```
Text Data Preprocessing
Text Data To Numerical Vector Conversion using:
Word2Vec
Pretrained GloVe
Pretrained BERT
We will follow the steps mentioned below:

Data Loading
Exploratory Data Analysis
Data Preparation
Train Test Split
Data Preprocessing (special characters, stop words, lower case, stemming, etc)
Converting text to numerical vector using Word2Vec, Pretrained GloVe and BERT
Preprocessing Test Data
Training on Train Data
Predictions on Test Data
Model Evaluation
```

Project Hint - Reading the Data from Database

In [1]:
```
import sqlite3
import pandas as pd
import numpy as np
```

Step 1 - Reading the Tables from Database file

In [2]:
```
# Read the code below and write your observation in the next cell

conn = sqlite3.connect('eng_subtitles_database.db')
cursor = conn.cursor()
cursor.execute("SELECT name FROM sqlite_master WHERE type='table'")
print(cursor.fetchall())
```

```
[('zipfiles',)]
```

In the above cell, I am able to read the table inside the database. As mentioned earlier, table name is `zipfiles` . We also know from README.txt that this table contains three columns: 'num', 'name' and 'content'.

Step 2 - Reading the columns of Table

In [3]:
```
cursor.execute("PRAGMA table_info('zipfiles')")
cols = cursor.fetchall()
for col in cols:
    print(col[1])
```

```
num
name
content
```

The above code helps in checking the column names in the database table.

Let's now use `SELECT * FROM zipfiles` to read all the data into a `df` variable.

Step 3 - Loading the Database Table inside a Pandas DataFrame

In [4]:
```
df = pd.read_sql_query("""SELECT * FROM zipfiles""", conn)
df.head()
```

Out[4]:

	num	name	content
0	9180533	the.message.(1976).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x1c\xa9\x...
1	9180583	here.comes.the.grump.s01.e09.joltin.jack.in.bo...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x17\xb9\x...
2	9180592	yumis.cells.s02.e13.episode.2.13.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00L\xb9\x99V...
3	9180594	yumis.cells.s02.e14.episode.2.14.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00U\xa9\x99V...
4	9180600	broker.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x001\xa9\x99V...

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 82498 entries, 0 to 82497
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   num      82498 non-null  int64 
 1   name     82498 non-null  object
 2   content  82498 non-null  object
dtypes: int64(1), object(2)
memory usage: 1.9+ MB
```

Looks like the `content` column donot contain the subtitles text. Instead as mentioned in README.txt, it might be latin-1 encoded.

Step 4 - Printing `content` of 0th Row

In [6]:
```
b_data = df.iloc[0, 2]

# here 2 represent the index of content column
# 0 represents the row number
```

In [7]:
```
print(b_data)
```

```
b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x1c\xa9\x99V\x9fx\x96\xf0\x8c\x9e\x00\x00\x86\x9b\x01\x00;\x00\x00\x00The.Message.1976.REMAS
TERED.1080p.BluRay.x264-PiGNUS.EN.srt\xad\xbdm\x93\xdc\xc6\x91.\xfa\x9d\x11\xfc\x0f-}\xe1=\x11-\x9d\x06P\x85\x17\x9d\x8d\xd5%%[\x
a4-Y>&u\x15>\xdf\xd0\xd3\x98\x19x\xfae\x0cts<\xfe\xf57\x9f\'\xb3\n\xd9\xa4\xbc\xbb\xf7\xc6Fl\xacELW\xa2\xaa\x90\x95\x95\xafO\x16/
_l6\xdf\xe0\xff\xea\xf5f\xb3Y}\xf5\xd5\xbf\xaf\xf4AQ\xae7Mx\xf9\xe2\xd7\xfe|s\xbf\xea\x8f\xcf\xab\x8f\xe3n8\xadN\xc7\xfdx\x1cVO\x
e3\xf9~\xf5\xf3\xe3p\xfc\xea\xfd/o>\xbc\xfb\xf0\xe3\xef\xde\xbf|\xf1\xfbi\x18Vo\xa6\xd3\xd3<L\xab\xe1\x1f\xe7\xe18\x8f\xa7\xe37\x
ab\xd3\xbc\xdb~-\xc3\x1e\xfe\xa7<|\xf9\xe2\xe5\x8bR_[~S\xd6\xeb\xa2k\xf3k\xe5A\xb7\xeeb\xf5\xf2\xc5\xbb\xe3\xea|?\xac\x8e\xfdaX\x
9dnW?\x9cvk>8\x9c\xe6\xf3\xean\xeao\xc6\xd3ev\x8f~\x1a\xa6\x9b\xf1\xf6\xb2\xff\x1a\xe4\xabD\xbe*d\x11\xa5#_U\xeb\xaa\xd9`\xa6\xa7
\xc3\xea\xa7\xcb}\x7f8\xf4F\xf9\xa7a\x9e\x87\xe3\x9d\xcc\\\xdf\x07B!\x13\xaa\xd61n<!\xd9\xaf\xd0\xbc|q>\xad\xde\x0e2\x9d\xfd\x98
\xa63\x1c\x1e\x87\xe9D*o\x9e\xff\xd9\x1f\xcf\xe3\xe5\xc0Y\xc5L\xac^\x87M\xe3\x89u\xebP\xcb\xbb\xeed\xff\xce\xe3\xf1n^\t\xd9\xfb\x
f1\xb0z\xba?\xad\xc6\x19d_\xbe\xb8=\xed\xf7\xa7\'\x9d\xde4\xde\xdd\x9f\x07\xd9\x81\xd5\xdde\xdc\xf5\xc7\x9b\x81o\xa8\xf3\x1b\xbau
\xd3V\xee\r\xa1\\\xc7\xae\x90m]m\xc7\xdd\xea\xf9t\xe1\x1b\x86~\xe2\x94w\xe3G|\xd3\x9b~\xaf\xfb\xd7$:\xa1Z\x17\xb5_v\x88\xeb\xb6h@
\xa7?\xe8\xe6\x7f\xb2k\xa0\x8b\xe7\x8f\xc3\xe9q?\xfc/Pk3\xb5z]U~\xdd\xa1\x95i\xca\xba\xfb\x9b\x9b\xe1\xf1\xbcz7\xef\x85\xea\xad\x
ec\x9dLpZ\xcd\xfd\xfec\x7f\x16\x06\xe2\x9c\xbaD%\x96\xeb\xb2*\x1c\x95\x18\xd6\xcdF\xfe\xfcvX\xcd\x8fC\xff0c*\xfd\xea8<\xad\x1e\xa
7\xd3\xe3\xfdp^\x8d\xc7\xd5\xeb\xa9\xdf\x8e=I\x15\x9bL+\xae7\xb5g\xbf\xd8\xac\xdb2\x82\xeb\xe7\xd5x^\xed\xc7\x87A\x96#\x9f\xe0\xe
9~8\xae\xfep\xba?\xeaG~\xd3?\x9e\xc7Yx\xf9\x06\x0c*k~\x90\xcf\x06NP\xbe)\x8a\xfc\x86v]\x84\x98\xdfP}\xb3)\xd6e#\xfc\x7f\xba\x9c1O
n\xff \xe7\xe5\xbc^\xddL\xcf\xa0\xd2o\xf1\xb7\xbc\xfaoI\xd0NL\xc5\x13S\xd5\x8e`\xd9\xae\xbb\x8dp\xfc\x87\x930\xf4\xdf\xfb\xa7\xbf
\x8f\xc2\x88\x7f\xee\xcf\xd3\xd8Orr\xe5\x1d\xaf\xf7\xc3?\xfa\xe3nJ\xcb\xaf\x12-\xe1\x8a\xe0\x0e\xbd<\x08\xeb\xba\x92\xef\xf5\xc7q
\x9e\xfa\xb5g\xe5?\x0f\xd3\x9c\xc6\x87<^v>6~|\xb3\x0e\x9dl_:Yd\xa9\x99\xfcF\x81\x81\xd5\xe2\x911\x8b\x92\x8b\x99\\\xb3n\x8a\xe0\x
c8\xc5\x8d\x9c:\xd9\xca\xd7\xff\t\x7f|\xad\x84\xeaDH\xc6\xb5\x9d_\x97\xf0L\x08\xc2"\xc3a+guPJ:\xa8\xc9\x83\xc2\xba\xac\xfdb\xe4\x
```

From the content, it appears to start with the bytes "PK\x03.....", which suggests that it might be a ZIP archive file. How do I know it? Experience! I have worked with something similar earlier.

Step 5 - Unzipping the content of 385th row and decoding using `latin-1`

In [8]:
```
import zipfile
import io

# Assuming 'content' is the binary data from your database
binary_data = df.iloc[385, 2]

# Decompress the binary data using the zipfile module
with io.BytesIO(binary_data) as f:
    with zipfile.ZipFile(f, 'r') as zip_file:
        # Reading only one file in the ZIP archive
        subtitle_content = zip_file.read(zip_file.namelist()[0])

# Now 'subtitle_content' should contain the extracted subtitle content
print(subtitle_content.decode('latin-1'))  # Assuming the content is latin-1 encoded text
```

```
1
00:00:06,000 --> 00:00:12,074
Watch any video online with Open-SUBTITLES
Free Browser extension: osdb.link/ext

2
00:00:15,370 --> 00:00:16,506
You lose everything, my girl.

3
00:00:16,530 --> 00:00:19,360
So you've said - four times.

4
00:00:20,330 --> 00:00:22,120
I definitely had
it on yesterday.

5
```

Look's like it worked.

Step 6 - Applying the above Function on the Entire Data

In [9]:
```
import zipfile
import io
import pandas as pd

def decode_method(binary_data):
    # Decompress the binary data using the zipfile module
    with io.BytesIO(binary_data) as f:
        with zipfile.ZipFile(f, 'r') as zip_file:
            # Assuming there's only one file in the ZIP archive
            subtitle_name = zip_file.namelist()[0]
            subtitle_content = zip_file.read(subtitle_name)

    # Now 'subtitle_content' should contain the extracted subtitle content
    subtitle_content_str = subtitle_content.decode('latin-1')  # Assuming the content is UTF-8 encoded text

    # Create a DataFrame containing the subtitle name and content
    df = pd.DataFrame({'Subtitle_Name': [subtitle_name], 'Subtitle_Content': [subtitle_content_str]})

    return subtitle_content_str, df
```

In [10]:
```
df.head()
```

Out[10]:

	num	name	content
0	9180533	the.message.(1976).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x1c\xa9\x...
1	9180583	here.comes.the.grump.s01.e09.joltin.jack.in.bo...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x17\xb9\x...
2	9180592	yumis.cells.s02.e13.episode.2.13.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00L\xb9\x99V...
3	9180594	yumis.cells.s02.e14.episode.2.14.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00U\xa9\x99V...
4	9180600	broker.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x001\xa9\x99V...

In [11]:
```
import zipfile
import io

count = 0

def decode_method(binary_data):
    global count
    # Decompress the binary data using the zipfile module
    # print(count, end=" ")
    count += 1
    with io.BytesIO(binary_data) as f:
        with zipfile.ZipFile(f, 'r') as zip_file:
            # Assuming there's only one file in the ZIP archive
            subtitle_content = zip_file.read(zip_file.namelist()[0])

    # Now 'subtitle_content' should contain the extracted subtitle content
    return subtitle_content.decode('latin-1')  # Assuming the content is UTF-8 encoded text
```

In [12]:
```
df['file_content'] = df['content'].apply(decode_method)

df.head()
```

Out[12]:

	num	name	content	file_content
0	9180533	the.message.(1976).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x1c\xa9\x...	1\r\n00:00:06,000 --> 00:00:12,074\r\nWatch an...
1	9180583	here.comes.the.grump.s01.e09.joltin.jack.in.bo...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x17\xb9\x...	1\r\n00:00:29,359 --> 00:00:32,048\r\nAh! Ther...
2	9180592	yumis.cells.s02.e13.episode.2.13.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00L\xb9\x99V...	1\r\n00:00:53,200 --> 00:00:56,030\r\n<i>Yumi'...
3	9180594	yumis.cells.s02.e14.episode.2.14.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00U\xa9\x99V...	1\r\n00:00:06,000 --> 00:00:12,074\r\nWatch an...
4	9180600	broker.(2022).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x001\xa9\x99V...	ï»¿1\r\n00:00:06,000 --> 00:00:12,074\r\nWatch...

In [13]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 82498 entries, 0 to 82497
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   num           82498 non-null  int64 
 1   name          82498 non-null  object
 2   content       82498 non-null  object
 3   file_content  82498 non-null  object
dtypes: int64(1), object(3)
memory usage: 2.5+ MB
```

In [14]:
```
df.tail()
```

Out[14]:

	num	name	content	file_content
82493	9521935	the.prophets.game.(2000).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\xb8\xa6\x...	ï»¿1\r\n00:01:16,284 --> 00:01:19,537\r\nGod,\...
82494	9521937	west.beirut.(1998).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x13\x97\x...	1\r\n00:00:06,000 --> 00:00:12,074\r\napi.Open...
82495	9521938	frankenstein.the.true.story.(1973).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00$\x97\x9aV...	1\r\n00:00:01,001 --> 00:00:04,630\r\n(Dramati...
82496	9521940	frankenstein.the.true.story.(1973).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x00\x97\x...	1\r\n00:00:06,000 --> 00:00:12,074\r\nAdvertis...
82497	9521941	zombie.island.massacre.(1984).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00,\x97\x9aV...	1\r\n00:00:01,919 --> 00:00:03,253\r\n(Sharp w...

Using Random 30k Sample

In [15]:
```
# Replace df with the name of your DataFrame if it's different
df = df.sample(n=30000, random_state=1).reset_index(drop=True)
```

In [16]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   num           30000 non-null  int64 
 1   name          30000 non-null  object
 2   content       30000 non-null  object
 3   file_content  30000 non-null  object
dtypes: int64(1), object(3)
memory usage: 937.6+ KB
```

In [17]:
```
df.head()
```

Out[17]:

	num	name	content	file_content
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x86}\x9aV...	ï»¿1\r\n00:00:01,035 --> 00:00:02,536\r\nRILEY...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00o\xa9\x99V...	1\r\n00:00:06,000 --> 00:00:12,074\r\nSupport ...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\xd9\x9b\x...	[Script Info]\r\nTitle: English (US)\r\nOrigin...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	b'PK\x03\x04\x14\x00\x00\x00\x08\x0014\x9aV#\x...	1\r\n00:00:00,870 --> 00:00:02,306\r\n[wind wh...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\n\x95\x9a...	ï»¿1\r\n00:00:06,000 --> 00:00:12,074\r\nWatch...

In [18]:
```
df.head()
```

Out[18]:

	num	name	content	file_content
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x86}\x9aV...	ï»¿1\r\n00:00:01,035 --> 00:00:02,536\r\nRILEY...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00o\xa9\x99V...	1\r\n00:00:06,000 --> 00:00:12,074\r\nSupport ...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\xd9\x9b\x...	[Script Info]\r\nTitle: English (US)\r\nOrigin...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	b'PK\x03\x04\x14\x00\x00\x00\x08\x0014\x9aV#\x...	1\r\n00:00:00,870 --> 00:00:02,306\r\n[wind wh...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\n\x95\x9a...	ï»¿1\r\n00:00:06,000 --> 00:00:12,074\r\nWatch...

In [19]:
```
df.shape
```

Out[19]:
```
(30000, 4)
```

Cleaning the data

Data Prepration - Text Preprocessing

In [20]:
```
import re

# Function to clean text
def clean_text(text):
    # Remove timestamps and other patterns (assuming timestamps are like [00:00:00])
    text = re.sub(r'\[\d+:\d+:\d+\]', '', text)

    # Remove special characters, numbers, and extra spaces
    text = re.sub(r'[^a-zA-Z\s]', '', text)

    # Replace dots with space
    text = re.sub(r'\.', ' ', text)

    # Convert to lowercase
    text = text.lower()

    # Remove unnecessary whitespace characters
    text = re.sub(r'\s+', ' ', text).strip()

    return text
```

In [22]:
```
from tqdm import tqdm
tqdm.pandas()
```

In [23]:
```
# Clean the 'name' and 'file_content' columns
df['cleaned_file_content'] = df['file_content'].progress_apply(clean_text)

# Display the cleaned columns
print(df[['cleaned_file_content']])
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 30000/30000 [03:25<00:00, 146.00it/s]

                                cleaned_file_content
0      riley ipreviously oni the client list derek yo...
1      support us and become vip member to remove all...
2      script info title english us original script p...
3      wind whooshing eerie synth music bats chirping...
4      watch any video online with opensubtitles free...
...                                                  ...
29995  please do not attempt to perform any of these ...
29996  crickets chirring music box music apiopensubti...
29997  i previously oni velma the last place my moms ...
29998  one two three four five numberblocks six seven...
29999  kid capri its the russell simmons def comedy j...

[30000 rows x 1 columns]
```

In [24]:
```
df.head()
```

Out[24]:

	num	name	content	file_content	cleaned_file_content
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\x86}\x9aV...	ï»¿1\r\n00:00:01,035 --> 00:00:02,536\r\nRILEY...	riley ipreviously oni the client list derek yo...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	b'PK\x03\x04\x14\x00\x00\x00\x08\x00o\xa9\x99V...	1\r\n00:00:06,000 --> 00:00:12,074\r\nSupport ...	support us and become vip member to remove all...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\xd9\x9b\x...	[Script Info]\r\nTitle: English (US)\r\nOrigin...	script info title english us original script p...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	b'PK\x03\x04\x14\x00\x00\x00\x08\x0014\x9aV#\x...	1\r\n00:00:00,870 --> 00:00:02,306\r\n[wind wh...	wind whooshing eerie synth music bats chirping...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	b'PK\x03\x04\x14\x00\x00\x00\x08\x00\n\x95\x9a...	ï»¿1\r\n00:00:06,000 --> 00:00:12,074\r\nWatch...	watch any video online with opensubtitles free...

In [25]:
```
df.drop(columns=['content', 'file_content'], inplace=True)
```

In [26]:
```
df.head()
```

Out[26]:

	num	name	cleaned_file_content
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...

In [27]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 3 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   num                   30000 non-null  int64 
 1   name                  30000 non-null  object
 2   cleaned_file_content  30000 non-null  object
dtypes: int64(1), object(2)
memory usage: 703.3+ KB
```

In [28]:
```
df.head()
```

Out[28]:

	num	name	cleaned_file_content
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...

In [1]:
```
# Save the data and Importing
```

In [29]:
```
df.to_csv('clean_data.csv')
```

In [ ]:
```
import pandas as pd
```

In [30]:
```
df1 = pd.read_csv('clean_data.csv')
```

In [31]:
```
df1.shape
```

Out[31]:
```
(30000, 4)
```

In [32]:
```
df1.head()
```

Out[32]:

	Unnamed: 0	num	name	cleaned_file_content
0	0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...
1	1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...
2	2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...
3	3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...
4	4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...

In [33]:
```
df1.drop(columns='Unnamed: 0', inplace =True)
```

In [34]:
```
df1.shape
```

Out[34]:
```
(30000, 3)
```

In [35]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 3 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   num                  30000 non-null  int64 
 1   name                 30000 non-null  object
 2   cleaned_file_content  30000 non-null  object
dtypes: int64(1), object(2)
memory usage: 703.3+ KB
```

Converting Text to Numerical vectors - Word2Vec Representation

Step 1 - Import Word2Vec module from gensim.models

Step 2 - Convert the sentences to the List of Words (i.e. List of Tokens)

Step 3 - Use Word2Vec to learn numerical vectors for each unique words. Word2Vec uses the list of tokens and generate 300Dimensional numerical vector for each unique word.

Step 4 - Convert the word vectors to document vectors.

In [36]:
```
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
```

In [37]:
```
# Download NLTK resources
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\manch\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\manch\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\manch\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[37]: True

In [38]:
```
from nltk.stem import WordNetLemmatizer, PorterStemmer  # Import PorterStemmer

# Initialize the lemmatizer and PorterStemmer
lemmatizer = WordNetLemmatizer()
stemmer = PorterStemmer()
```

In [39]:
```
# Initialize the lemmatizer
lemmatizer = WordNetLemmatizer()

# Define a function to remove stop words, tokenize, and lemmatize text
def tokenize_text(text):
    # Tokenize the text
    tokens = word_tokenize(text)

    # Remove stop words
    stop_words = set(stopwords.words('english'))
    tokens = [token for token in tokens if token.lower() not in stop_words]

    # Lemmatize tokens
    tokens = [lemmatizer.lemmatize(token) for token in tokens]

    return tokens
```

In [ ]:
```
# 15 mins
```

In [40]:
```
# Tokenize, remove stop words, and lemmatize the 'cleaned_file_content' column
df1['tokenised_content_lemma'] = df1['cleaned_file_content'].progress_apply(tokenize_text)

df1['content_size_lemma'] = df1['tokenised_content_lemma'].progress_apply(len)
```

```
100%|██████████████████████████████████████████████████████████████████████| 30000/30000 [13:56<00:00, 35.87it/s]
100%|██████████████████████████████████████████████████████████████████████| 30000/30000 [00:00<00:00, 830171.67it/s]
```

In [41]:
```
# Display the tokenized content
print(df1[['tokenised_content_lemma', 'content_size_lemma']])
```

```
                                tokenised_content_lemma  content_size_lemma
0      [riley, ipreviously, oni, client, list, derek,...                3244
1      [support, u, become, vip, member, remove, ad, ...                2853
2      [script, info, title, english, u, original, sc...                1732
3      [wind, whooshing, eerie, synth, music, bat, ch...                2948
4      [watch, video, online, opensubtitles, free, br...                1482
...                                                  ...                 ...
29995  [please, attempt, perform, stunt, activity, sh...                3055
29996  [cricket, chirring, music, box, music, apiopen...                1805
29997  [previously, oni, velma, last, place, mom, cel...                1797
29998  [one, two, three, four, five, numberblocks, si...                 267
29999  [kid, capri, russell, simmons, def, comedy, ja...                1578

[30000 rows x 2 columns]
```

In [42]:
```
df1.head()
```

Out[42]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	[riley, ipreviously, oni, client, list, derek,...	3244
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	[support, u, become, vip, member, remove, ad, ...	2853
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	[script, info, title, english, u, original, sc...	1732
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	[wind, whooshing, eerie, synth, music, bat, ch...	2948
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	[watch, video, online, opensubtitles, free, br...	1482

chunking the data

In [73]:
```
import pandas as pd

# Assuming df contains a column named 'srt_content' containing the content of each .srt file
# You may need to adjust column names accordingly

# Define a function to create overlapping chunks for each .srt file content
def overlapping_chunks(cleaned_file_content, chunk_size=4, overlap_size=50):
    chunks = []
    lines = cleaned_file_content.split('\n')
    total_lines = len(lines)
    for i in range(0, total_lines - chunk_size + 1, chunk_size - overlap_size):
        chunk = '\n'.join(lines[i:i + chunk_size])
        chunks.append(chunk)
    return chunks

# Apply the overlapping chunks function to each .srt file content in the DataFrame
df1['overlapping_chunks_file_content'] = df1['cleaned_file_content'].apply(lambda x: overlapping_chunks(x))

# Now df['overlapping_chunks'] contains a list of overlapping chunks for each .srt file
# Each chunk is a string containing a part of the .srt content with overlap
df1.head()
```

Out[73]:

	name	file_content	cleaned_file_content	tokenised_content_lemma	content_size_lemma	overlapping_chunks_file_cont
0	the.client.list.s02.e06.unanswered.prayers.(20...	ï»¿1\r\n00:00:01,035 --> 00:00:02,536\r\nRILEY...	riley ipreviously oni the client list derek yo...	[riley, ipreviously, oni, client, list, derek,...	3244	[riley ipreviously oni the client derek
1	the.devil.is.a.woman.(1935).eng.1cd	1\r\n00:00:06,000 --> 00:00:12,074\r\nSupport ...	support us and become vip member to remove all...	[support, u, become, vip, member, remove, ad, ...	2853	[support us and become member to remove
2	kingdom.s04.e18.rivercrossing.battle.(2022).en...	[Script Info]\r\nTitle: English (US)\r\nOrigin...	script info title english us original script p...	[script, info, title, english, u, original, sc...	1732	[script info title english us orig scrip
3	sorority.babes.in.the.slimeball.bowlorama.2.(2...	1\r\n00:00:00,870 --> 00:00:02,306\r\n[wind wh...	wind whooshing eerie synth music bats chirping...	[wind, whooshing, eerie, synth, music, bat, ch...	2948	[wind whooshing eerie synth mu bats chirp
4	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	ï»¿1\r\n00:00:06,000 --> 00:00:12,074\r\nWatch...	watch any video online with opensubtitles free...	[watch, video, online, opensubtitles, free, br...	1482	[watch any video online opensubtitles fi

In [43]:
```
df1.shape
```

Out[43]:
```
(30000, 5)
```

In [44]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   num                     30000 non-null  int64 
 1   name                    30000 non-null  object
 2   cleaned_file_content    30000 non-null  object
 3   tokenised_content_lemma  30000 non-null  object
 4   content_size_lemma      30000 non-null  int64 
dtypes: int64(2), object(3)
memory usage: 1.1+ MB
```

In [45]:
```
df1.to_csv('lemma_data.csv')
```

In [ ]:
```
df1.info()
```

In [ ]:
```
import pandas as pd
```

In [46]:
```
df2 = pd.read_csv('lemma_data.csv')
```

In [47]:
```
df2.shape
```

Out[47]:
```
(30000, 6)
```

In [48]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Unnamed: 0              30000 non-null  int64
 1   num                     30000 non-null  int64
 2   name                    30000 non-null  object
 3   cleaned_file_content    30000 non-null  object
 4   tokenised_content_lemma  30000 non-null  object
 5   content_size_lemma      30000 non-null  int64
dtypes: int64(3), object(3)
memory usage: 1.4+ MB
```

In [49]:
```
df2.head()
```

Out[49]:

	Unnamed: 0	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma
0	0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244
1	1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853
2	2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732
3	3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948
4	4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482

In [51]:
```
df2.drop(columns='Unnamed: 0', inplace =True)
```

In [52]:
```
df2.head()
```

Out[52]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482

In [53]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   num                     30000 non-null  int64 
 1   name                    30000 non-null  object
 2   cleaned_file_content    30000 non-null  object
 3   tokenised_content_lemma  30000 non-null  object
 4   content_size_lemma      30000 non-null  int64 
dtypes: int64(2), object(3)
memory usage: 1.1+ MB
```

In [54]:
```
import gensim

print(gensim.__version__)
from gensim.models import Word2Vec
```

```
4.3.2
```

In [57]:
```
# 15 mins
```

In [56]:
```
model = Word2Vec(list(df2.tokenised_content_lemma), vector_size=300, min_count=1)

print(model)
```

```
Word2Vec<vocab=31, vector_size=300, alpha=0.025>
```

In [58]:
```
# Checking the shape of vectors learned by the model

print(model.wv.__getitem__(model.wv.index_to_key).shape)
```

```
(31, 300)
```

In [60]:
```
def encode_text(text, model):
    # Assuming text is a list of tokens
    embedding = sum(model.wv[token] for token in text if token in model.wv) / len(text)
    return embedding

df2['doc_vector'] = df2.tokenised_content_lemma.progress_apply(lambda x: encode_text(x, model))
```

```
100%|██████████████████████████████████████████████████████████████████████████| 30000/30000 [54:02<00:00,  9.25it/s]
```

In [61]:
```
df2.head()
```

Out[61]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-0.10934329, -0.0768003, -0.18056774, -0.2326...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-0.0983889, -0.075304784, -0.1655197, -0.2400...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-0.048926443, -0.13909523, -0.18452683, -0.28...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-0.098996446, -0.06885647, -0.16654159, -0.23...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199, -0.062561914, -0.15584423, -0.21...

In [63]:
```
df2.shape
```

Out[63]:
```
(30000, 6)
```

In [64]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   num                     30000 non-null  int64 
 1   name                    30000 non-null  object
 2   cleaned_file_content    30000 non-null  object
 3   tokenised_content_lemma  30000 non-null  object
 4   content_size_lemma      30000 non-null  int64 
 5   doc_vector              30000 non-null  object
dtypes: int64(2), object(4)
memory usage: 1.4+ MB
```

In [65]:
```
df2.to_csv('w2v_data.csv')
```

In [ ]:
```

```

Pretrained BERT for Sentence Vectors

In [1]:
```
import pandas as pd
```

In [2]:
```
df3 = pd.read_csv('w2v_data.csv')
```

In [3]:
```
df3.shape
```

Out[3]:
```
(30000, 7)
```

In [4]:
```
df3.head()
```

Out[4]:

	Unnamed: 0	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector
0	0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...
1	1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...
2	2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...
3	3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...
4	4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...

In [5]:
```
df3.drop(columns='Unnamed: 0', inplace =True)
```

In [6]:
```
df3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   num                     30000 non-null  int64 
 1   name                    30000 non-null  object
 2   cleaned_file_content    30000 non-null  object
 3   tokenised_content_lemma  30000 non-null  object
 4   content_size_lemma      30000 non-null  int64 
 5   doc_vector              30000 non-null  object
dtypes: int64(2), object(4)
memory usage: 1.4+ MB
```

In [7]:
```
df3.head()
```

Out[7]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...

In [8]:
```
from sentence_transformers import SentenceTransformer, util

model = SentenceTransformer('all-MiniLM-L6-v2')
```

In [9]:
```
from tqdm import tqdm
tqdm.pandas()
```

In [10]:
```
# from sentence_transformers import SentenceTransformer, util

# model = SentenceTransformer('all-MiniLM-L6-v2')
def encode_text(text):
  embedding = model.encode(text)
  return embedding
```

In [11]:
```
df11 = df3.loc[:2000,:]
df11.head()
```

Out[11]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...

In [12]:
```
df11['sentance_content_bert'] = df11['cleaned_file_content'].progress_apply(encode_text)
df11.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2001/2001 [11:07<00:00,  3.00it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\712963940.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df11['sentance_content_bert'] = df11['cleaned_file_content'].progress_apply(encode_text)
```

Out[12]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [13]:
```
df12 = df3.loc[2001:4000,:]
df12.head()
```

Out[13]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector
2001	9242706	the.yakuzas.guide.to.babysitting.s01.e11.firew...	script info title english us original script b...	['script', 'info', 'title', 'english', 'u', 'o...	1529	[-3.07599939e-02 -1.56279713e-01 -1.56834319e-...
2002	9428559	the.lord.of.the.rings.the.fellowship.of.the.ri...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	5994	[-0.10344059 -0.08081925 -0.16428243 -0.213812...
2003	9271256	bullet.train.(2022).eng.1cd	apiopensubtitlesorg is deprecated please imple...	['apiopensubtitlesorg', 'deprecated', 'please'...	5755	[-8.83942544e-02 -7.16175139e-02 -1.83909342e-...
2004	9297763	blockbuster.s01.e09.thimble.(2022).eng.1cd	all right gang huddle up this is for you guys ...	['right', 'gang', 'huddle', 'guy', 'oh', 'didn...	1989	[-8.56300741e-02 -7.75319487e-02 -1.63196027e-...
2005	9344265	the.most.beautiful.flower.(2022).eng.1cd	netflix presents biasa high school green beans...	['netflix', 'present', 'biasa', 'high', 'schoo...	1513	[-1.04019031e-01 -7.96759427e-02 -1.75259918e-...

In [14]:
```
df12['sentance_content_bert'] = df12['cleaned_file_content'].progress_apply(encode_text)
df12.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2000/2000 [24:19<00:00,  1.37it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\2228528983.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df12['sentance_content_bert'] = df12['cleaned_file_content'].progress_apply(encode_text)
```

Out[14]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
2001	9242706	the.yakuzas.guide.to.babysitting.s01.e11.firew...	script info title english us original script b...	['script', 'info', 'title', 'english', 'u', 'o...	1529	[-3.07599939e-02 -1.56279713e-01 -1.56834319e-...	[-0.0034559942, -0.014797437, 0.07932088, -0.0...
2002	9428559	the.lord.of.the.rings.the.fellowship.of.the.ri...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	5994	[-0.10344059 -0.08081925 -0.16428243 -0.213812...	[-0.06378279, 0.0143626975, -0.05514103, -0.05...
2003	9271256	bullet.train.(2022).eng.1cd	apiopensubtitlesorg is deprecated please imple...	['apiopensubtitlesorg', 'deprecated', 'please'...	5755	[-8.83942544e-02 -7.16175139e-02 -1.83909342e-...	[-0.03831797, -0.07563398, 0.009899806, 0.0112...
2004	9297763	blockbuster.s01.e09.thimble.(2022).eng.1cd	all right gang huddle up this is for you guys ...	['right', 'gang', 'huddle', 'guy', 'oh', 'didn...	1989	[-8.56300741e-02 -7.75319487e-02 -1.63196027e-...	[-0.13685279, 0.02369652, -0.0012757194, -0.01...
2005	9344265	the.most.beautiful.flower.(2022).eng.1cd	netflix presents biasa high school green beans...	['netflix', 'present', 'biasa', 'high', 'schoo...	1513	[-1.04019031e-01 -7.96759427e-02 -1.75259918e-...	[-0.0951997, -0.08066843, 0.031748097, -0.0581...

In [15]:
```
new_df = pd.concat([df11,df12],ignore_index=True, axis=0)
new_df.head()
```

Out[15]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [16]:
```
df13 = df3.loc[4001:6000,:]
df13['sentance_content_bert'] = df13['cleaned_file_content'].progress_apply(encode_text)
df13.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████| 2000/2000 [30:20<00:00,  1.10it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\2024776208.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df13['sentance_content_bert'] = df13['cleaned_file_content'].progress_apply(encode_text)
```

Out[16]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
4001	9316759	one.tree.hill.s09.e02.in.the.room.where.you.sl...	ipreviously oni one tree hill i had a fire at ...	['ipreviously', 'oni', 'one', 'tree', 'hill', ...	2572	[-0.10138439 -0.07383844 -0.17823057 -0.231539...	[0.0145669775, -0.061377663, -0.022976166, -0....
4002	9253363	pioneers.of.television.s01.e02.late.night.(200...	let me close my eyes and just every little inf...	['let', 'close', 'eye', 'every', 'little', 'in...	3876	[-9.47634354e-02 -7.94382989e-02 -1.65449485e-...	[-0.054164283, -0.0451637, 0.004766886, -0.061...
4003	9341779	army.wives.s04.e02.scars.and.stripes.(2010).en...	ipreviously oni army wives our son tried to ki...	['ipreviously', 'oni', 'army', 'wife', 'son', ...	2643	[-1.04013450e-01 -7.13478178e-02 -1.65078789e-...	[-0.08255247, -0.059647605, 0.09548979, -0.004...
4004	9460559	silk.stalkings.s02.e17.night.games.(1993).eng.1cd	whats her story wild turkey straight up she a ...	['whats', 'story', 'wild', 'turkey', 'straight...	2806	[-9.82831270e-02 -7.54871890e-02 -1.85113847e-...	[-0.07851616, -0.019380031, 0.029582461, 0.021...
4005	9451172	grand.designs.s03.e08.the.traditional.cottage....	its a bit scary that its all our money its eve...	['bit', 'scary', 'money', 'everything', 'could...	3748	[-1.03748016e-01 -7.13696256e-02 -1.56368300e-...	[-0.025606273, -0.02973921, 0.043111995, -0.06...

In [17]:
```
new_df = pd.concat([new_df,df13],ignore_index=True, axis=0)
new_df.head()
```

Out[17]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [18]:
```
df14 = df3.loc[6001:8000,:]
df14['sentance_content_bert'] = df14['cleaned_file_content'].progress_apply(encode_text)
df14.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████| 2000/2000 [04:47<00:00,  6.96it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\2276790312.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df14['sentance_content_bert'] = df14['cleaned_file_content'].progress_apply(encode_text)
```

Out[18]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
6001	9208014	impractical.jokers.s09.e26.brooke.shields.(202...	laughter screams lets just get this started ar...	['laughter', 'scream', 'let', 'get', 'started'...	1703	[-0.09946816 -0.07624239 -0.17200854 -0.233802...	[-0.07807629, -0.02912761, 0.049470894, 0.0248...
6002	9487219	invisible.s01.e05.episode.1.5.(2022).eng.1cd	antakeiri just confessed to the murder of emi ...	['antakeiri', 'confessed', 'murder', 'emi', 's...	1909	[-7.92224407e-02 -7.66554773e-02 -1.37319043e-...	[-0.08749884, 0.015282028, -0.049739387, -0.01...
6003	9392290	gentleman.(2020).eng.1cd	resynched by nonsequitorseries once again my h...	['resynched', 'nonsequitorseries', 'heart', 'f...	88	[-9.16472226e-02 -7.39626810e-02 -1.43047005e-...	[-0.06800948, -0.021019911, 0.078424126, 0.002...
6004	9468301	csi.miami.s03.e19.sex.taxes.(2005).eng.1cd	hey get away from my boat thief youre not gonn...	['hey', 'get', 'away', 'boat', 'thief', 'youre...	2172	[-9.53904092e-02 -7.64079690e-02 -1.72718689e-...	[-0.10429397, 0.005755268, -0.015018639, -0.06...
6005	9463846	our.blooming.youth.s01.e10.episode.1.10.(2023)...	park hyungsik jeon sonee pyo yejin yun jongseo...	['park', 'hyungsik', 'jeon', 'sonee', 'pyo', '...	2723	[-0.08806863 -0.07335749 -0.16878524 -0.218788...	[-0.08385247, 0.04649207, 0.10863581, 0.006626...

In [19]:
```
new_df = pd.concat([new_df,df14],ignore_index=True, axis=0)
new_df.head()
```

Out[19]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [20]:
```
df15 = df3.loc[8001:10000,:]
df15['sentance_content_bert'] = df15['cleaned_file_content'].progress_apply(encode_text)
df15.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2000/2000 [08:24<00:00,  3.97it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\845179494.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df15['sentance_content_bert'] = df15['cleaned_file_content'].progress_apply(encode_text)
```

Out[20]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
8001	9351448	the.banshees.of.inisherin.(2022).eng.1cd	apiopensubtitlesorg is deprecated please imple...	['apiopensubtitlesorg', 'deprecated', 'please'...	4459	[-8.64537656e-02 -9.04118344e-02 -1.79184362e-...	[0.0032120578, -0.09480547, 0.052504268, -0.06...
8002	9401841	american.dad.s04.e15.wife.insurance.(2009).eng...	good morning usa i got a feeling that its gonn...	['good', 'morning', 'usa', 'got', 'feeling', '...	1585	[-0.08580665 -0.07550405 -0.16680738 -0.235261...	[-0.013176337, -0.031332716, 0.04096018, 0.041...
8003	9256158	house.of.wax.(1953).eng.1cd	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	3869	[-9.80368853e-02 -6.34661317e-02 -1.57282323e-...	[-0.02410146, -0.031204488, 0.03174951, -0.112...
8004	9508645	kojak.s02.e03.hush.now.dont.you.die.(1974).eng...	music music music music music skip class run a...	['music', 'music', 'music', 'music', 'music', ...	2861	[-0.09006649 -0.06592161 -0.17272265 -0.235484...	[-0.050546892, -0.014168973, 0.061474536, -0.0...
8005	9219633	this.is.us.s02.e16.vegas.baby.(2018).eng.1cd	i previously oni this is us you know when i wa...	['previously', 'oni', 'u', 'know', 'kid', 'obs...	3002	[-1.10897392e-01 -7.63068497e-02 -1.73733607e-...	[-0.110696316, -0.10000415, 0.0813859, -0.0829...

In [21]:
```
new_df = pd.concat([new_df,df15],ignore_index=True, axis=0)
new_df.head()
```

Out[21]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [22]:
```
df16 = df3.loc[10001:12000,:]
df16['sentance_content_bert'] = df16['cleaned_file_content'].progress_apply(encode_text)
df16.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████| 2000/2000 [04:45<00:00,  7.01it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\1748578328.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df16['sentance_content_bert'] = df16['cleaned_file_content'].progress_apply(encode_text)
```

Out[22]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content
10001	9208054	hotties.s01.e03.light.my.cherries.(2022).eng.1cd	oh hey its jade or as im known by my peers why...	['oh', 'hey', 'jade', 'im', 'known', 'peer', '...	1854	[-1.02389939e-01 -7.64683858e-02 -1.81584641e-...	[-0.0595 -0.083 -0.010130153, -0.
10002	9367990	this.is.gwar.(2021).eng.1cd	birds chirping light scraping vehicles passing...	['bird', 'chirping', 'light', 'scraping', 'veh...	9308	[-7.90914867e-02 -7.62532208e-02 -1.71830501e-...	[-0.00912 0.006967 0.07747704, -(
10003	9267940	welcome.to.demonschool.irumakun.s01.e19.i.want...	script info title english us original script n...	['script', 'info', 'title', 'english', 'u', 'o...	1908	[0.05002058 -0.1523164 -0.13656582 -0.299883...	[0.001647 -0.02871 0.060121845,
10004	9414144	legion.of.superheroes.(2023).eng.1cd	apiopensubtitlesorg is deprecated please imple...	['apiopensubtitlesorg', 'deprecated', 'please'...	3278	[-0.07513556 -0.07696131 -0.14968702 -0.229555...	[-0.06155 -0.06403 0.037950598,
10005	9196507	the.chi.s05.e07.angels.(2022).eng.1cd	apiopensubtitlesorg is deprecated please imple...	['apiopensubtitlesorg', 'deprecated', 'please'...	2791	[-0.10215868 -0.07280289 -0.19138134 -0.231998...	[-0.03 -0.05293 0.012356777, -0.0

In [23]:
```
new_df = pd.concat([new_df,df16],ignore_index=True, axis=0)
new_df.head()
```

Out[23]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [24]:
```
df17 = df3.loc[12001:14000,:]
df17['sentance_content_bert'] = df17['cleaned_file_content'].progress_apply(encode_text)
df17.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2000/2000 [04:42<00:00,  7.07it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\3909431988.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df17['sentance_content_bert'] = df17['cleaned_file_content'].progress_apply(encode_text)
```

Out[24]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_ber
12001	9429954	the.office.s06.e08.koi.pond.(2009).eng.1cd	and then i think im going to go to the garlic ...	['think', 'im', 'going', 'go', 'garlic', 'fest...	2078	[-0.09583216 -0.08166428 -0.18516263 -0.241945...	[-0.03761893 -0.008056768 0.04135403, -0.071..
12002	9191704	madam.secretary.s01.e09.so.it.goes.(2014).eng.1cd	the framework for the next phase of the talks ...	['framework', 'next', 'phase', 'talk', 'place'...	2807	[-8.05756450e-02 -6.79903850e-02 -1.45589471e-...	[-0.015036208 0.02320732 0.021067487, 0.0125..
12003	9460596	silk.stalkings.s04.e08.ask.the.dust.(1994).eng...	apiopensubtitlesorg is deprecated please imple...	['apiopensubtitlesorg', 'deprecated', 'please'...	2148	[-0.09280361 -0.07814712 -0.15919352 -0.234746...	[0.0006849572 0.005551415 0.039690502, -0.01..
12004	9198778	empire.s02.e11.death.will.have.his.day.(2016)....	ipreviously oni empire when hakeem wins ii get...	['ipreviously', 'oni', 'empire', 'hakeem', 'wi...	2505	[-9.70948935e-02 -8.11146349e-02 -1.73993245e-...	[-0.04885235 -0.076392196 0.005214465, -0.15..
12005	9353400	the.santa.clauses.s01.e06.chapter.six.a.christ...	simon reading twas the night before christmas ...	['simon', 'reading', 'twas', 'night', 'christm...	1694	[-1.05213828e-01 -6.94110617e-02 -1.71953604e-...	[-0.0825478 -0.046360213 0.040324092, -0.076..

In [25]:
```
new_df = pd.concat([new_df,df17],ignore_index=True, axis=0)
new_df.head()
```

Out[25]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [26]:
```
df18 = df3.loc[14001:16000,:]
df18['sentance_content_bert'] = df18['cleaned_file_content'].progress_apply(encode_text)
df18.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2000/2000 [07:54<00:00,  4.22it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\2205419664.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df18['sentance_content_bert'] = df18['cleaned_file_content'].progress_apply(encode_text)
```

Out[26]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_
14001	9454041	as.time.goes.by.s01.e01.you.must.remember.this...	theme song iyou must remember thisi ia kiss is...	['theme', 'song', 'iyou', 'must', 'remember', ...	1516	[-0.0951754 -0.06677904 -0.15973537 -0.233513...	[-0.09709 -0.042123 -0.000746434, -(
14002	9242226	vampire.academy.s01.e02.earth.air.water.fire.(...	apiopensubtitlesorg is deprecated please imple...	['apiopensubtitlesorg', 'deprecated', 'please'...	1948	[-8.05898607e-02 -7.44136274e-02 -1.43815517e-...	[-0.07854 -0.059621 -0.0029286009, -
14003	9348661	unforgettable.s04.e12.bad.company.(2016).eng.1cd	cant get a bite to save my life thats cause yo...	['cant', 'get', 'bite', 'save', 'life', 'thats...	2774	[-9.40676257e-02 -7.02347010e-02 -1.66766748e-...	[-0.13495 0.11449 0.073383726, -0.01
14004	9499292	byeonhyeokui.sarang.s01.e05.episode.1.5.(2017)...	anwhats going on anjun what brings you by anju...	['anwhats', 'going', 'anjun', 'brings', 'anjun...	2555	[-0.08638296 -0.07009944 -0.1701365 -0.235732...	[-0.075270 0.016777 0.0357354, -0.02
14005	9429873	futurama.s03.e13.bendin.in.the.wind.(2001).eng...	iman one two one two threei advertise your pro...	['iman', 'one', 'two', 'one', 'two', 'threei',...	1505	[-8.59041363e-02 -7.70007297e-02 -1.64791748e-...	[-0.08019 0.0023510 -0.018381495, -(

In [27]:
```
new_df = pd.concat([new_df,df18],ignore_index=True, axis=0)
new_df.head()
```

Out[27]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [28]:
```
df19 = df3.loc[16001:18000,:]
df19['sentance_content_bert'] = df19['cleaned_file_content'].progress_apply(encode_text)
df19.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2000/2000 [12:38<00:00,  2.64it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\2336338799.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df19['sentance_content_bert'] = df19['cleaned_file_content'].progress_apply(encode_text)
```

Out[28]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
16001	9420862	hot.in.cleveland.s02.e07.dog.tricks.sex.flicks...	i hot in clevelandi is recorded in front of a ...	['hot', 'clevelandi', 'recorded', 'front', 'li...	1456	[-1.02737613e-01 -7.21664950e-02 -1.70121193e-...	[-0.10080973, -0.047080465, 0.008609233, 0.000...
16002	9505136	rise.fall.s01.e14.episode.1.14.(2023).eng.1cd	this programme contains strong language strang...	['programme', 'contains', 'strong', 'language'...	3500	[-1.03391163e-01 -6.51381239e-02 -1.68129444e-...	[-0.043388773, -0.09597999, 0.0344118, -0.0732...
16003	9499253	ghosts.s02.e05.bump.in.the.night.(2020).eng.1cd	script info title default file scripttype v wr...	['script', 'info', 'title', 'default', 'file',...	2711	[1.15722828e-01 -2.96496063e-01 -2.40038857e-...	[-0.013964106, -0.038802456, 0.027211094, 0.01...
16004	9397271	outlander.s04.e06.blood.of.my.blood.(2018).eng...	claire ipreviouslyi weve chosen a name baby kl...	['claire', 'ipreviouslyi', 'weve', 'chosen', '...	2522	[-8.89287144e-02 -6.19221479e-02 -1.43336743e-...	[-0.085511304, -0.0071783094, 0.03698805, 0.00...
16005	9270581	padre.no.hay.mas.que.uno.3.(2022).eng.1cd	advertise your product or brand here contact w...	['advertise', 'product', 'brand', 'contact', '...	6099	[-0.10184556 -0.06776034 -0.1644768 -0.228548...	[-0.031446703, -0.00909334, 0.018630132, -0.06...

In [29]:
```
new_df = pd.concat([new_df,df19],ignore_index=True, axis=0)
new_df.head()
```

Out[29]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [30]:
```
df20 = df3.loc[18001:20000,:]
df20['sentance_content_bert'] = df20['cleaned_file_content'].progress_apply(encode_text)
df20.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2000/2000 [04:48<00:00,  6.93it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\346014879.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df20['sentance_content_bert'] = df20['cleaned_file_content'].progress_apply(encode_text)
```

Out[30]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_b
18001	9367904	90210.s02.e14.girl.fight.(2010).eng.1cd	well this has been a productive afternoon yeah...	['well', 'productive', 'afternoon', 'yeah', 'i...	2409	[-1.15172960e-01 -6.62367940e-02 -1.75372392e-...	[-0.086169(-0.0199698 0.07527088, -0.05
18002	9399171	bungo.stray.dogs.s04.e03.the.secret.founding.o...	script info title english us original script k...	['script', 'info', 'title', 'english', 'u', 'o...	2327	[-0.04008507 -0.12580973 -0.16890977 -0.250085...	[-0.00670030(-0.0435957(0.111013696, 0.(
18003	9421324	taboo.s01.e01.episode.1.1.(2017).eng.1cd	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1865	[-9.49173197e-02 -6.58254698e-02 -1.42497107e-...	[-0.0520514 0.0186769 -0.0930824, -0.11
18004	9450399	quantico.s01.e07.go.(2015).eng.1cd	script info title default file scripttype v wr...	['script', 'info', 'title', 'default', 'file',...	5207	[1.08780330e-01 -2.88579018e-01 -2.39247512e-...	[-0.035149(-0.03576(0.013126599, -0.014
18005	9399775	women.at.war.s01.e04.episode.4.(2022).eng.1cd	netflix presents come on one more step were al...	['netflix', 'present', 'come', 'one', 'step', ...	1291	[-0.09539285 -0.06760713 -0.15721595 -0.226088...	[-0.062581(-0.070646 -0.04207025, -0.05(

In [31]:
```
new_df = pd.concat([new_df,df20],ignore_index=True, axis=0)
new_df.head()
```

Out[31]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [32]:
```
df21 = df3.loc[20001:22000,:]
df21['sentance_content_bert'] = df21['cleaned_file_content'].progress_apply(encode_text)
df21.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2000/2000 [04:45<00:00,  7.01it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\90145499.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df21['sentance_content_bert'] = df21['cleaned_file_content'].progress_apply(encode_text)
```

Out[32]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
20001	9231080	flatland.s01.e05.a.gentle.rain.(2002).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	1622	[-0.10122931 -0.07119595 -0.17184035 -0.226168...	[-0.09338511, -0.04040897, -0.010522487, -0.05...
20002	9332116	battle.for.saipan.(2022).eng.1cd	use the free code joinnow at wwwplayshipseu dr...	['use', 'free', 'code', 'joinnow', 'wwwplayshi...	2936	[-0.08494363 -0.06882316 -0.16988894 -0.232197...	[-0.08687571, -0.07450393, 0.060404934, -0.074...
20003	9362436	the.great.s01.e09.love.hurts.(2020).eng.1cd	apiopensubtitlesorg is deprecated please imple...	['apiopensubtitlesorg', 'deprecated', 'please'...	2764	[-0.08708873 -0.07407045 -0.16646565 -0.235027...	[-0.025401864, -0.029697467, 0.023453519, 0.00...
20004	9495498	tiny.beautiful.things.s01.e07.go.(2023).eng.1cd	advertise your product or brand here contact w...	['advertise', 'product', 'brand', 'contact', '...	1903	[-0.11263889 -0.07305092 -0.17696753 -0.229608...	[-0.1058409, -0.09009985, 0.022006258, -0.0389...
20005	9210526	survivor.s20.e10.going.down.in.flames.(2010).e...	jeff probst previously on isurvivori with thei...	['jeff', 'probst', 'previously', 'isurvivori',...	2928	[-0.10461924 -0.07693632 -0.1731256 -0.227250...	[-0.07107766, -0.028268559, 0.009724177, -0.07...

In [33]:
```
new_df = pd.concat([new_df,df21],ignore_index=True, axis=0)
new_df.head()
```

Out[33]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [34]:
```
df22 = df3.loc[22001:24000,:]
df22['sentance_content_bert'] = df22['cleaned_file_content'].progress_apply(encode_text)
df22.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████| 2000/2000 [04:47<00:00,  6.95it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\92166505.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df22['sentance_content_bert'] = df22['cleaned_file_content'].progress_apply(encode_text)
```

Out[34]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_l
22001	9481052	blue.lock.s01.e24.the.time.has.come.(2023).eng...	script info title english us scaledborderandsh...	['script', 'info', 'title', 'english', 'u', 's...	2431	[-3.25537324e-02 -1.57744169e-01 -1.80679888e-...	[-0.024049 -0.042097 0.07836842, -0.0
22002	9287989	new.amsterdam.s05.e06.give.me.a.sign.(2022).en...	previously on new amsterdam sighs dramatic mus...	['previously', 'new', 'amsterdam', 'sigh', 'dr...	2754	[-9.20306742e-02 -7.07242861e-02 -1.68428078e-...	[0.021759 -0.085767 0.098071024, -0.0
22003	9376188	secrets.of.summer.s02.e04.open.sky.(2022).eng.1cd	a netflix series tony wont you say goodbye to ...	['netflix', 'series', 'tony', 'wont', 'say', '...	1505	[-9.78885069e-02 -7.20411688e-02 -1.58579871e-...	[-0.008194 -0.072017 0.086092174, -0.0
22004	9271186	the.amazing.race.canada.s08.e10.is.that.a.wild...	previously on the amazing race canada bathurst...	['previously', 'amazing', 'race', 'canada', 'b...	3255	[-1.10624880e-01 -6.91541806e-02 -1.66781545e-...	[-0.0739(-0.003904(0.0401196, -0.049
22005	9443359	mom.s05.e18.spaghetti.sauce.and.a.dumpster.fir...	script info title default file scripttype v wr...	['script', 'info', 'title', 'default', 'file',...	2737	[1.19423978e-01 -3.00981909e-01 -2.44951963e-...	[-0.03568 -0.028523 0.00823984, 0.00

In [35]:
```
new_df = pd.concat([new_df,df22],ignore_index=True, axis=0)
new_df.head()
```

Out[35]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [36]:
```
df23 = df3.loc[24001:26000,:]
df23['sentance_content_bert'] = df23['cleaned_file_content'].progress_apply(encode_text)
df23.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████| 2000/2000 [05:04<00:00,  6.56it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\3941159069.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df23['sentance_content_bert'] = df23['cleaned_file_content'].progress_apply(encode_text)
```

Out[36]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
24001	9260231	glomdalsbruden.(1926).eng.1cd	the bride of glomdalen adapted from the novel ...	['bride', 'glomdalen', 'adapted', 'novel', 'ja...	624	[-1.08158268e-01 -5.45340367e-02 -1.45908609e-...	[-0.077882946 -0.009881992 0.05770749, -0.03..
24002	9450146	the.young.riders.s02.e08.requiem.for.a.hero.(1...	this place gets bigger every time i come here ...	['place', 'get', 'bigger', 'every', 'time', 'c...	1649	[-0.09585558 -0.08901681 -0.17811936 -0.239453...	[-0.048022598 -0.046969272 0.013737078, -0.0..
24003	9345394	garireo.s02.e03.episode.2.3.(2013).eng.1cd	script info script generated by aegisub httpww...	['script', 'info', 'script', 'generated', 'aeg...	3679	[-8.10206458e-02 -1.23595968e-01 -1.52987540e-...	[0.04607229 -0.0291352 -0.05017469 -0.01454..
24004	9186989	rubeusutori.in.habeodeu.s01.e13.episode.1.13.(...	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	1445	[-1.04831234e-01 -6.63323551e-02 -1.71677843e-...	[-0.058432914 0.0032893969 0.023912683, 0.03..
24005	9376687	iki.wo.hisomete.s01.e08.episode.1.8.(2021).eng...	script info script generated by aegisub httpww...	['script', 'info', 'script', 'generated', 'aeg...	1317	[-3.91445979e-02 -1.51133835e-01 -1.86584115e-...	[0.013775244 -0.03558014 -0.040352218, -0.04..

In [37]:
```
new_df = pd.concat([new_df,df23],ignore_index=True, axis=0)
new_df.head()
```

Out[37]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [38]:
```
df24 = df3.loc[26001:28000,:]
df24['sentance_content_bert'] = df24['cleaned_file_content'].progress_apply(encode_text)
df24.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 2000/2000 [04:44<00:00,  7.04it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\980672348.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df24['sentance_content_bert'] = df24['cleaned_file_content'].progress_apply(encode_text)
```

Out[38]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
26001	9452000	ghosts.s02.e16.isaacs.book.(2023).eng.1cd	in summary while this book may be about a forg...	['summary', 'book', 'may', 'forgotten', 'found...	1894	[-8.79711285e-02 -8.01629499e-02 -1.72001362e-...	[-0.055523966, -0.015167514, 0.0048165806, -0....
26002	9348779	a.womans.vengeance.(1948).eng.1cd	bell clanging watch any video online with open...	['bell', 'clanging', 'watch', 'video', 'online...	6085	[-9.71338451e-02 -6.82802349e-02 -1.57850355e-...	[-0.040452708, -0.037264597, -0.007890472, -0....
26003	9210966	tears.on.fire.s01.e09.the.truth.(2021).eng.1cd	if you think i dont support you enough as your...	['think', 'dont', 'support', 'enough', 'chief'...	2027	[-9.88564119e-02 -6.25671074e-02 -1.43308371e-...	[-0.016954886, 0.071878165, 0.11949639, 0.0173...
26004	9348716	murder.one.s01.e19.chapter.nineteen.(1996).eng...	ithe victimi murder in hollywood last night yo...	['ithe', 'victimi', 'murder', 'hollywood', 'la...	3286	[-8.29386190e-02 -7.43170679e-02 -1.44189298e-...	[-0.06130754, -0.018050581, 0.031040868, -0.05...
26005	9271741	scrubs.s05.e10.her.story.ii.(2006).eng.1cd	the turtlesi happy togetheri ijd behold julie ...	['turtlesi', 'happy', 'togetheri', 'ijd', 'beh...	1652	[-8.85888562e-02 -7.68031478e-02 -1.70708701e-...	[-0.06376065, 0.021168213, 0.03759763, -0.0556...

In [39]:
```
new_df = pd.concat([new_df,df24],ignore_index=True, axis=0)
new_df.head()
```

Out[39]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [40]:
```
df25 = df3.loc[28001:,:]
df25['sentance_content_bert'] = df25['cleaned_file_content'].progress_apply(encode_text)
df25.head()
```

```
100%|██████████████████████████████████████████████████████████████████████████| 1999/1999 [04:45<00:00,  6.99it/s]
C:\Users\manch\AppData\Local\Temp\ipykernel_9688\1861430132.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df25['sentance_content_bert'] = df25['cleaned_file_content'].progress_apply(encode_text)
```

Out[40]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
28001	9348327	thriller.s01.e08.file.it.under.fear.(1973).eng...	bright orchestra music eerie music apiopensubt...	['bright', 'orchestra', 'music', 'eerie', 'mus...	3370	[-9.55126733e-02 -7.44086653e-02 -1.63964465e-...	[-0.12684873, -0.07917543, 0.059293527, -0.030...
28002	9406119	american.dad.s13.e19.top.of.the.steve.(2019).e...	rain day rain day todays the day we count the ...	['rain', 'day', 'rain', 'day', 'today', 'day',...	1636	[-1.01312645e-01 -6.04631975e-02 -1.55363187e-...	[-0.06302291, -0.031551152, 0.110944584, 0.014...
28003	9512932	mrs.davis.s01.e01.episode.1.1.(2023).eng.1cd	isomber orchestral musici i i crowd jeering i ...	['isomber', 'orchestral', 'musici', 'crowd', '...	2128	[-0.08635066 -0.07564896 -0.16523147 -0.231817...	[-0.06632358, -0.10276244, 0.06814727, -0.1523...
28004	9510879	criminal.minds.s02.e20.honor.among.thieves.(20...	isuspenseful theme playingi in russian bones c...	['isuspenseful', 'theme', 'playingi', 'russian...	2580	[-8.08129609e-02 -8.39678049e-02 -1.55411586e-...	[-0.062118325, 0.038890865, -0.06073351, -0.03...
28005	9236051	growing.up.s01.e08.episode.1.8.(2022).eng.1cd	igrowing up is so many thingsi iwell get throu...	['igrowing', 'many', 'thingsi', 'iwell', 'get'...	1450	[-9.10296142e-02 -9.31662172e-02 -1.52700618e-...	[-0.054308116, -0.02300869, 0.020534001, 0.040...

In [41]:
```
new_df = pd.concat([new_df,df25],ignore_index=True, axis=0)
new_df.head()
```

Out[41]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....

In [42]:
```
new_df.shape
```

Out[42]:
```
(30000, 7)
```

In [43]:
```
new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   num                      30000 non-null  int64 
 1   name                     30000 non-null  object
 2   cleaned_file_content     30000 non-null  object
 3   tokenised_content_lemma  30000 non-null  object
 4   content_size_lemma       30000 non-null  int64 
 5   doc_vector               30000 non-null  object
 6   sentance_content_bert    30000 non-null  object
dtypes: int64(2), object(5)
memory usage: 1.6+ MB
```

In [44]:
```
new_df.to_csv('bert_data.csv')
```

chunking the data

In [45]:
```
# Example function to chunk a document
def chunk_document(document, chunk_size, overlap):
    chunks = []
    start_idx = 0
    while start_idx < len(document):
        end_idx = min(start_idx + chunk_size, len(document))
        chunks.append(document[start_idx:end_idx])
        start_idx += chunk_size - overlap
    return chunks

# Parameters
chunk_size = 500  # Set the chunk size
overlap = 100      # Set the overlap size

# Apply chunking to content
new_df['content_chunks'] = new_df['cleaned_file_content'].apply(lambda x: chunk_document(x, chunk_size, overlap))
```

Final data set like below

In [46]:
```
new_df.head()
```

Out[46]:

	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	sentance_content_bert	co
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-0.11649825, -0.044796105, 0.026207969, -0.10...	[r o
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-0.017634636, -0.06649098, 0.008615398, -0.07...	[
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-0.017091548, -0.044068865, 0.04417335, -0.04...	c
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-0.06590253, -0.02547965, 0.048389476, -0.016...	[w
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-0.068130724, -0.024480091, -0.025265323, -0....	[w

In [47]:
```
new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   num                      30000 non-null  int64 
 1   name                     30000 non-null  object
 2   cleaned_file_content     30000 non-null  object
 3   tokenised_content_lemma  30000 non-null  object
 4   content_size_lemma       30000 non-null  int64 
 5   doc_vector               30000 non-null  object
 6   sentance_content_bert    30000 non-null  object
 7   content_chunks           30000 non-null  object
dtypes: int64(2), object(6)
memory usage: 1.8+ MB
```

In [48]:
```
new_df.to_csv('final_data.csv')
```

Using the final data set

In [2]:
```
import pandas as pd
```

In [3]:
```
final_df = pd.read_csv('final_data.csv')
```

In [4]:
```
final_df.head
```

Out[4]:
```
<bound method NDFrame.head of        Unnamed: 0      num                                                name  \
0               0  9487821  the.client.list.s02.e06.unanswered.prayers.(20...
1               1  9183678                the.devil.is.a.woman.(1935).eng.1cd
2               2  9258855  kingdom.s04.e18.rivercrossing.battle.(2022).en...
3               3  9421238  sorority.babes.in.the.slimeball.bowlorama.2.(2...
4               4  9518525  ada.twist.scientist.s03.e03.dadbotinto.the.bra...
...           ...      ...                                                ...
29995       29995  9473109  king.of.the.road.s03.e03.chained.up.and.trippe...
29996       29996  9487995                             badoet.(2015).eng.1cd
29997       29997  9400464          velma.s01.e03.velma.kai.(2023).eng.1cd
29998       29998  9489287       numberblocks.s03.e26.thirteen.(2019).eng.1cd
29999       29999  9305337  def.comedy.jam.s02.e09.episode.2.9.(1992).eng.1cd

                                  cleaned_file_content  \
0      riley ipreviously oni the client list derek yo...
1      support us and become vip member to remove all...
2      script info title english us original script p...
3      wind whooshing eerie synth music bats chirping...
4      watch any video online with opensubtitles free...
...                                                  ...
29995  please do not attempt to perform any of these ...
29996  crickets chirring music box music apiopensubti...
29997  i previously oni velma the last place my moms ...
29998  one two three four five numberblocks six seven...
29999  kid capri its the russell simmons def comedy j...

                                tokenised_content_lemma  content_size_lemma  \
0      ['riley', 'ipreviously', 'oni', 'client', 'lis...                3244
1      ['support', 'u', 'become', 'vip', 'member', 'r...                2853
2      ['script', 'info', 'title', 'english', 'u', 'o...                1732
3      ['wind', 'whooshing', 'eerie', 'synth', 'music...                2948
4      ['watch', 'video', 'online', 'opensubtitles', ...                1482
...                                                  ...                 ...
29995  ['please', 'attempt', 'perform', 'stunt', 'act...                3055
29996  ['cricket', 'chirring', 'music', 'box', 'music...                1805
29997  ['previously', 'oni', 'velma', 'last', 'place'...                1797
29998  ['one', 'two', 'three', 'four', 'five', 'numbe...                 267
29999  ['kid', 'capri', 'russell', 'simmons', 'def', ...                1578

                                           doc_vector  \
0      [-1.09343290e-01 -7.68003017e-02 -1.80567741e-...
1      [-9.83889028e-02 -7.53047839e-02 -1.65519699e-...
2      [-4.89264429e-02 -1.39095232e-01 -1.84526831e-...
3      [-9.89964455e-02 -6.88564703e-02 -1.66541591e-...
4      [-0.10334199 -0.06256191 -0.15584423 -0.213279...
...                                                  ...
```

```
29995  [-1.06857479e-01 -7.70704076e-02 -1.95597991e-...
29996  [-0.06744428 -0.07895508 -0.16554289 -0.238525...
29997  [-0.09154191 -0.08225142 -0.17166042 -0.234753...
29998  [-9.71302763e-02 -2.97208223e-02 -1.45906299e-...
29999  [-0.08471369 -0.07977783 -0.18513532 -0.242066...

                                  sentance_content_bert  \
0      [-1.16498247e-01 -4.47961055e-02  2.62079686e-...
1      [-1.76346358e-02 -6.64909780e-02  8.61539785e-...
2      [-1.70915481e-02 -4.40688655e-02  4.41733487e-...
3      [-6.59025311e-02 -2.54796501e-02  4.83894758e-...
4      [-6.81307241e-02 -2.44800914e-02 -2.52653230e-...
...                                                  ...
29995  [ 7.56419310e-03 -5.54080568e-02  2.06312984e-...
29996  [ 1.17038190e-02 -2.66452506e-02  2.01791693e-...
29997  [-8.32793862e-02 -6.34732619e-02  5.84286172e-...
29998  [-8.33337158e-02 -1.16495080e-01 -2.15825532e-...
29999  [-4.05492522e-02 -7.91372731e-02  1.41583364e-...

                                         content_chunks
0      ['riley ipreviously oni the client list derek ...
1      ['support us and become vip member to remove a...
2      ['script info title english us original script...
3      ['wind whooshing eerie synth music bats chirpi...
4      ['watch any video online with opensubtitles fr...
...                                                  ...
29995  ['please do not attempt to perform any of thes...
29996  ['crickets chirring music box music apiopensub...
29997  ['i previously oni velma the last place my mom...
29998  ['one two three four five numberblocks six sev...
29999  ['kid capri its the russell simmons def comedy...

[30000 rows x 9 columns]>
```

In [5]:
```
final_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Unnamed: 0               30000 non-null  int64 
 1   num                      30000 non-null  int64 
 2   name                     30000 non-null  object
 3   cleaned_file_content     30000 non-null  object
 4   tokenised_content_lemma  30000 non-null  object
 5   content_size_lemma       30000 non-null  int64 
 6   doc_vector               30000 non-null  object
 7   sentance_content_bert    30000 non-null  object
 8   content_chunks           30000 non-null  object
dtypes: int64(3), object(6)
memory usage: 2.1+ MB
```

In [6]:
```
final_df = final_df.rename(columns={'sentance_content_bert': 'embeddings'})
```

In [7]:
```
final_df.head()
```

Out[7]:

	Unnamed: 0	num	name	cleaned_file_content	tokenised_content_lemma	content_size_lemma	doc_vector	embeddings
0	0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	3244	[-1.09343290e-01 -7.68003017e-02 -1.80567741e-...	[-1.16498247e-01 -4.47961055e-02 2.62079686e-..
1	1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	2853	[-9.83889028e-02 -7.53047839e-02 -1.65519699e-...	[-1.76346358e-02 -6.64909780e-02 8.61539785e-..
2	2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	1732	[-4.89264429e-02 -1.39095232e-01 -1.84526831e-...	[-1.70915481e-02 -4.40688655e-02 4.41733487e-..
3	3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	2948	[-9.89964455e-02 -6.88564703e-02 -1.66541591e-...	[-6.59025311e-02 -2.54796501e-02 4.83894758e-..
4	4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	1482	[-0.10334199 -0.06256191 -0.15584423 -0.213279...	[-6.81307241e-02 -2.44800914e-02 -2.52653230e-..

In [8]:
```
# Assuming 'df' is your DataFrame
emb_df = final_df[['num', 'name', 'content_chunks', 'embeddings']].copy()
```

In [9]:
```
# Assuming 'df' is your DataFrame
emb_df1 = final_df[['num', 'name', 'content_chunks', 'embeddings']].copy()
```

In [10]:
```
emb_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 4 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   num             30000 non-null  int64 
 1   name            30000 non-null  object
 2   content_chunks  30000 non-null  object
 3   embeddings      30000 non-null  object
dtypes: int64(1), object(3)
memory usage: 937.6+ KB
```

In [11]:
```
emb_df.head()
```

Out[11]:

	num	name	content_chunks	embeddings
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	['riley ipreviously oni the client list derek ...	[-1.16498247e-01 -4.47961055e-02 2.62079686e-...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	['support us and become vip member to remove a...	[-1.76346358e-02 -6.64909780e-02 8.61539785e-...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	['script info title english us original script...	[-1.70915481e-02 -4.40688655e-02 4.41733487e-...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	['wind whooshing eerie synth music bats chirpi...	[-6.59025311e-02 -2.54796501e-02 4.83894758e-...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	['watch any video online with opensubtitles fr...	[-6.81307241e-02 -2.44800914e-02 -2.52653230e-...

In [12]:
```
import sqlite3

# Specify the path to your desired location (e.g., /content/drive/MyDrive)
db_path = 'chromadb.db'

# Connect to the SQLite database (creates the file if it doesn't exist)
conn = sqlite3.connect(db_path)
conn.close()  # Close the connection

print(f"SQLite database '{db_path}' created successfully.")
```

```
SQLite database 'chromadb.db' created successfully.
```

In [13]:
```
import sqlite3

try:
    # Specify the path to your desired location (e.g., /content/drive/MyDrive)
    db_path = r'R:\Data Science\ZDS\Intership\Innomatics\Assignment8_Search Engine Project\chromadb.db'

    # Connect to the SQLite database (creates the file if it doesn't exist)
    conn = sqlite3.connect(db_path)
    conn.close()  # Close the connection

    print(f"SQLite database '{db_path}' created successfully.")
except Exception as e:
    print("An error occurred:", e)
```

```
SQLite database 'R:\Data Science\ZDS\Intership\Innomatics\Assignment8_Search Engine Project\chromadb.db' created successfully.
```

In [14]:
```
import chromadb
chroma_client = chromadb.Client()
```

In [15]:
```
collection = chroma_client.create_collection(name="my_collection")
```

In [16]:
```
emb_df.head()
```

Out[16]:

	num	name	content_chunks	embeddings
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	['riley ipreviously oni the client list derek ...	[-1.16498247e-01 -4.47961055e-02 2.62079686e-...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	['support us and become vip member to remove a...	[-1.76346358e-02 -6.64909780e-02 8.61539785e-...
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	['script info title english us original script...	[-1.70915481e-02 -4.40688655e-02 4.41733487e-...
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	['wind whooshing eerie synth music bats chirpi...	[-6.59025311e-02 -2.54796501e-02 4.83894758e-...
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	['watch any video online with opensubtitles fr...	[-6.81307241e-02 -2.44800914e-02 -2.52653230e-...

Create a file in Cromadb

In [18]:

```
import sqlite3
import pandas as pd
import json

# SQLite database path
db_path = 'chromadb1136.db'

# Connect to the SQLite database
conn = sqlite3.connect(db_path)

# Convert lists to strings
emb_df['content_chunks'] = emb_df['content_chunks'].apply(json.dumps)
emb_df['embeddings'] = emb_df['embeddings'].apply(json.dumps)

# Save DataFrame to SQLite
emb_df.to_sql("dd", conn, if_exists="replace", index=False)

print(f"DataFrame from {emb_df} is successfully saved to ChromaDB using SQLite: {db_path}")
```

```
DataFrame from            num                                                 name  \
0      9487821  the.client.list.s02.e06.unanswered.prayers.(20...
1      9183678               the.devil.is.a.woman.(1935).eng.1cd
2      9258855  kingdom.s04.e18.rivercrossing.battle.(2022).en...
3      9421238  sorority.babes.in.the.slimeball.bowlorama.2.(2...
4      9518525  ada.twist.scientist.s03.e03.dadbotinto.the.bra...
...        ...                                                 ...
29995  9473109  king.of.the.road.s03.e03.chained.up.and.trippe...
29996  9487995                             badoet.(2015).eng.1cd
29997  9400464            velma.s01.e03.velma.kai.(2023).eng.1cd
29998  9489287      numberblocks.s03.e26.thirteen.(2019).eng.1cd
29999  9305337  def.comedy.jam.s02.e09.episode.2.9.(1992).eng.1cd

                                       content_chunks  \
0      "\"['riley ipreviously oni the client list der...
1      "\"['support us and become vip member to remov...
2      "\"['script info title english us original scr...
3      "\"['wind whooshing eerie synth music bats chi...
4      "\"['watch any video online with opensubtitles...
...                                                  ...
29995  "\"['please do not attempt to perform any of t...
29996  "\"['crickets chirring music box music apiopen...
29997  "\"['i previously oni velma the last place my ...
29998  "\"['one two three four five numberblocks six ...
29999  "\"['kid capri its the russell simmons def com...

                                          embeddings
0      "[-1.16498247e-01 -4.47961055e-02  2.62079686e...
1      "[-1.76346358e-02 -6.64909780e-02  8.61539785e...
2      "[-1.70915481e-02 -4.40688655e-02  4.41733487e...
3      "[-6.59025311e-02 -2.54796501e-02  4.83894758e...
4      "[-6.81307241e-02 -2.44800914e-02 -2.52653230e...
...                                                 ...
29995  "[ 7.56419310e-03 -5.54080568e-02  2.06312984e...
29996  "[ 1.17038190e-02 -2.66452506e-02  2.01791693e...
29997  "[-8.32793862e-02 -6.34732619e-02  5.84286172e...
29998  "[-8.33337158e-02 -1.16495080e-01 -2.15825532e...
29999  "[-4.05492522e-02 -7.91372731e-02  1.41583364e...

[30000 rows x 4 columns] is successfully saved to ChromaDB using SQLite: chromadb1136.db
```

In [ ]:

In [ ]:

Experiment on 5k data set

In [34]:
```
new_df.head()
```

Out[34]:

	num	name	cleaned_file_content	tokenised_content_lemma	con
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	riley ipreviously oni the client list derek yo...	['riley', 'ipreviously', 'oni', 'client', 'lis...	
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	support us and become vip member to remove all...	['support', 'u', 'become', 'vip', 'member', 'r...	
2	9258855	kingdom.s04.e18.rivercrossing.battle.(2022).en...	script info title english us original script p...	['script', 'info', 'title', 'english', 'u', 'o...	
3	9421238	sorority.babes.in.the.slimeball.bowlorama.2.(2...	wind whooshing eerie synth music bats chirping...	['wind', 'whooshing', 'eerie', 'synth', 'music...	
4	9518525	ada.twist.scientist.s03.e03.dadbotinto.the.bra...	watch any video online with opensubtitles free...	['watch', 'video', 'online', 'opensubtitles', ...	

In [35]:
```
final_df = new_df.copy()
```

In [52]:
```
df1 = new_df.copy()
```

Create a file in Cromadb

In [38]:
```
import sqlite3

# Specify the path to your desired location (e.g., /content/drive/MyDrive)
db_path = 'chromadb.db'

# Connect to the SQLite database (creates the file if it doesn't exist)
conn = sqlite3.connect(db_path)
conn.close()  # Close the connection

print(f"SQLite database '{db_path}' created successfully.")
```

```
SQLite database 'chromadb.db' created successfully.
```

In [39]:
```
import sqlite3

try:
    # Specify the path to your desired location (e.g., /content/drive/MyDrive)
    db_path = r'R:\Data Science\ZDS\Intership\Innomatics\Assignment8_Search Engine Pro

    # Connect to the SQLite database (creates the file if it doesn't exist)
    conn = sqlite3.connect(db_path)
    conn.close()  # Close the connection

    print(f"SQLite database '{db_path}' created successfully.")
except Exception as e:
    print("An error occurred:", e)
```

```
SQLite database 'R:\Data Science\ZDS\Intership\Innomatics\Assignment8_Search Engine
Project\chromadb.db' created successfully.
```

In [40]:
```
import chromadb
chroma_client = chromadb.Client()
```

In [41]:
```
collection = chroma_client.create_collection(name="my_collection")
```

In [53]:
```
df1 = df1.rename(columns={'sentance_content_bert': 'embeddings'})
```

In [69]:
```
emb_df1 = df1.copy()
```

In [54]:
```
# Assuming 'df' is your DataFrame
emb_df1 = df1[['num', 'name', 'content_chunks', 'embeddings']].copy()
```

In [56]:
```
emb_df1.head(2)
```

Out[56]:

	num	name	content_chunks	embeddings
0	9487821	the.client.list.s02.e06.unanswered.prayers.(20...	[riley ipreviously oni the client list derek y...	[-0.022161696, -0.111041605, 0.02133926, 0.050...
1	9183678	the.devil.is.a.woman.(1935).eng.1cd	[support us and become vip member to remove al...	[-0.0005046579, -0.060602482, -0.03618558, -0....

In [ ]:
```

```

In [ ]:
```
collection = chroma_client.create_collection(name="my_collection")
```

In [ ]:
```
df1 = df1.rename(columns={'sentance_content_bert': 'embeddings'})
```

In [66]:
```
import sqlite3
import pandas as pd
import json

# SQLite database path
db_path = 'chromadb1133.db'

# Connect to the SQLite database
conn = sqlite3.connect(db_path)

# Convert lists to strings
emb_df1['content_chunks'] = emb_df1['content_chunks'].apply(json.dumps)
emb_df1['embeddings'] = emb_df1['embeddings'].apply(lambda x: json.dumps(x.tolist()))

# Save DataFrame to SQLite
emb_df1.to_sql("dd", conn, if_exists="replace", index=False)

print(f"DataFrame from {emb_df1} is successfully saved to ChromaDB using SQLite: {db_p
```

```
DataFrame from          num                                                name  \
0     9487821  the.client.list.s02.e06.unanswered.prayers.(20...
1     9183678               the.devil.is.a.woman.(1935).eng.1cd
2     9258855  kingdom.s04.e18.rivercrossing.battle.(2022).en...
3     9421238  sorority.babes.in.the.slimeball.bowlorama.2.(2...
4     9518525  ada.twist.scientist.s03.e03.dadbotinto.the.bra...
...       ...                                                ...
4996  9237926         blue.gender.s01.e16.a.sign.(2000).eng.1cd
4997  9375637    the.equalizer.s01.e10.reckoning.(2021).eng.1cd
4998  9226928  rap.sht.s01.e08.something.for.the.road.(2022)....
4999  9316079  dead.to.me.s03.e08.well.find.a.way.(2022).eng.1cd
5000  9305324  def.comedy.jam.s01.e04.episode.1.4.(1992).eng.1cd

                                          content_chunks  \
0     "\"\\\"[\\\\\\\"riley ipreviously oni the clie...
1     "\"\\\"[\\\\\\\"support us and become vip memb...
2     "\"\\\"[\\\\\\\"script info title english us o...
3     "\"\\\"[\\\\\\\"wind whooshing eerie synth mus...
4     "\"\\\"[\\\\\\\"watch any video online with op...
...                                                 ...
4996  "\"\\\"[\\\\\\\"watch any video online with op...
4997  "\"\\\"[\\\\\\\"iim the one you call when you ...
4998  "\"\\\"[\\\\\\\"rap music playing wondering ho...
4999  "\"\\\"[\\\\\\\"wwwopensubtitlescom hey mama h...
5000  "\"\\\"[\\\\\\\"theme music playing kid capri ...

                                             embeddings
0     "[-0.02216169610619545, -0.1110416054725647, 0...
1     "[-0.0005046579171903431, -0.06060248240828514...
2     "[-0.03646128252148628, -0.030952945351600647,...
3     "[0.02794903703033924, -0.07734321057796478, 0...
4     "[0.0011420863447710872, -0.11879019439220428,...
...                                                 ...
4996  "[0.007031263317912817, -0.10100991278886795, ...
4997  "[0.006331723649054766, -0.08844021707773209, ...
4998  "[0.010032282210886478, -0.14277908205986023, ...
4999  "[-0.04041879624128342, -0.07212220132350922, ...
5000  "[0.014225239865481853, -0.08701223880052567, ...

[5001 rows x 4 columns] is successfully saved to ChromaDB using SQLite: chromadb113
3.db
```

In [ ]:

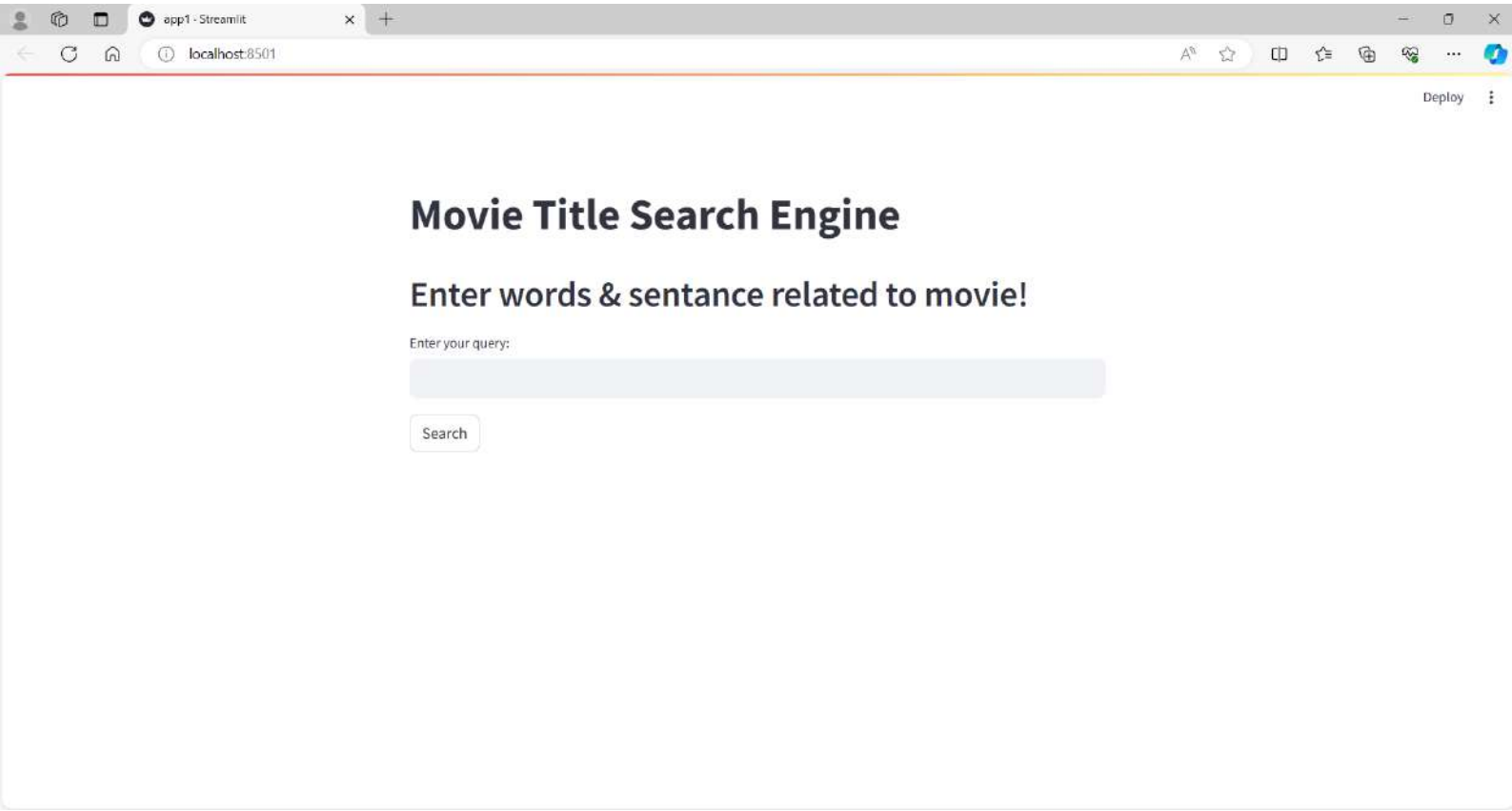
In [68]:
```
emb_df1.content_chunks[66]
```

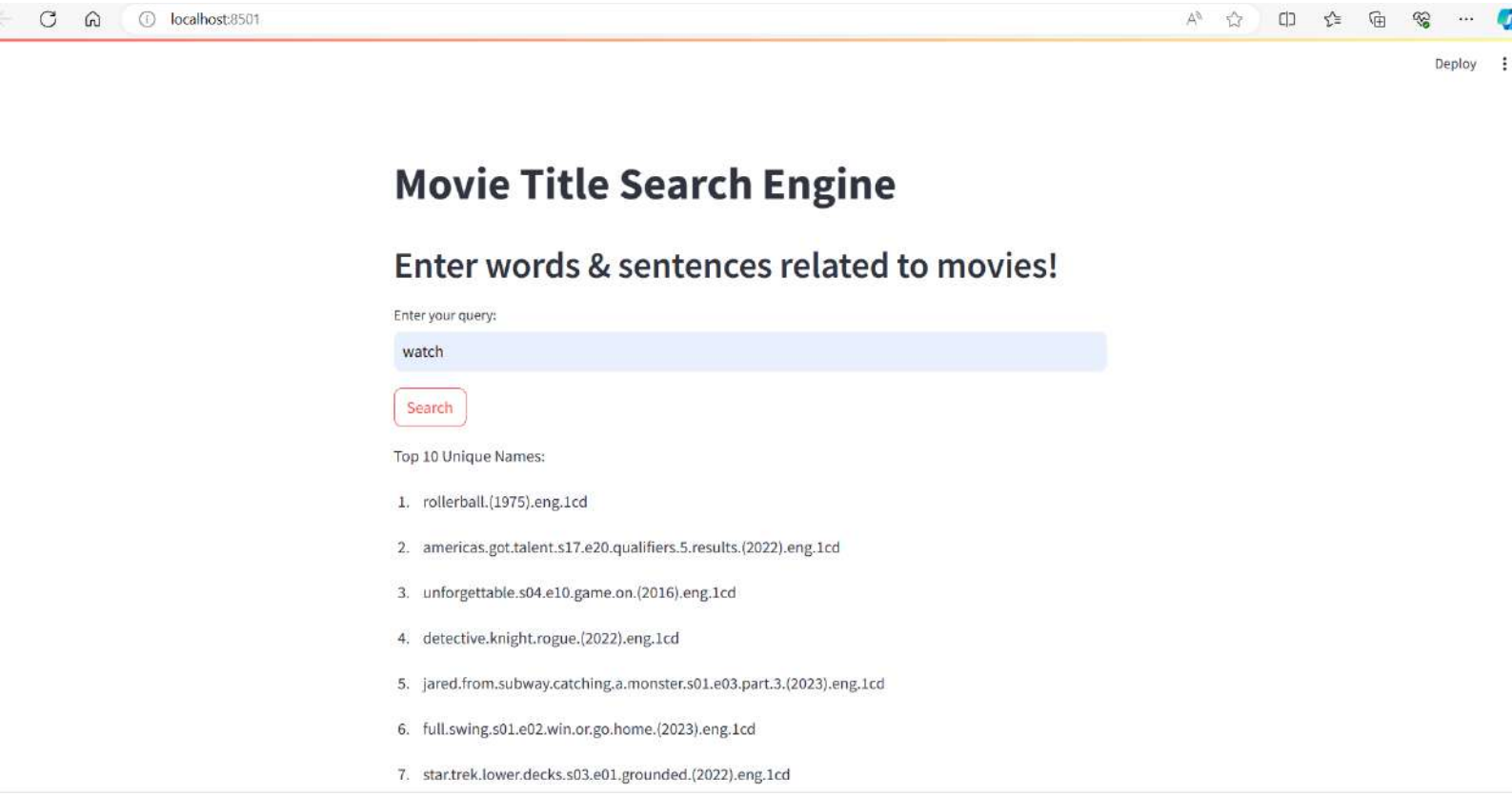
```
with washington but it does not look promising hmm whats the big whoop we can keep
the talismans here in one of uncles magic boxes the risk is too great the talisman
s will be safest from dark forces on sacred ground we must take them to the benshu
i temple but but isnt that like near china jade the talismans are unpredictable da
ngerous you should not be playing with \\\\\\\\\\\\\\", \\\\\\\\\\\\\\" that like
near china jade the talismans are unpredictable dangerous you should not be playin
g with them in the first place blows raspberry snip testing im in im gonna miss yo
u guys how about one last hurrah for old times sake hmm hydraulic humming uhoh huh
aah ooh groaning man section s a long way from washington captain black black i pr
omise your trip will be worthwhile sir gasps black follow me once you see just how
powerful these talismans are im sure youll agree its imperative we keep them
\\\\\\\\\\\\\\", \\\\\\\\\\\\\\" once you see just how powerful these talismans ar
e im sure youll agree its imperative we keep them here under lock and laser whew t
hose guys must be blind observe the power of invisibility astonishing isnt it i co
uld be over here or perhaps im over here hmm must not be doing it right never mind
lets try levitation grunting uhh i dont understand its running a section can be ve
ry stressful captain black im authorizing weeks of r and r effective immediately i
suggest you get as far away from here \\\\\\\\\\\\\\", \\\\\\\\\\\\\\"black im aut
horizing weeks of r and r effective immediately i suggest you get as far away from
here as possible im very sorry about this you should be sorry we are not used to t
```

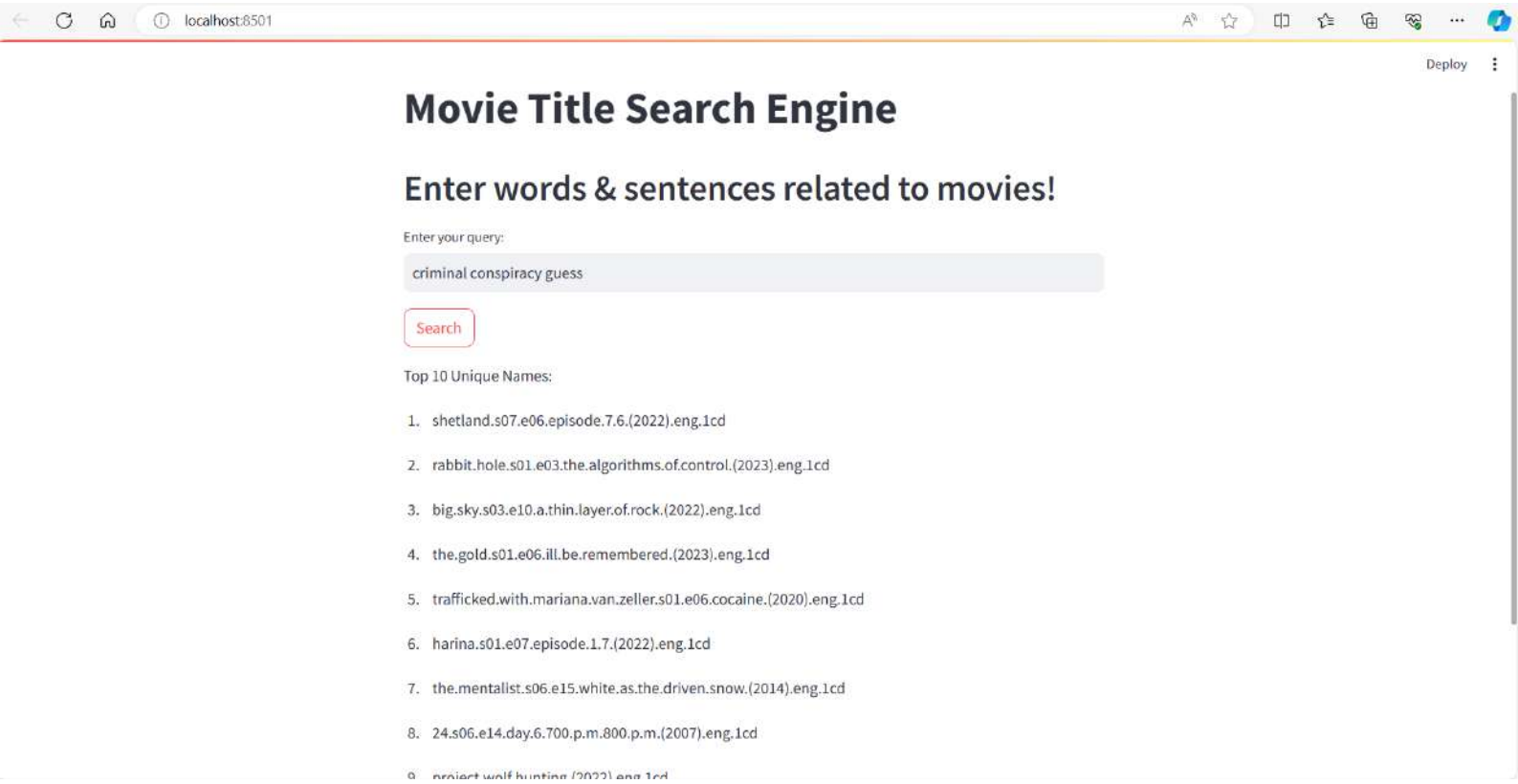
In [63]:
```
import sqlite3

# Connect to the SQLite database
conn = sqlite3.connect('chromadb1133.db')
cursor = conn.cursor()

# Fetch the table schema
cursor.execute("PRAGMA table_info(dd)")
columns = cursor.fetchall()

# Print the table schema
for column in columns:
    print(column)

# Close the cursor and connection
cursor.close()
conn.close()
```

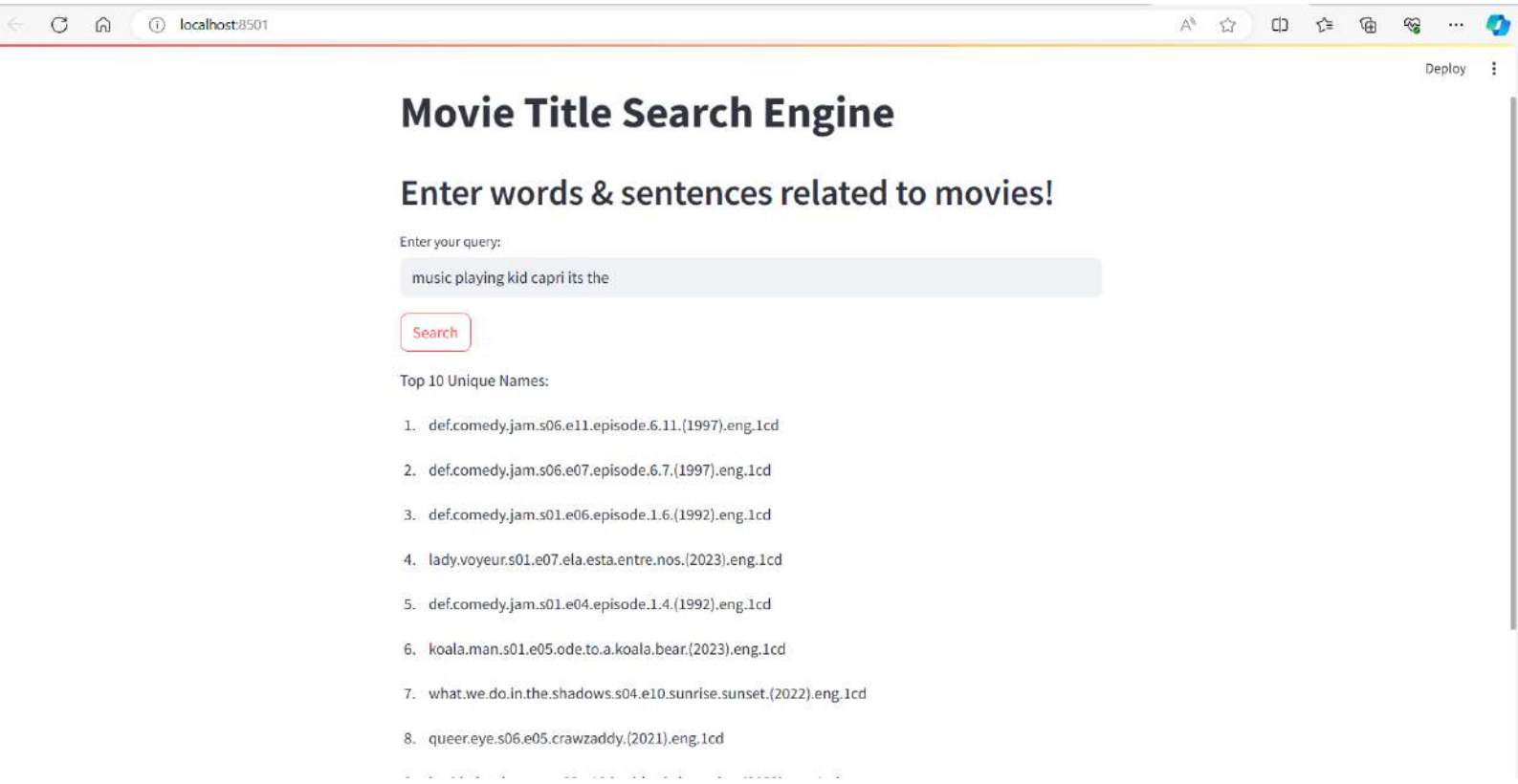
```
(0, 'num', 'INTEGER', 0, None, 0)
(1, 'name', 'TEXT', 0, None, 0)
(2, 'content_chunks', 'TEXT', 0, None, 0)
(3, 'embeddings', 'TEXT', 0, None, 0)
```

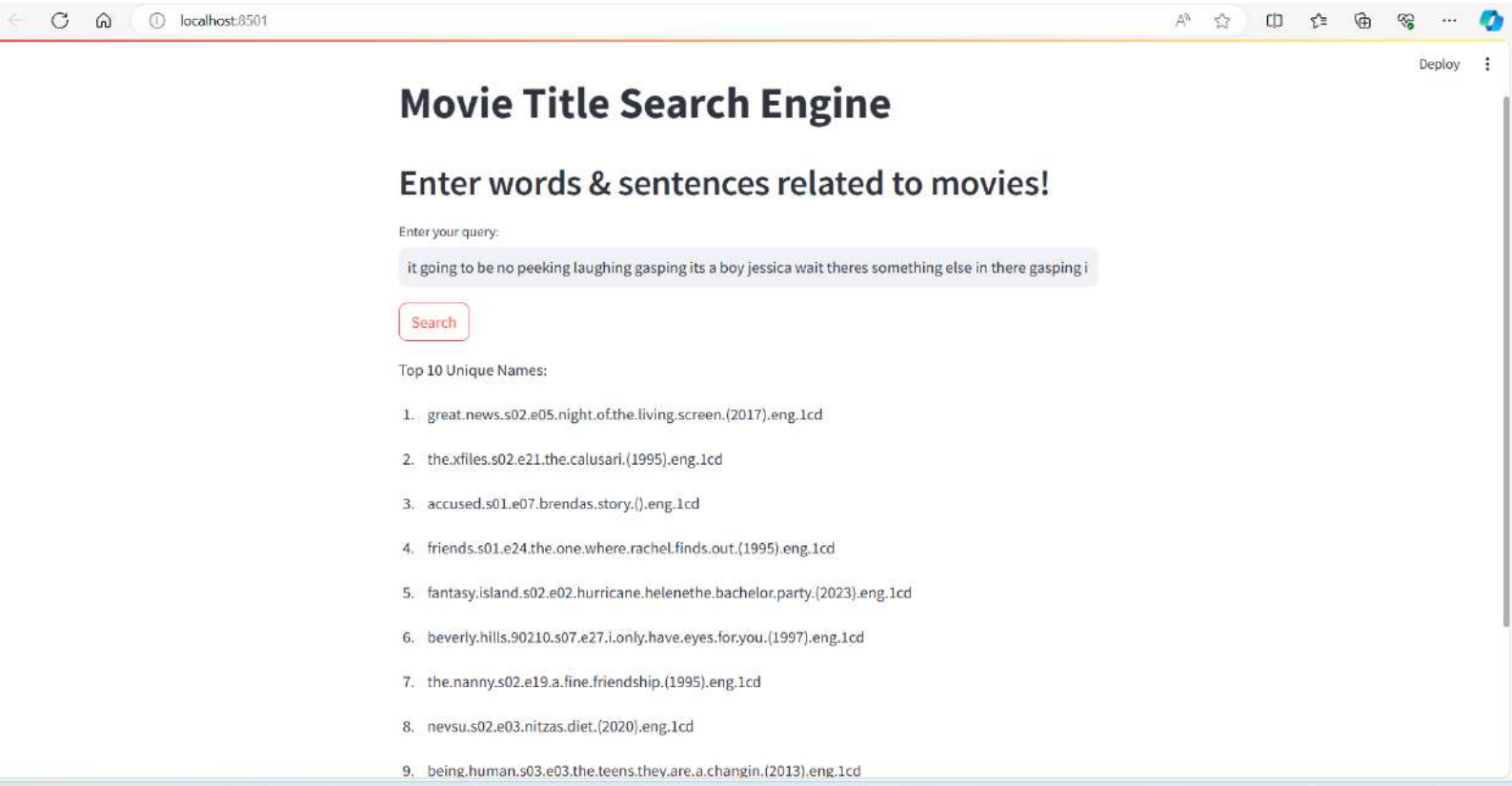
In [ ]:

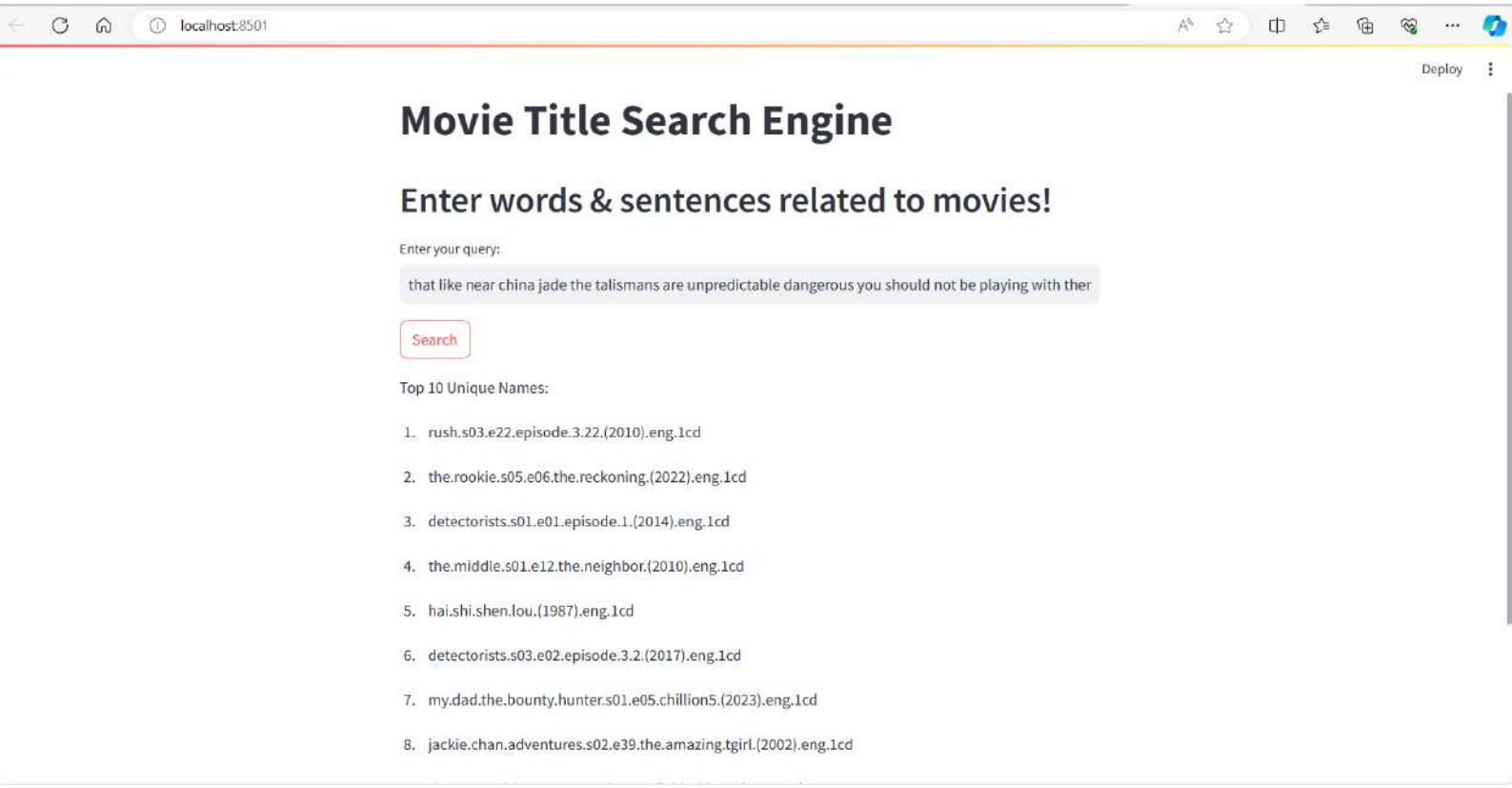

app1 - Streamlit
localhost:8501
Deploy
Movie Title Search Engine
Enter words & sentance related to movie!
Enter your query:
Search


localhost:8501
Deploy
Movie Title Search Engine
Enter words & sentences related to movies!
Enter your query:
watch
Search
Top 10 Unique Names:
1. rollerball.(1975).eng.1cd
2. americas.got.talent.s17.e20.qualifiers.5.results.(2022).eng.1cd
3. unforgettable.s04.e10.game.on.(2016).eng.1cd
4. detective.knight.rogue.(2022).eng.1cd
5. jared.from.subway.catching.a.monster.s01.e03.part.3.(2023).eng.1cd
6. full.swing.s01.e02.win.or.go.home.(2023).eng.1cd
7. star.trek.lower.decks.s03.e01.grounded.(2022).eng.1cd

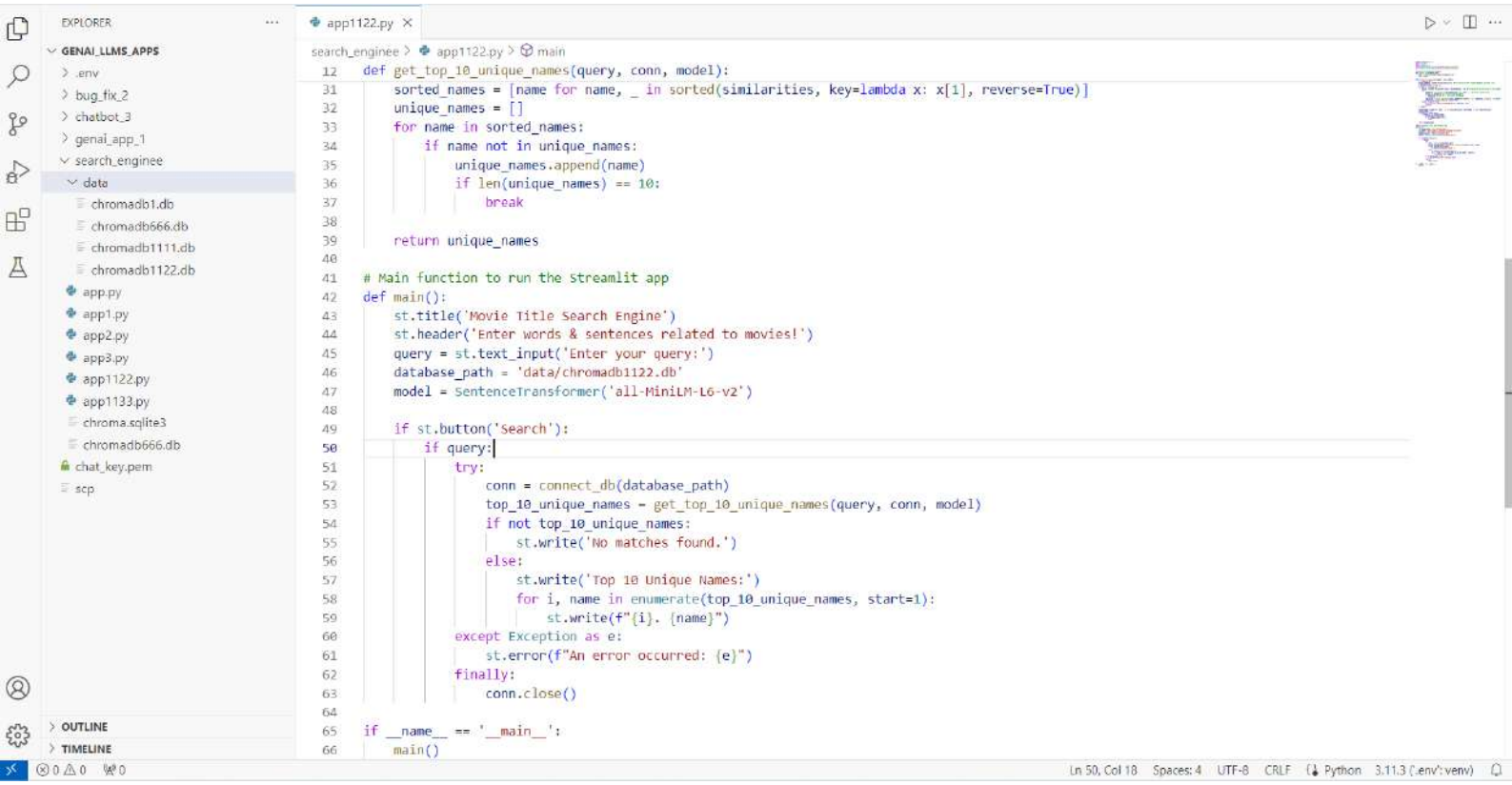

localhost:8501
Deploy
Movie Title Search Engine
Enter words & sentences related to movies!
Enter your query:
criminal conspiracy guess
Search
Top 10 Unique Names:
1. shetland.s07.e06.episode.7.6.(2022).eng.1cd
2. rabbit.hole.s01.e03.the.algorithms.of.control.(2023).eng.1cd
3. big.sky.s03.e10.a.thin.layer.of.rock.(2022).eng.1cd
4. the.gold.s01.e06.ill.be.remembered.(2023).eng.1cd
5. trafficked.with.mariana.van.zeller.s01.e06.cocaine.(2020).eng.1cd
6. harina.s01.e07.episode.1.7.(2022).eng.1cd
7. the.mentalist.s06.e15.white.as.the.driven.snow.(2014).eng.1cd
8. 24.s06.e14.day.6.700.p.m.800.p.m.(2007).eng.1cd


localhost:8501
Deploy
Movie Title Search Engine
Enter words & sentences related to movies!
Enter your query:
music playing kid capri its the
Search
Top 10 Unique Names:
1. def.comedy.jam.s06.e11.episode.6.11.(1997).eng.1cd
2. def.comedy.jam.s06.e07.episode.6.7.(1997).eng.1cd
3. def.comedy.jam.s01.e06.episode.1.6.(1992).eng.1cd
4. lady.voyeur.s01.e07.ela.esta.entre.nos.(2023).eng.1cd
5. def.comedy.jam.s01.e04.episode.1.4.(1992).eng.1cd
6. koala.man.s01.e05.ode.to.a.koala.bear.(2023).eng.1cd
7. what.we.do.in.the.shadows.s04.e10.sunrise.sunset.(2022).eng.1cd
8. queer.eye.s06.e05.crawzaddy.(2021).eng.1cd


localhost:8501
Deploy
Movie Title Search Engine
Enter words & sentences related to movies!
Enter your query:
it going to be no peeking laughing gasping its a boy jessica wait theres something else in there gasping i
Search
Top 10 Unique Names:
1. great.news.s02.e05.night.of.the.living.screen.(2017).eng.1cd
2. the.xfiles.s02.e21.the.calusari.(1995).eng.1cd
3. accused.s01.e07.brendas.story.().eng.1cd
4. friends.s01.e24.the.one.where.rachel.finds.out.(1995).eng.1cd
5. fantasy.island.s02.e02.hurricane.helenethe.bachelor.party.(2023).eng.1cd
6. beverly.hills.90210.s07.e27.i.only.have.eyes.for.you.(1997).eng.1cd
7. the.nanny.s02.e19.a.fine.friendship.(1995).eng.1cd
8. nevsu.s02.e03.nitzas.diet.(2020).eng.1cd
9. being.human.s03.e03.the.teens.they.are.a.changin.(2013).eng.1cd


localhost:8501
Deploy
Movie Title Search Engine
Enter words & sentences related to movies!
Enter your query:
that like near china jade the talismans are unpredictable dangerous you should not be playing with ther
Search
Top 10 Unique Names:
1. rush.s03.e22.episode.3.22.(2010).eng.1cd
2. the.rookie.s05.e06.the.reckoning.(2022).eng.1cd
3. detectorists.s01.e01.episode.1.(2014).eng.1cd
4. the.middle.s01.e12.the.neighbor.(2010).eng.1cd
5. hai.shi.shen.lou.(1987).eng.1cd
6. detectorists.s03.e02.episode.3.2.(2017).eng.1cd
7. my.dad.the.bounty.hunter.s01.e05.chillion5.(2023).eng.1cd
8. jackie.chan.adventures.s02.e39.the.amazing.tgirl.(2002).eng.1cd


EXPLORER
GENAI_LLMS_APPS
.env
bug_fix_2
chatbot_3
genai_app_1
search_enginee
data
chromadb1.db
chromadb666.db
chromadb1111.db
chromadb1122.db
app.py
app1.py
app2.py
app3.py
app1122.py
app1133.py
chroma.sqlite3
chromadb666.db
chat_key.pem
scp
OUTLINE
TIMELINE
app1122.py
search_enginee > app1122.py > main
import streamlit as st
import sqlite3
import numpy as np
from sklearn.metrics.pairwise import cosine_similarity
from sentence_transformers import SentenceTransformer
# Connect to the SQLite database
def connect_db(database_path):
conn = sqlite3.connect("data/chromadb1122.db")
return conn
def get_top_10_unique_names(query, conn, model):
c = conn.cursor()
query_embedding = model.encode([query])[0] # Use [0] to get the single embedding from the list
similarities = []
c.execute("SELECT * FROM dd")
for row in c.fetchall():
dd_num, dd_name, dd_content_chunks, dd_embeddings = row # Assuming the fourth column is not needed
try:
embeddings = np.fromstring(dd_embeddings[1:-1], sep=', ') # Parse as numpy array
# Check if embeddings is a list of valid numbers
if embeddings.size == 0: # Skip empty embeddings
continue
similarity = cosine_similarity(query_embedding.reshape(1, -1), embeddings.reshape(1, -1))[0][0]
similarities.append((dd_name, similarity))
except Exception as e:
print(f"Error processing embeddings for {dd_name}: {e}")
continue
c.close()
sorted_names = [name for name, _ in sorted(similarities, key=lambda x: x[1], reverse=True)]
unique_names = []
for name in sorted_names:
if name not in unique_names:
unique_names.append(name)
if len(unique_names) == 10:
break


EXPLORER
app1122.py
GENAI_LLMS_APPS
.env
bug_fix_2
chatbot_3
genai_app_1
search_enginee
data
chromadb1.db
chromadb666.db
chromadb1111.db
chromadb1122.db
app.py
app1.py
app2.py
app3.py
app1122.py
app1133.py
chroma.sqlite3
chromadb666.db
chat_key.pem
scp
OUTLINE
TIMELINE
search_enginee > app1122.py > main
def get_top_10_unique_names(query, conn, model):
sorted_names = [name for name, _ in sorted(similarities, key=lambda x: x[1], reverse=True)]
unique_names = []
for name in sorted_names:
if name not in unique_names:
unique_names.append(name)
if len(unique_names) == 10:
break
return unique_names
# Main function to run the Streamlit app
def main():
st.title('Movie Title Search Engine')
st.header('Enter words & sentences related to movies!')
query = st.text_input('Enter your query:')
database_path = 'data/chromadb1122.db'
model = SentenceTransformer('all-MiniLM-L6-v2')
if st.button('Search'):
if query:
try:
conn = connect_db(database_path)
top_10_unique_names = get_top_10_unique_names(query, conn, model)
if not top_10_unique_names:
st.write('No matches found.')
else:
st.write('Top 10 Unique Names:')
for i, name in enumerate(top_10_unique_names, start=1):
st.write(f"{i}. {name}")
except Exception as e:
st.error(f"An error occurred: {e}")
finally:
conn.close()
if __name__ == '__main__':
main()
Ln 50, Col 18
Spaces: 4
UTF-8
CRLF
Python
3.11.3 ('.env': venv)