of Plagiarism: -> 4

A and B, Probability : 10.826%
A and C, Probability : 9.333%
B and D, Probability : 0.0%
C and D, Probability : 0.0%

PS C:\Users\kumar\Documents\Code\github\anti-plagiarism-tool\text_crosscheck>
```

```
        first =
        first =
        second
        second
        print(f
```

```
cheating can make students very unproductive
```

```
Plagiarism will not be allowed
```

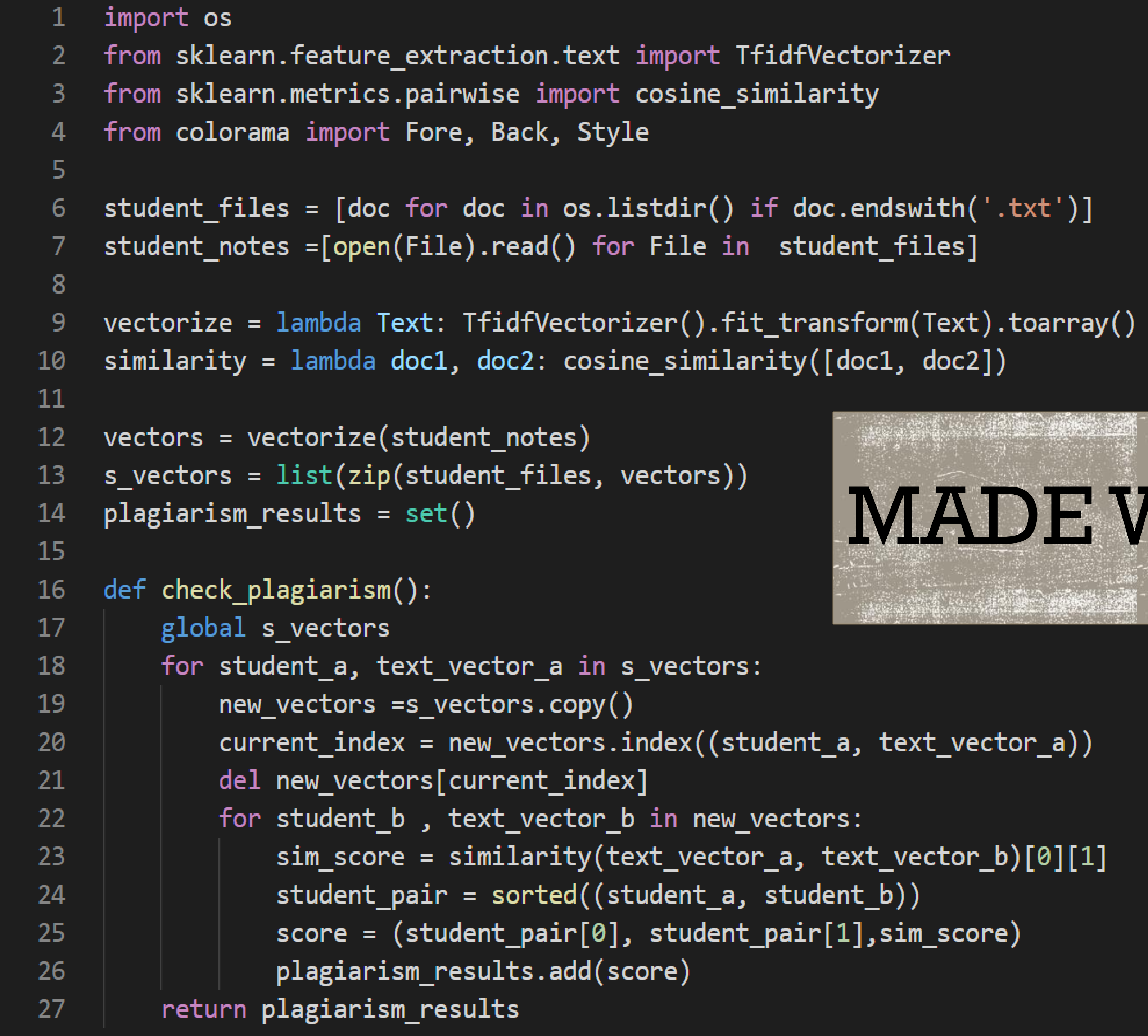
HERE IS HOW THE RESULTS ARE SEEN.

Four different texts by people A,B,C,D is been compared and the similarity percentage is displayed.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

THE ALGORITHM IS BASED ON MATHEMATICAL TEXT SIMILARITY FORMULA:

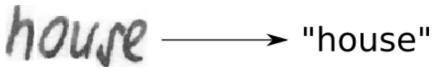
- Where:
- A -> Textual Content of writer A
- B -> Textual Content of writer B

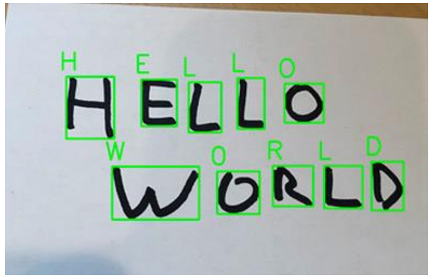


```python
import os
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from colorama import Fore, Back, Style

student_files = [doc for doc in os.listdir() if doc.endswith('.txt')]
student_notes =[open(File).read() for File in  student_files]

vectorize = lambda Text: TfidfVectorizer().fit_transform(Text).toarray()
similarity = lambda doc1, doc2: cosine_similarity([doc1, doc2])

vectors = vectorize(student_notes)
s_vectors = list(zip(student_files, vectors))
plagiarism_results = set()

def check_plagiarism():
    global s_vectors
    for student_a, text_vector_a in s_vectors:
        new_vectors =s_vectors.copy()
        current_index = new_vectors.index((student_a, text_vector_a))
        del new_vectors[current_index]
        for student_b , text_vector_b in new_vectors:
            sim_score = similarity(text_vector_a, text_vector_b)[0][1]
            student_pair = sorted((student_a, student_b))
            score = (student_pair[0], student_pair[1],sim_score)
            plagiarism_results.add(score)
    return plagiarism_results
```

MADE WITH PYTHON3

SUPPORTS BOTH TYPED AND HANDWRITTEN TEXT





MACHINE LEARNING IS USED FOR OPTICAL CHARACTER RECOGNITION (OCR)

MACHINE LEARNING IN SIMPLE TERMS

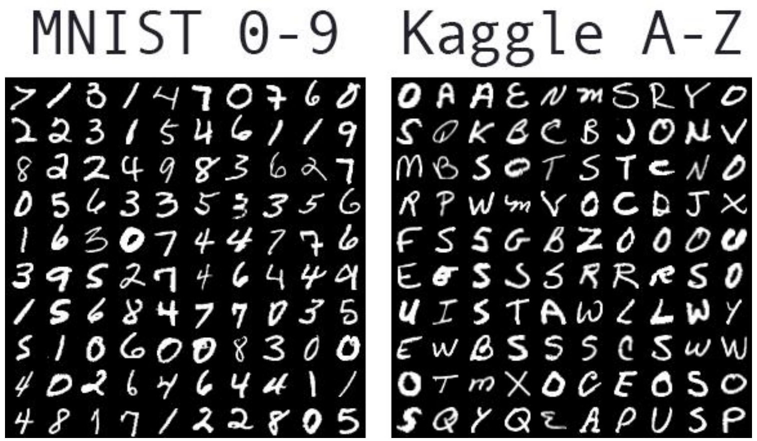
Dataset provided to Algorithm

75% of data is used for training the algorithm

25% of data is used to test the trained algorithm

The algorithm now used produces high efficiency

DATASETS USED :



ADVANTAGES

- Students maintaining academic honesty.
- Students using their creativity instead of copying ideas for assignments.
- Fair competition in the field of Literature where content should be genuine.
- Original authors will get their fair recognition.

REFERENCES:

1. Cosine Similarity - Text Similarity Metric - Machine Learning Tutorials (studymachinelearning.com)
2. tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository) (github.com)
3. OpenCV: OCR of Hand-written Data using kNN
4. scikit-learn: machine learning in Python — scikit-learn 0.24.1 documentation (scikit-learn.org)
5. Stock Images from | Unsplash

CODE REPOSITORY AVAILABLE AT

- https://github.com/ashwinexe/anti-plagiarism-tool

VIDEO AVAILABLE AT

https://youtu.be/oRiEpIBWS9Y

THANKYOU