SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

FACULTY OF SCIENCE AND TECHNOLOGY

(Formerly SRM University, Under section 3 of UGC Act, 1956)

S.R.M NAGAR, KATTANKULATHUR – 603 203,

KANCHEEPURAM DISTRICT

SCHOOL OF COMPUTING

DEPARTMENT OF NETWORKING AND COMMUNICATIONS

Course Code:	18CSE305J
Course Name:	Artificial Intelligence

Course Project

Title: Plagiarism Checker

Team Members:

1. RA1911030010069 – Praveen Kumar
2. RA1911030010090 – Tejas Ashok
3. RA1911030010103 – Vinoth S

Date: 18-04-2022

Title: Plagiarism Checker - Python

Problem Statement:

We all know that computers are good at numbers, so in order to compute the similarity between on two text documents, the textual raw data is transformed into vectors => arrays of numbers and then from that we are going to use a basic knowledge vector to compute the similarity between them.

This python application identifies similarities between a test.txt file that can be created in the local directory and a website. The link to the website can be provided at the start of the program and the application scrapes the website using requests and beautifulsoup. The driver code then compares the scraped data and test.txt for similarities and produces the results.

Working:

The Plagiarism Checker uses the request module to send requests to a website (Wikipedia currently supported), to receive all the html data from the website.

Beautifulsoup then parses the raw html and organizes the data so that it could be manipulated.

The data is then split according to headings on the website and then are stored in sperate text files according to their headings.

The data of all the files are then compared with each other by converting the data to vectors and then comparing it with cosine_similarity.

The output is then displayed. The files with similarity close to 1 are considered copied.

Code:

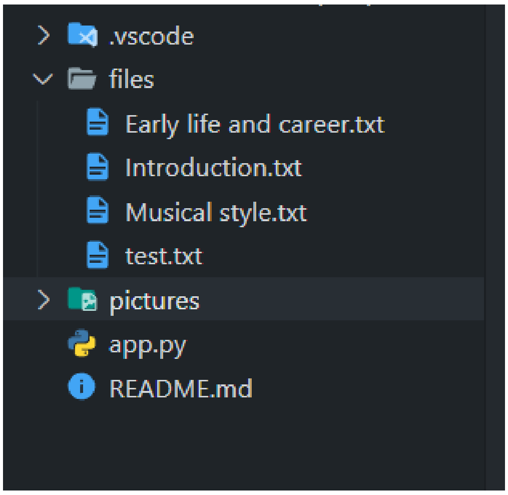
```
import os

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.metrics.pairwise import cosine_similarity

import requests
```

```
from bs4 import BeautifulSoup as bs

# Directory of the files
DIR = "files"

# URL and Headers
url = input("Enter the URL (Wikipedia Only): ")

headers = {
    "Host": "en.wikipedia.org",
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:99.0) Gecko/20100101 Firefox/99.0"
}

# Request the url with headers and convert to text
r = requests.get(url, headers=headers).text

# Creates a bs object to perform parsing
soup = bs(r, 'lxml')
TITLE = soup.find('h1').get_text()

# Dictionary to store parsed string
parsed_dic = {}

# Parsing
para = soup.find('p', class_=None)

for d in para.find_all('sup'):
    d.decompose()
parsed_dic["Introduction"] = para.get_text()


for tag in soup.find_all('h2'):
    sib = tag.find_next_sibling('p')
    if sib is None:
        continue
    p = ""
    while(sib is not None and
sib.find_previous_sibling('h2').find('span').get_text() ==
tag.find('span').get_text()):
        for d in sib.find_all('sup'):
            d.decompose()


        p += sib.get_text()
        sib = sib.find_next_sibling('p')
    parsed_dic[tag.find('span').get_text()] = p

# Creates seperate txt files for every heading in Wikipedia
for key, value in parsed_dic.items():
    with open(DIR + os.sep + key + '.txt', 'w', encoding="utf-8") as f:
        f.write(value)

# Creates a list of files and its data
student_files = [doc for doc in os.listdir(DIR) if doc.endswith('.txt')]

student_notes = [open(DIR + os.sep + _file, encoding='utf-8').read()
                 for _file in student_files]
```

```
# Creates vectors of the data of each file
def vectorize(Text): return TfidfVectorizer().fit_transform(Text).toarray()
def similarity(doc1, doc2): return cosine_similarity([doc1, doc2])

# Compares every files vector with each other
vectors = vectorize(student_notes)
s_vectors = list(zip(student_files, vectors))
plagiarism_results = set()

# Function to compare the vectors
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

# Print the result
print("Not very Similar:")
for data in check_plagiarism():
    if data[1].split('.')[0] == 'test' and data[2] <= 0.5:
        print(data)
print()
print("Are kind of Similar:")
for data in check_plagiarism():
    if data[1].split('.')[0] == 'test' and data[2] > 0.5 and data[2] <= 0.75:
        print(data)
print()
print("A lot Similar:")
for data in check_plagiarism():
    if data[1].split('.')[0] == 'test' and data[2] > 0.75:
        print(data)
```

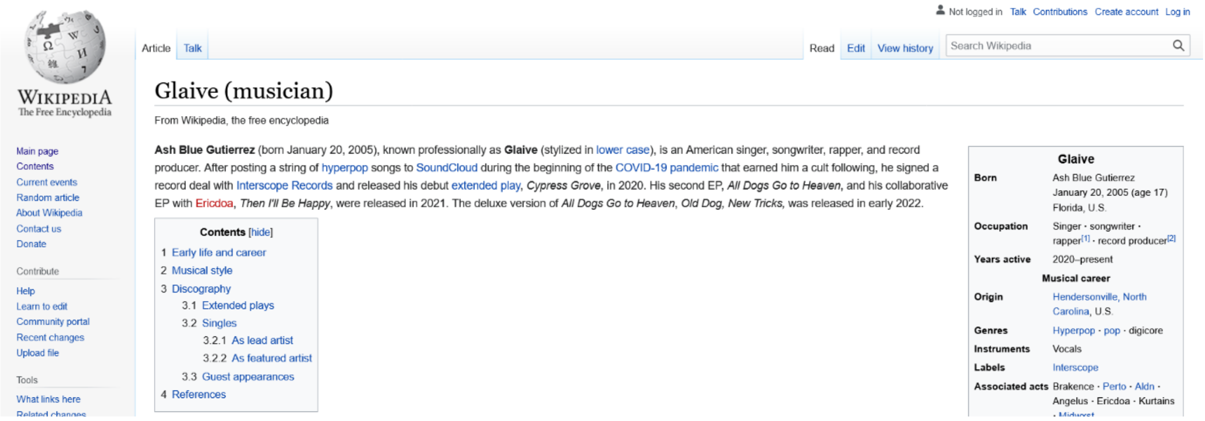
<u>Test Case #1:</u>

Directory



```
> .vscode
v files
    Early life and career.txt
    Introduction.txt
    Musical style.txt
    test.txt
> pictures
  app.py
  README.md
```

Webpage



Glaive (musician)

From Wikipedia, the free encyclopedia

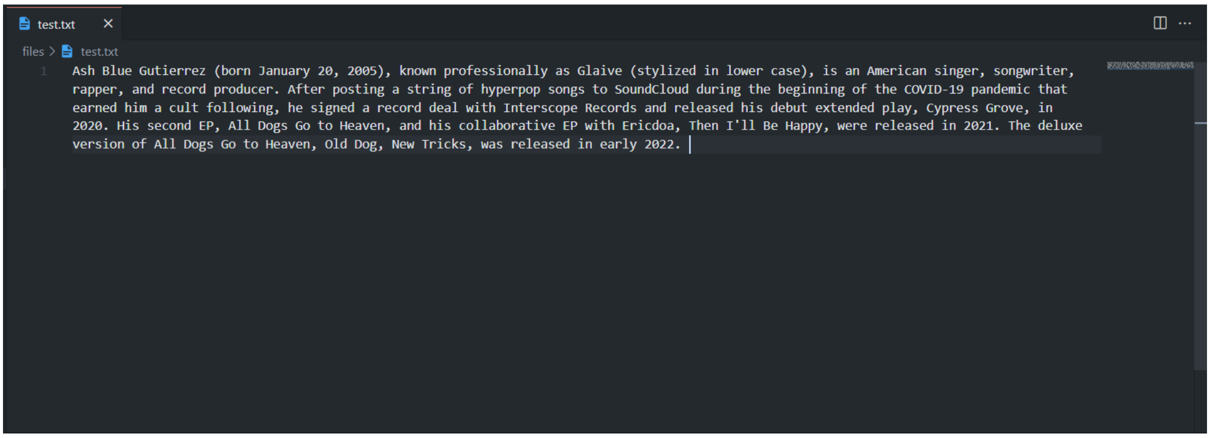
Ash Blue Gutierrez (born January 20, 2005), known professionally as **Glaive** (stylized in lower case), is an American singer, songwriter, rapper, and record producer. After posting a string of hyperpop songs to SoundCloud during the beginning of the COVID-19 pandemic that earned him a cult following, he signed a record deal with Interscope Records and released his debut extended play, *Cypress Grove*, in 2020. His second EP, *All Dogs Go to Heaven*, and his collaborative EP with Ericdoa, *Then I'll Be Happy*, were released in 2021. The deluxe version of *All Dogs Go to Heaven*, *Old Dog, New Tricks*, was released in early 2022.

Contents [hide]

1 Early life and career
2 Musical style
3 Discography
  3.1 Extended plays
  3.2 Singles
    3.2.1 As lead artist
    3.2.2 As featured artist
  3.3 Guest appearances
4 References

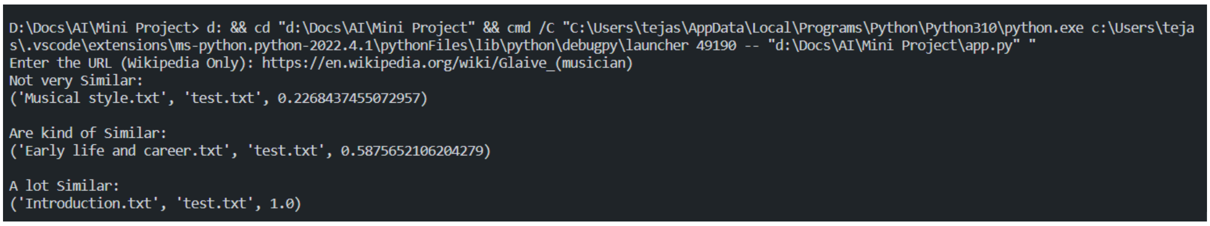
Glaive	
Born	Ash Blue Gutierrez January 20, 2005 (age 17) Florida, U.S.
Occupation	Singer · songwriter · rapper[1] · record producer[2]
Years active	2020–present
Musical career	
Origin	Hendersonville, North Carolina, U.S.
Genres	Hyperpop · pop · digicore
Instruments	Vocals
Labels	Interscope
Associated acts	Brakence · Perto · Aldn · Angelus · Ericdoa · Kurtains · Midwxst

Input



```
Ash Blue Gutierrez (born January 20, 2005), known professionally as Glaive (stylized in lower case), is an American singer, songwriter,
rapper, and record producer. After posting a string of hyperpop songs to SoundCloud during the beginning of the COVID-19 pandemic that
earned him a cult following, he signed a record deal with Interscope Records and released his debut extended play, Cypress Grove, in
2020. His second EP, All Dogs Go to Heaven, and his collaborative EP with Ericdoa, Then I'll Be Happy, were released in 2021. The deluxe
version of All Dogs Go to Heaven, Old Dog, New Tricks, was released in early 2022.
```

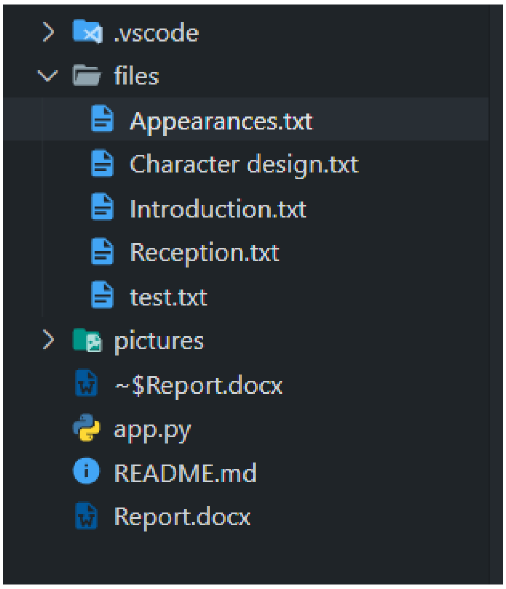
Output



```
D:\Docs\AI\Mini Project> d: && cd "d:\Docs\AI\Mini Project" && cmd /C "C:\Users\tejas\AppData\Local\Programs\Python\Python310\python.exe c:\Users\teja
s\.vscode\extensions\ms-python.python-2022.4.1\pythonFiles\lib\python\debugpy\launcher 49190 -- "d:\Docs\AI\Mini Project\app.py" "
Enter the URL (Wikipedia Only): https://en.wikipedia.org/wiki/Glaive_(musician)
Not very Similar:
('Musical style.txt', 'test.txt', 0.2268437455072957)

Are kind of Similar:
('Early life and career.txt', 'test.txt', 0.5875652106204279)

A lot Similar:
('Introduction.txt', 'test.txt', 1.0)
```

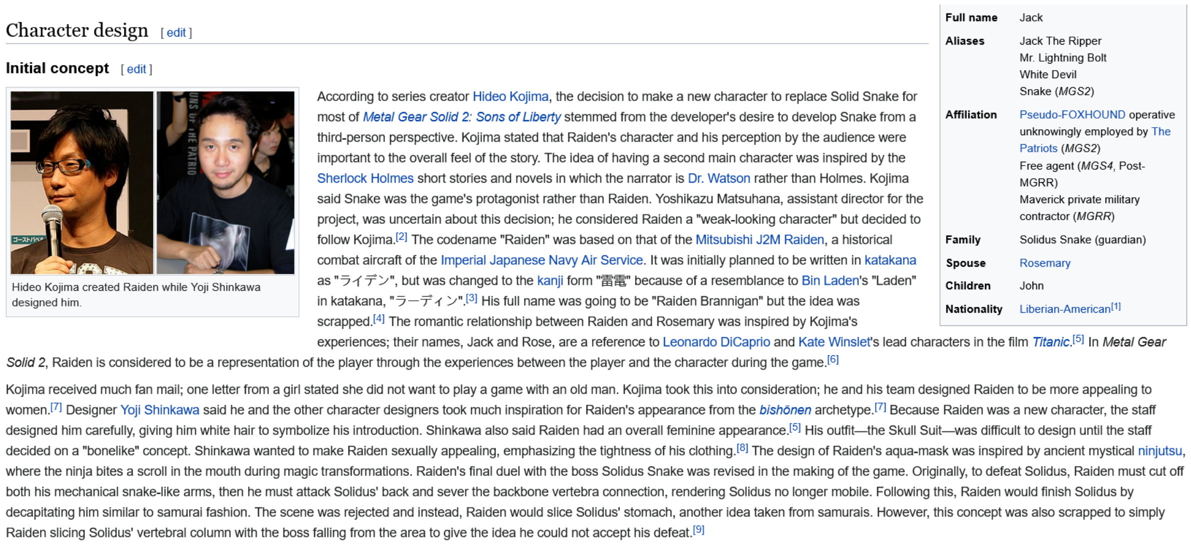
Test Case #2:

Directory



```
> .vscode
v files
    Appearances.txt
    Character design.txt
    Introduction.txt
    Reception.txt
    test.txt
> pictures
  ~$Report.docx
  app.py
  README.md
  Report.docx
```

Webpage



Character design [edit]

Initial concept [edit]

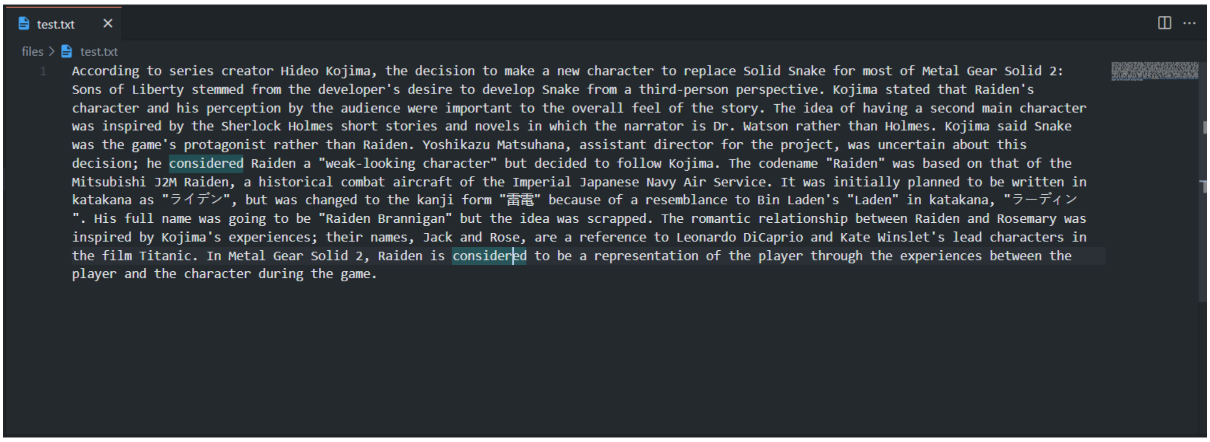
Full name	Jack
Aliases	Jack The Ripper Mr. Lightning Bolt White Devil Snake (*MGS2*)
Affiliation	Pseudo-FOXHOUND operative unknowingly employed by The Patriots (*MGS2*) Free agent (*MGS4*, Post-MGRR) Maverick private military contractor (*MGRR*)
Family	Solidus Snake (guardian)
Spouse	Rosemary
Children	John
Nationality	Liberian-American[1]

Hideo Kojima created Raiden while Yoji Shinkawa designed him.

According to series creator Hideo Kojima, the decision to make a new character to replace Solid Snake for most of *Metal Gear Solid 2: Sons of Liberty* stemmed from the developer's desire to develop Snake from a third-person perspective. Kojima stated that Raiden's character and his perception by the audience were important to the overall feel of the story. The idea of having a second main character was inspired by the Sherlock Holmes short stories and novels in which the narrator is Dr. Watson rather than Holmes. Kojima said Snake was the game's protagonist rather than Raiden. Yoshikazu Matsuhana, assistant director for the project, was uncertain about this decision; he considered Raiden a "weak-looking character" but decided to follow Kojima.[2] The codename "Raiden" was based on that of the Mitsubishi J2M Raiden, a historical combat aircraft of the Imperial Japanese Navy Air Service. It was initially planned to be written in katakana as "ライデン", but was changed to the kanji form "雷電" because of a resemblance to Bin Laden's "Laden" in katakana, "ラーディン".[3] His full name was going to be "Raiden Brannigan" but the idea was scrapped.[4] The romantic relationship between Raiden and Rosemary was inspired by Kojima's experiences; their names, Jack and Rose, are a reference to Leonardo DiCaprio and Kate Winslet's lead characters in the film *Titanic*.[5] In *Metal Gear Solid 2*, Raiden is considered to be a representation of the player through the experiences between the player and the character during the game.[6]

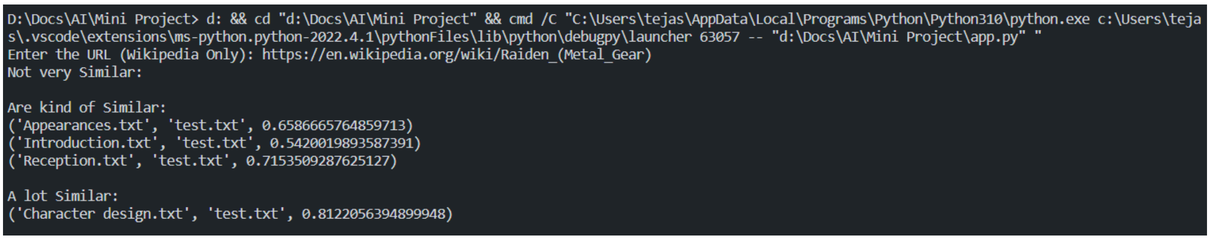
Kojima received much fan mail; one letter from a girl stated she did not want to play a game with an old man. Kojima took this into consideration; he and his team designed Raiden to be more appealing to women.[7] Designer Yoji Shinkawa said he and the other character designers took much inspiration for Raiden's appearance from the *bishōnen* archetype.[7] Because Raiden was a new character, the staff designed him carefully, giving him white hair to symbolize his introduction. Shinkawa also said Raiden had an overall feminine appearance.[5] His outfit—the Skull Suit—was difficult to design until the staff decided on a "bonelike" concept. Shinkawa wanted to make Raiden sexually appealing, emphasizing the tightness of his clothing.[8] The design of Raiden's aqua-mask was inspired by ancient mystical ninjutsu, where the ninja bites a scroll in the mouth during magic transformations. Raiden's final duel with the boss Solidus Snake was revised in the making of the game. Originally, to defeat Solidus, Raiden must cut off both his mechanical snake-like arms, then he must attack Solidus' back and sever the backbone vertebra connection, rendering Solidus no longer mobile. Following this, Raiden would finish Solidus by decapitating him similar to samurai fashion. The scene was rejected and instead, Raiden would slice Solidus' stomach, another idea taken from samurais. However, this concept was also scrapped to simply Raiden slicing Solidus' vertebral column with the boss falling from the area to give the idea he could not accept his defeat.[9]

Input



```
According to series creator Hideo Kojima, the decision to make a new character to replace Solid Snake for most of Metal Gear Solid 2:
Sons of Liberty stemmed from the developer's desire to develop Snake from a third-person perspective. Kojima stated that Raiden's
character and his perception by the audience were important to the overall feel of the story. The idea of having a second main character
was inspired by the Sherlock Holmes short stories and novels in which the narrator is Dr. Watson rather than Holmes. Kojima said Snake
was the game's protagonist rather than Raiden. Yoshikazu Matsuhana, assistant director for the project, was uncertain about this
decision; he considered Raiden a "weak-looking character" but decided to follow Kojima. The codename "Raiden" was based on that of the
Mitsubishi J2M Raiden, a historical combat aircraft of the Imperial Japanese Navy Air Service. It was initially planned to be written in
katakana as "ライデン", but was changed to the kanji form "雷電" because of a resemblance to Bin Laden's "Laden" in katakana, "ラーディン
". His full name was going to be "Raiden Brannigan" but the idea was scrapped. The romantic relationship between Raiden and Rosemary was
inspired by Kojima's experiences; their names, Jack and Rose, are a reference to Leonardo DiCaprio and Kate Winslet's lead characters in
the film Titanic. In Metal Gear Solid 2, Raiden is considered to be a representation of the player through the experiences between the
player and the character during the game.
```

Output



```
D:\Docs\AI\Mini Project> d: && cd "d:\Docs\AI\Mini Project" && cmd /C "C:\Users\tejas\AppData\Local\Programs\Python\Python310\python.exe c:\Users\teja
s\.vscode\extensions\ms-python.python-2022.4.1\pythonFiles\lib\python\debugpy\launcher 63057 -- "d:\Docs\AI\Mini Project\app.py" "
Enter the URL (Wikipedia Only): https://en.wikipedia.org/wiki/Raiden_(Metal_Gear)
Not very Similar:

Are kind of Similar:
('Appearances.txt', 'test.txt', 0.6586665764859713)
('Introduction.txt', 'test.txt', 0.5420019893587391)
('Reception.txt', 'test.txt', 0.7153509287625127)

A lot Similar:
('Character design.txt', 'test.txt', 0.8122056394899948)
```

GitHub Repository:

https://github.com/thesh4de/AI-mini-project