SCTR's Pune Institute of Computer Technology
Dhankawadi, Pune

AN INTERNSHIP REPORT ON

Content Based Music Recommendation System

SUBMITTED BY
Name: Ashutosh Thite
Class: TE - 4
Roll no: 31462

Under the guidance of
Dr. A. R. Deshpande

DEPARTMENT OF COMPUTER ENGINEERING
ACADEMIC YEAR 2021-22

Acknowledgement

It gives me great pleasure in presenting the internship report on "Content Based Music Recommendation System".

First of all I would like to thank our guide and mentor Dr. A. R. Deshpande ma'am for giving me this opportunity to do this internship. I would like to thank her for giving me all the help and guidance needed. I am really grateful for her kind support and valuable suggestions that proved to be beneficial in the overall completion of this internship.

I am thankful to our Head of Computer Engineering Department, Dr. G.V.Kale, for her indispensable support and suggestions throughout the internship work.

I would also genuinely like to express my gratitude to the Department Internship Coordinator, Prof.P.P.Joshi, for her constant guidance and support and for the timely resolution of the doubts related to the internship process.

Finally, I would again like to thank my mentor, Dr. A. R. Deshpande for her constant help and support during the overall internship process.

Contents

1 Title

Content Based Music Recommendation System

2 Introduction

This inhouse internship at Big Data and Text Analytics Group, PICT started on 12-02-2022 and will end on 25-04-2022. This was a research-based internship, where the problem statement was to develop a content based music recommendation system.

Name of inhouse internship :**Big Data and Text Analytics Group, PICT**
Address : Pune Institute of Computer Technology,Survey No. 27, Near Trimurti Chowk Dhankwadi, Pune – 411043
Institutional Mentor :- **Dr. Arati Deshpande, Asst. Prof. Dept. of Computer Engg.**
Guide name :- **Dr. Arati Deshpande, Asst. Prof. Dept. of Computer Engg.**

Start Date : **12-02-2022**
End Date : **25-04-2022**
Paid/ Unpaid : **Unpaid**
Mode of Internship : **Hybrid**

The whole internship progress was subdivided into smaller tasks in order to complete the internship on time.

3 Problem Statement

Design and Develop a recommendation system with collaborative/content/hybrid algorithm with context. Test on the accuracy and build a website/mobile application for the same.

4 Objectives and Scope

Objectives

- To develop a recommendation system to generate relevant recommendations from given dataset.
- To use various tools and techniques in data science e.g. scikit learn, pandas, nlpd.
- To use given dataset and use data science tools and techniques to generate relevant music recommendations.

Scope

- In this era of information overload, having the correct and relevant piece of information is necessary.
- Same is the case for music. Finding right music to play after current song, using information about current song's artist and other data is useful.
- This Content- Based Music Recommendation system will help user to genrate relevant music recommendations based on Artist Name, Genre of the current Song and Playlist in which the song is added.

5 Methodological Details

- **Content Based Recommendations**
  Content-based methods gives recommendations based on the similarity of two song contents or attributes while collaborative methods make a prediction on posible preferences using a matrix with ratings on different songs.

- **Stemming and Lemmatization**
  Stemming and lemmatization are methods used by search engines to analyze the meaning behind a word. Stemming uses the stem of the word, while lemmatization uses the context in which the word is being used.

- **Vectorisation**
  Vectorisation is a classic approach of converting input data from its raw format (i.e. text) into vectors of real numbers which is the format that ML models support.

  In Machine Learning, vectorization is a step in feature extraction. The idea is to get some distinct features out of the text for the model to train on, by converting text to numerical vectors.

- **Feature extraction using CountVectorizer**
  the CountVectorizer is used to vectorize and check for similarity in the input song title and the artist, genres and playlist name. We can specify the number of songs we want to be recommended/ displayed based on this similarity.

- **Cosine Similarity**
  Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis. Cosine similarity is checked and the songs matching the most are showed in decreasing order of their cosine similarity.

6 Modern engineering tools used

- **Jupyter Notebook**

  Jupyter Notebook is a web-based interactive computational environment for creating notebook documents. A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media.

- **Matplotlib**

  Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

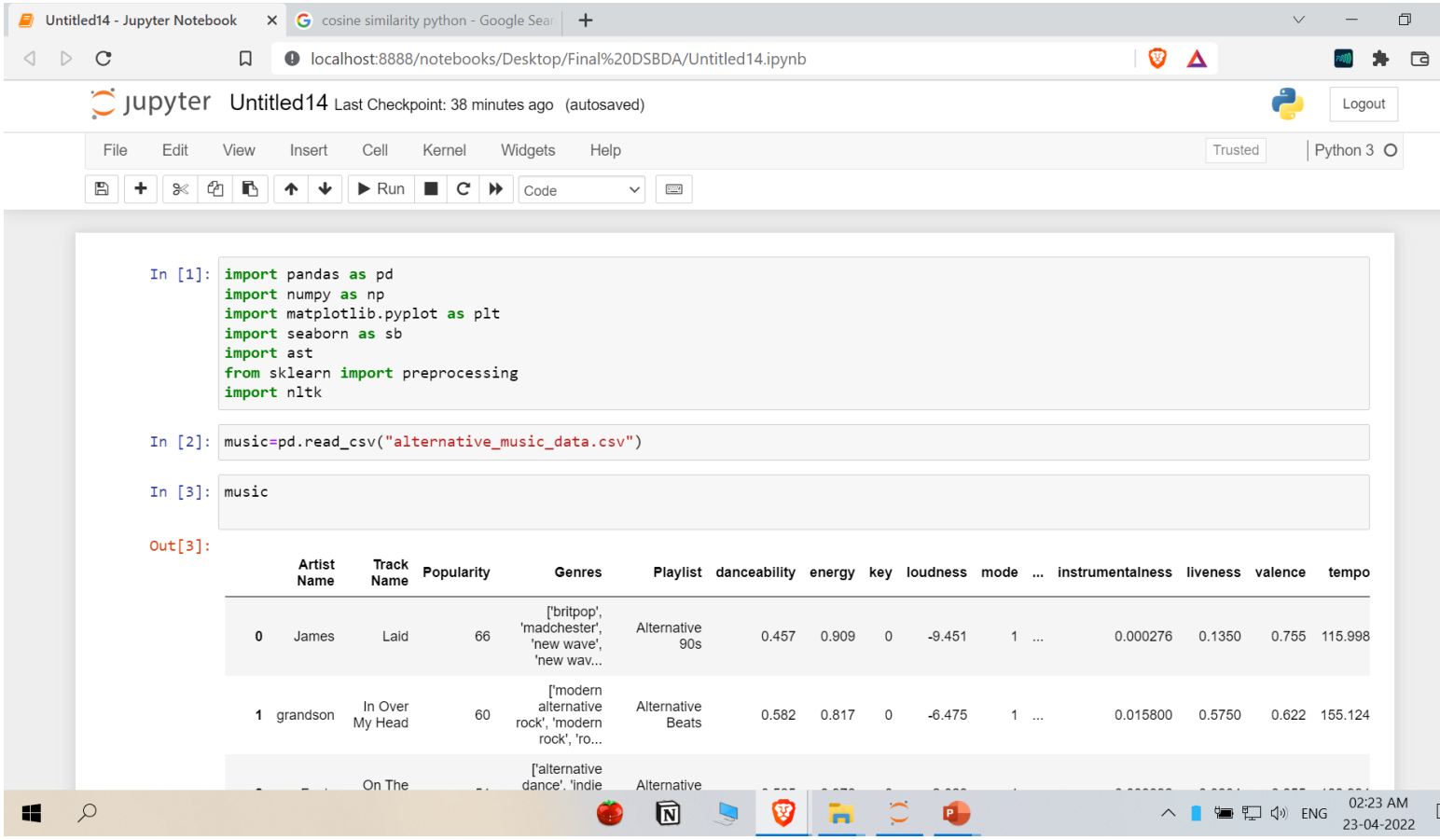
- **Scikit-learn**

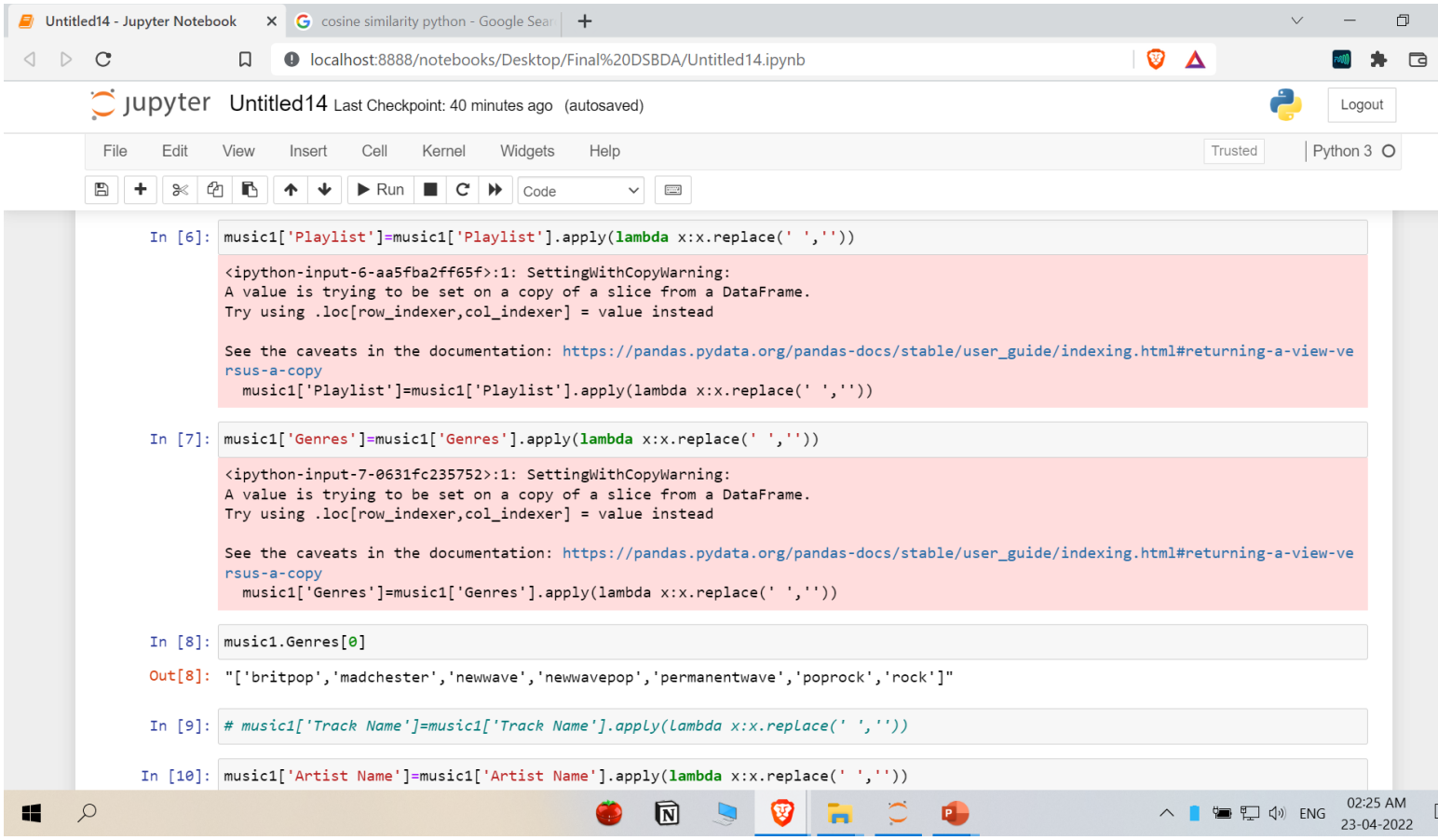
  Scikit-learn is a machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

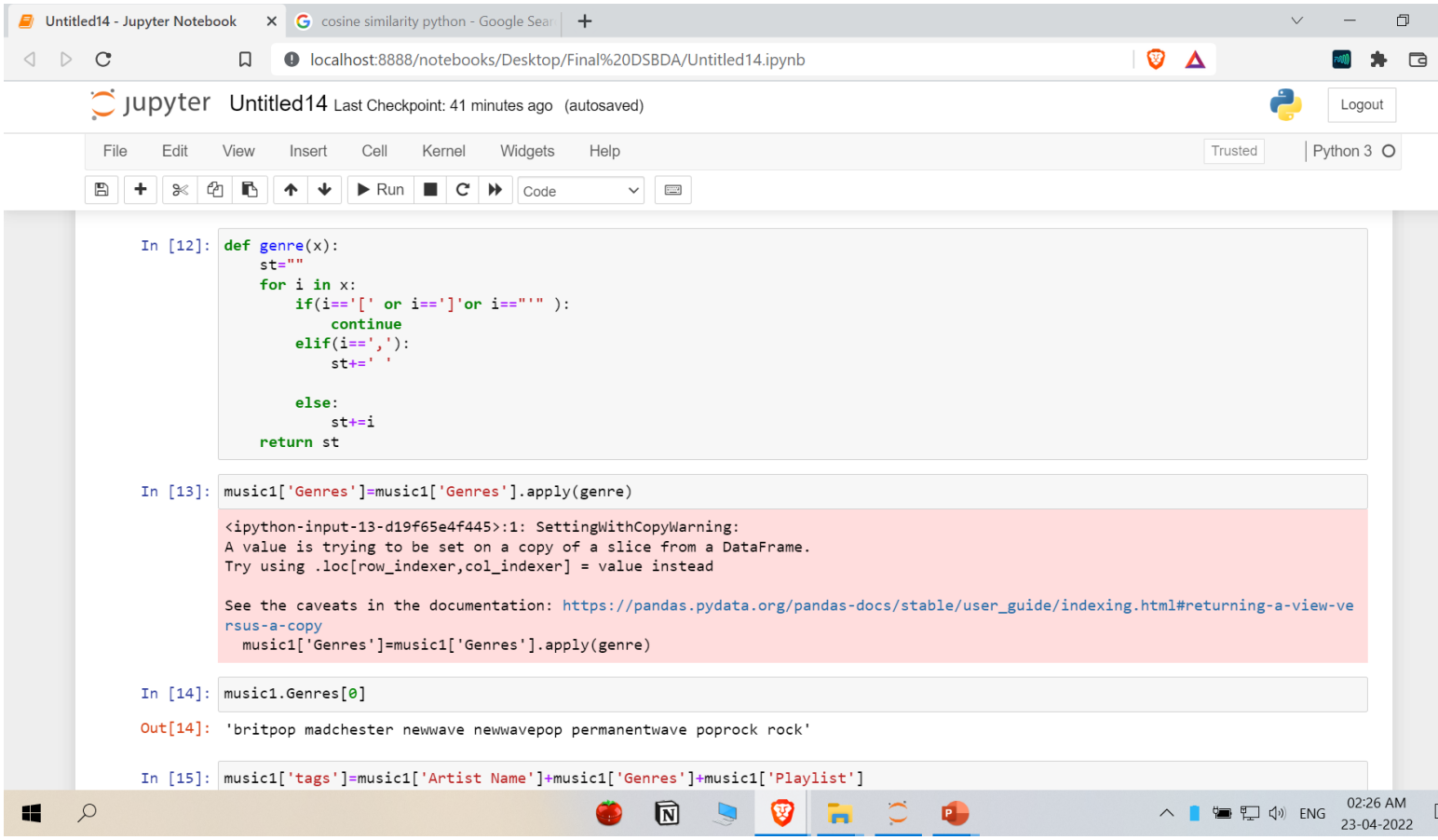
- **NLTK - Natural Language Toolkit**

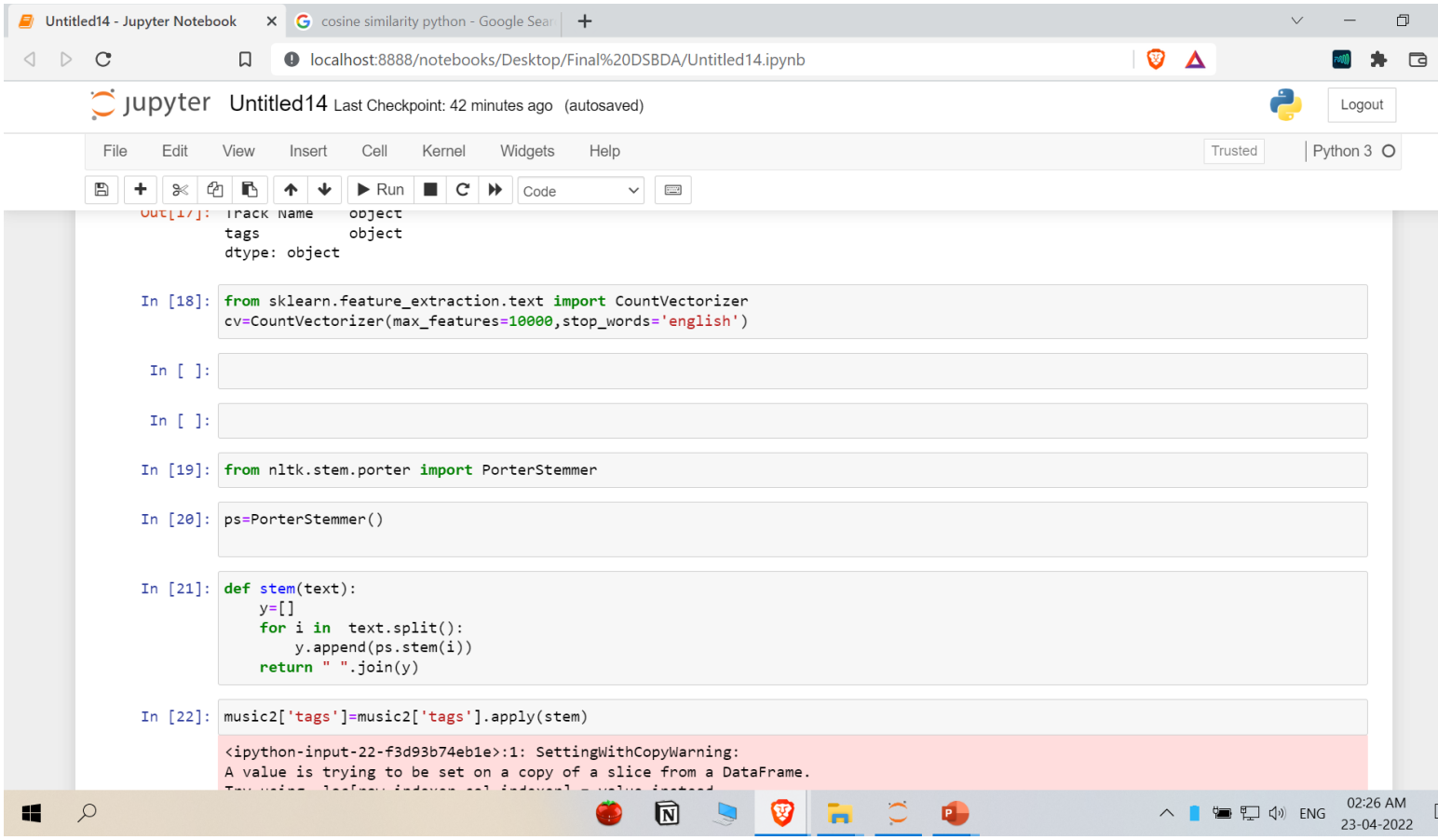
  The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

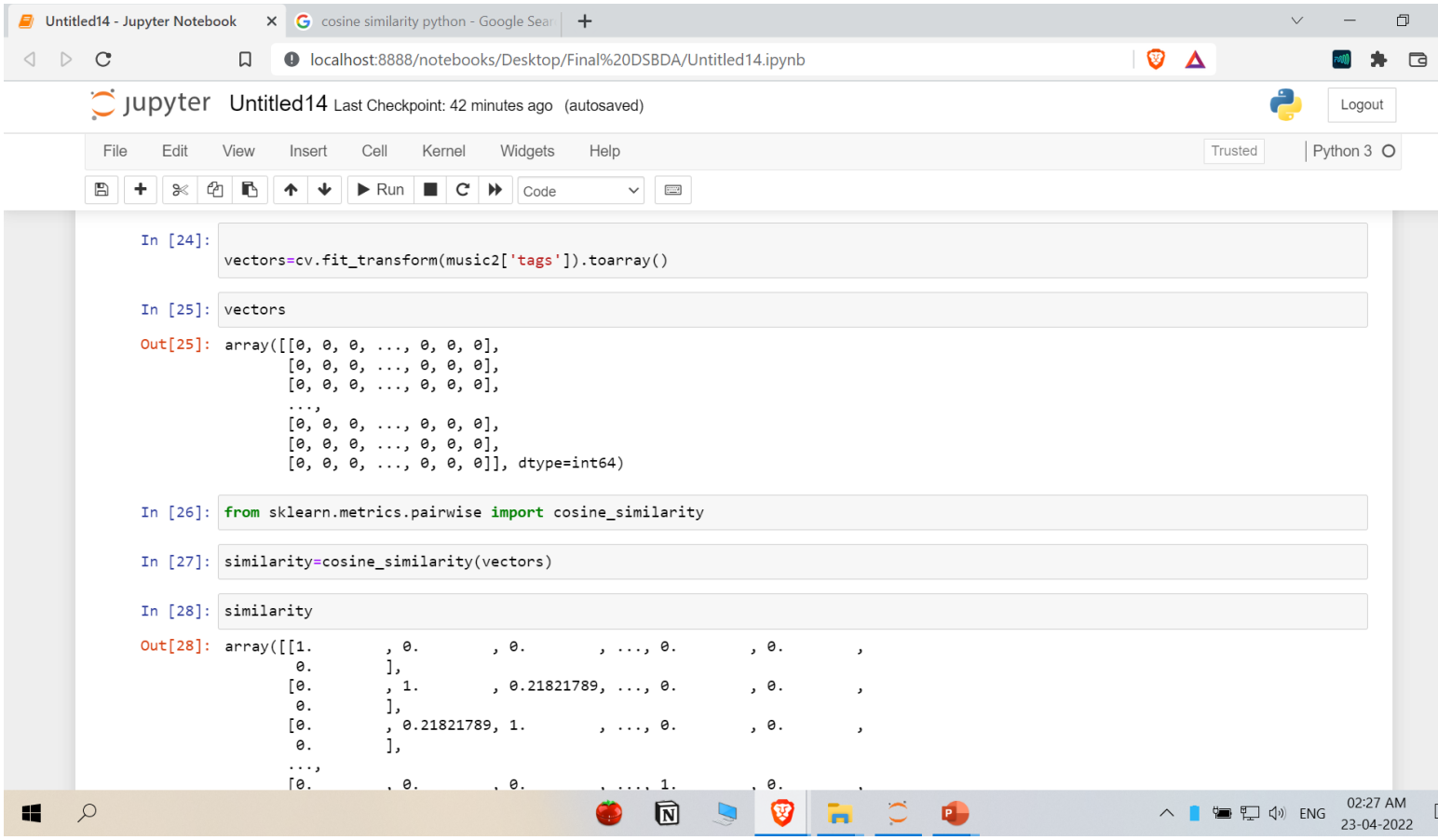
7 Outcome/ results of internship work (screenshots of work done)

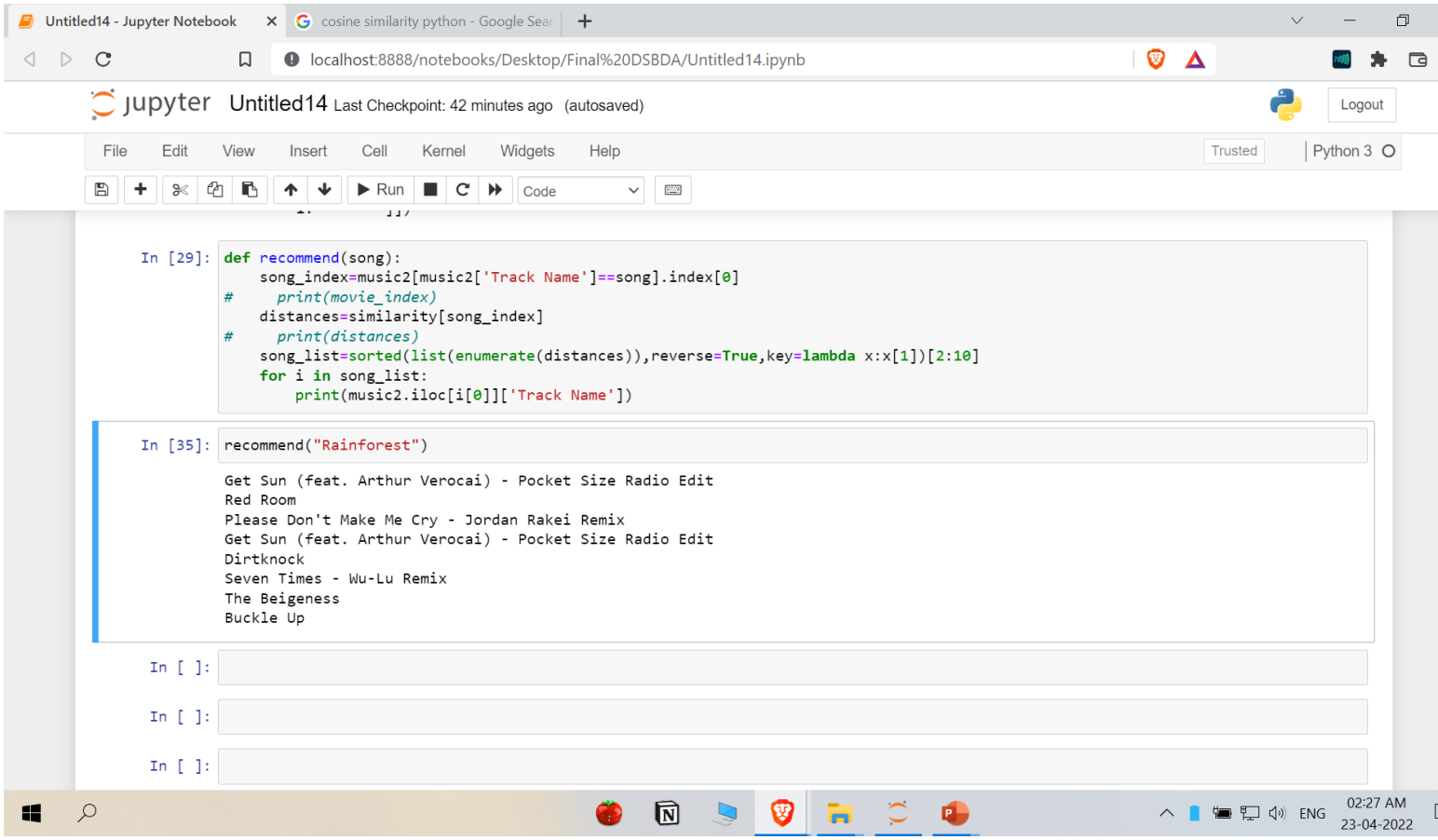


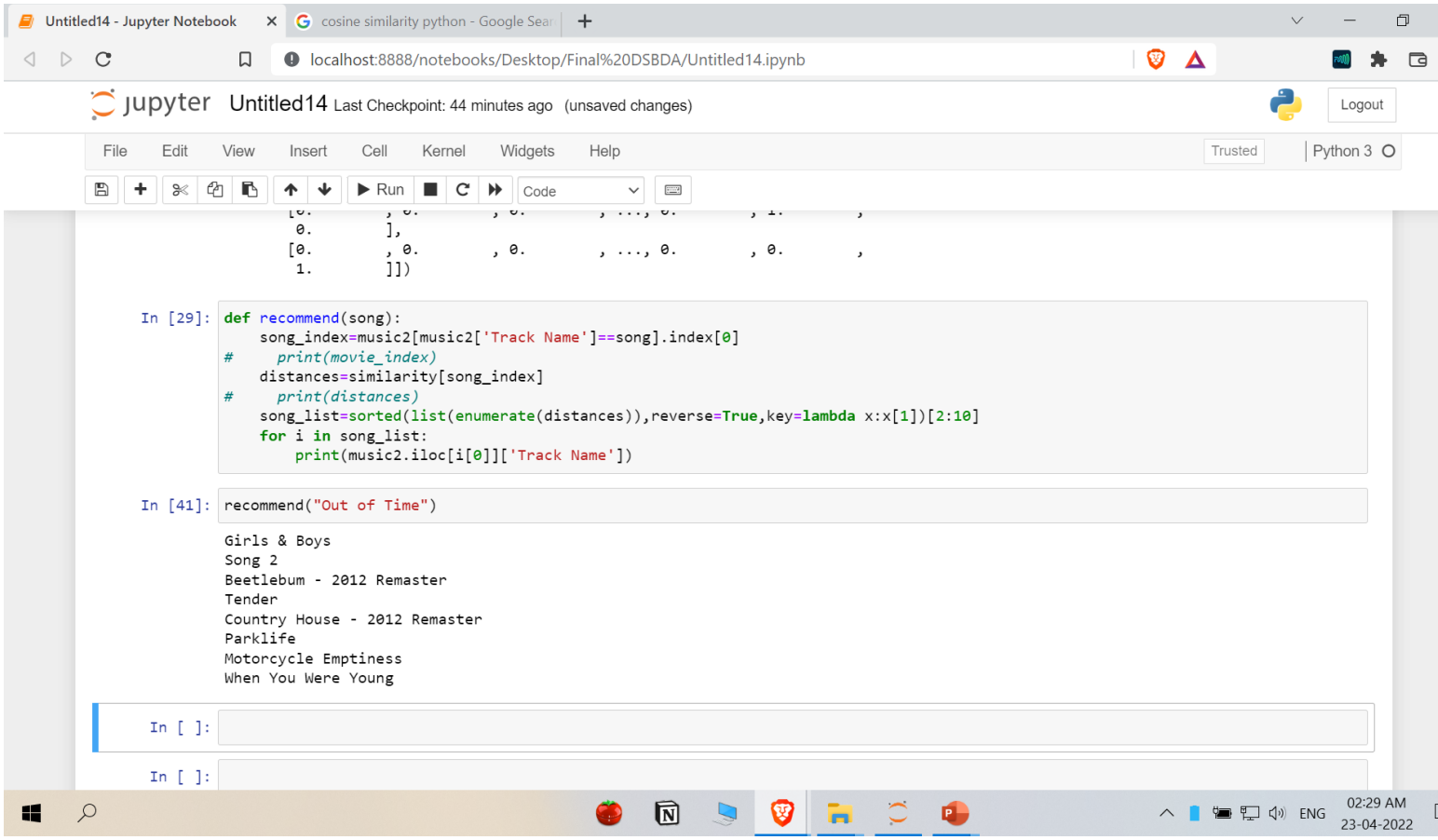

Untitled14 - Jupyter Notebook
cosine similarity python - Google Sear
localhost:8888/notebooks/Desktop/Final%20DSBDA/Untitled14.ipynb
jupyter Untitled14 Last Checkpoint: 40 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [6]: music1['Playlist']=music1['Playlist'].apply(lambda x:x.replace(' ',''))
<ipython-input-6-aa5fba2ff65f>: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
music1['Playlist']=music1['Playlist'].apply(lambda x:x.replace(' ',''))
In [7]: music1['Genres']=music1['Genres'].apply(lambda x:x.replace(' ',''))
<ipython-input-7-0631fc235752>: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
music1['Genres']=music1['Genres'].apply(lambda x:x.replace(' ',''))
In [8]: music1.Genres[0]
Out[8]: "['britpop','madchester','newwave','newwavepop','permanentwave','poprock','rock']"
In [9]: # music1['Track Name']=music1['Track Name'].apply(lambda x:x.replace(' ',''))
In [10]: music1['Artist Name']=music1['Artist Name'].apply(lambda x:x.replace(' ',''))
ENG
02:25 AM
23-04-2022

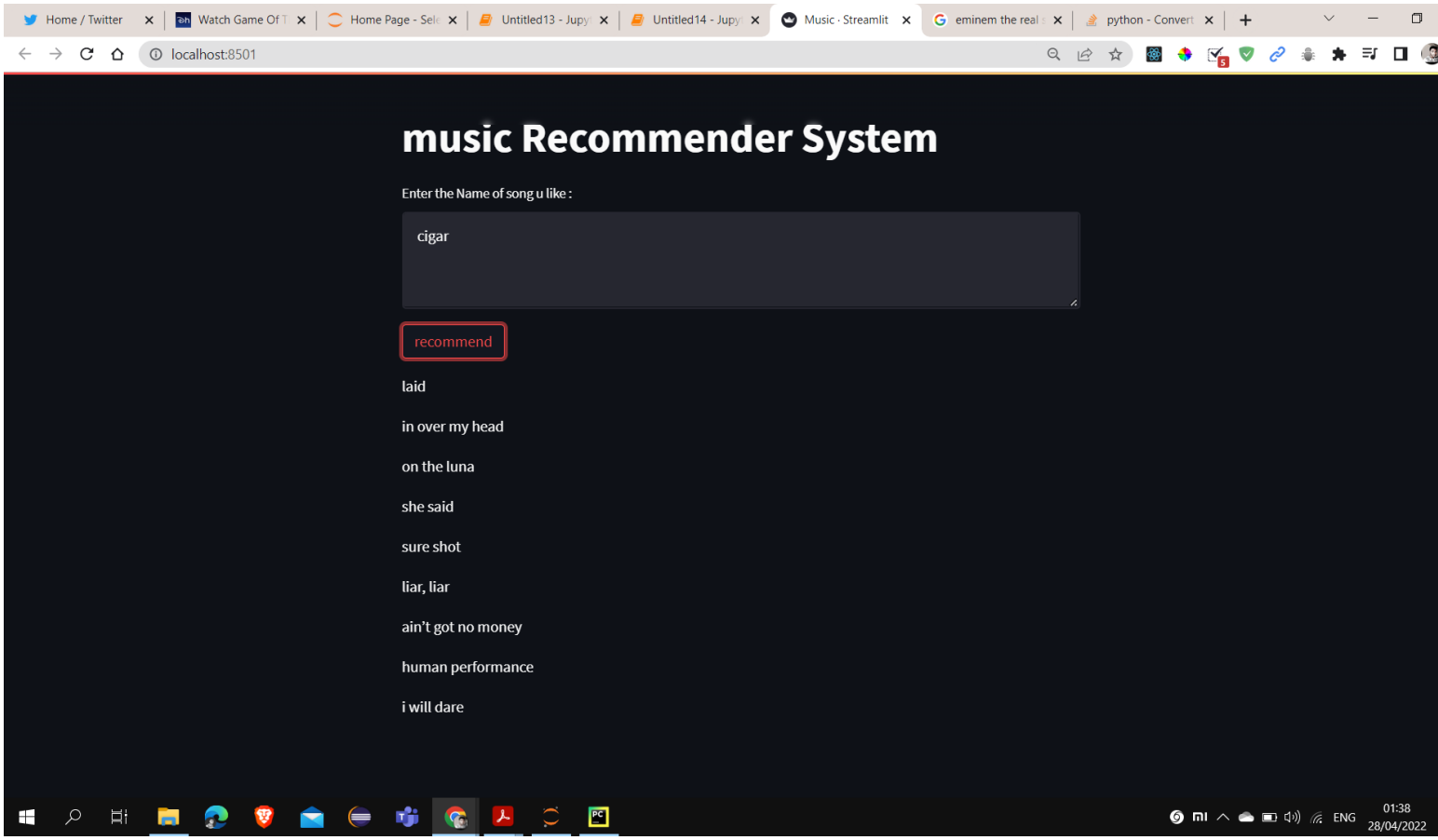

Untitled14 - Jupyter Notebook
cosine similarity python - Google Sear
localhost:8888/notebooks/Desktop/Final%20DSBDA/Untitled14.ipynb
jupyter Untitled14 Last Checkpoint: 41 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [12]:
def genre(x):
st=""
for i in x:
if(i=='[' or i==']'or i=="'"):
continue
elif(i==','):
st+=' '
else:
st+=i
return st
In [13]:
music1['Genres']=music1['Genres'].apply(genre)
<ipython-input-13-d19f65e4f445>: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
music1['Genres']=music1['Genres'].apply(genre)
In [14]:
music1.Genres[0]
Out[14]:
'britpop madchester newwave newwavepop permanentwave poprock rock'
In [15]:
music1['tags']=music1['Artist Name']+music1['Genres']+music1['Playlist']
ENG
02:26 AM
23-04-2022

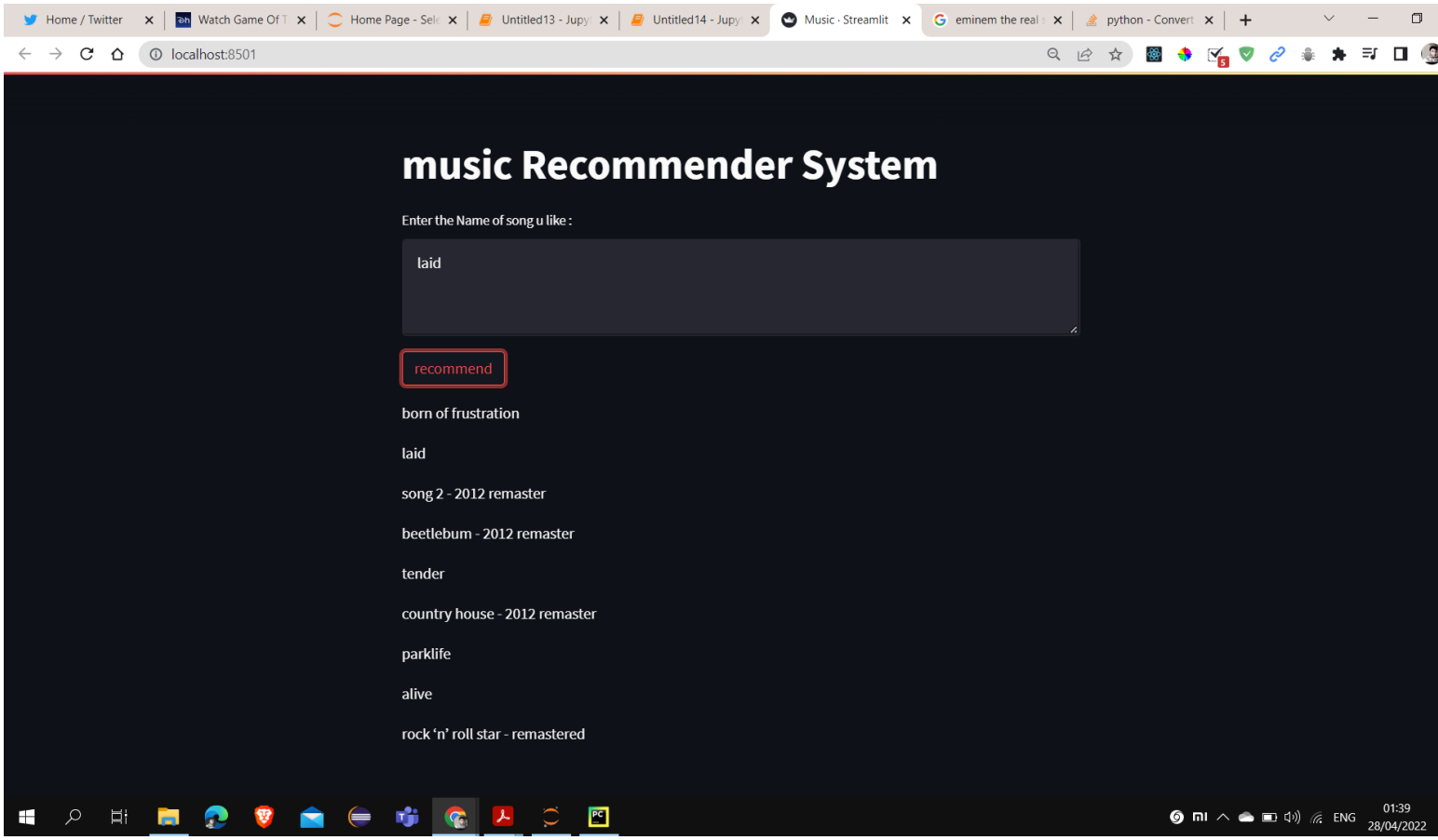

Untitled14 - Jupyter Notebook
cosine similarity python - Google Sear
localhost:8888/notebooks/Desktop/Final%20DSBDA/Untitled14.ipynb
jupyter Untitled14 Last Checkpoint: 42 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Out[17]: Track Name object
tags object
dtype: object
In [18]: from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer(max_features=10000,stop_words='english')
In [ ]:
In [ ]:
In [19]: from nltk.stem.porter import PorterStemmer
In [20]: ps=PorterStemmer()
In [21]: def stem(text):
y=[]
for i in text.split():
y.append(ps.stem(i))
return " ".join(y)
In [22]: music2['tags']=music2['tags'].apply(stem)
<ipython-input-22-f3d93b74eb1e>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
ENG
02:26 AM
23-04-2022


Untitled14 - Jupyter Notebook
cosine similarity python - Google Sear
localhost:8888/notebooks/Desktop/Final%20DSBDA/Untitled14.ipynb
jupyter Untitled14 Last Checkpoint: 42 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [24]:
vectors=cv.fit_transform(music2['tags']).toarray()
In [25]:
vectors
Out[25]:
array([[0, 0, 0, ..., 0, 0, 0],
[0, 0, 0, ..., 0, 0, 0],
[0, 0, 0, ..., 0, 0, 0],
...,
[0, 0, 0, ..., 0, 0, 0],
[0, 0, 0, ..., 0, 0, 0],
[0, 0, 0, ..., 0, 0, 0]], dtype=int64)
In [26]:
from sklearn.metrics.pairwise import cosine_similarity
In [27]:
similarity=cosine_similarity(vectors)
In [28]:
similarity
Out[28]:
array([[1. , 0. , 0. , ..., 0. , 0. ,
0.],
[0. , 1. , 0.21821789, ..., 0. , 0. ,
0.],
[0. , 0.21821789, 1. , ..., 0. , 0. ,
0.],
...,
[0. , 0. , 0. , ..., 1. , 0. ,
ENG
02:27 AM
23-04-2022


Untitled14 - Jupyter Notebook
cosine similarity python - Google Sear
localhost:8888/notebooks/Desktop/Final%20DSBDA/Untitled14.ipynb
jupyter Untitled14 Last Checkpoint: 42 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [29]:
def recommend(song):
song_index=music2[music2['Track Name']==song].index[0]
# print(movie_index)
distances=similarity[song_index]
# print(distances)
song_list=sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[2:10]
for i in song_list:
print(music2.iloc[i[0]]['Track Name'])
In [35]:
recommend("Rainforest")
Get Sun (feat. Arthur Verocai) - Pocket Size Radio Edit
Red Room
Please Don't Make Me Cry - Jordan Rakei Remix
Get Sun (feat. Arthur Verocai) - Pocket Size Radio Edit
Dirtknock
Seven Times - Wu-Lu Remix
The Beigeness
Buckle Up
In [ ]:
In [ ]:
In [ ]:
ENG
02:27 AM
23-04-2022


Untitled14 - Jupyter Notebook
cosine similarity python - Google Sear
localhost:8888/notebooks/Desktop/Final%20DSBDA/Untitled14.ipynb
jupyter Untitled14 Last Checkpoint: 44 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
0.],
[0. , 0. , 0. , ..., 0. , 0. ,
1.]])
In [29]:
def recommend(song):
song_index=music2[music2['Track Name']==song].index[0]
# print(movie_index)
distances=similarity[song_index]
# print(distances)
song_list=sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[2:10]
for i in song_list:
print(music2.iloc[i[0]]['Track Name'])
In [41]:
recommend("Out of Time")
Girls & Boys
Song 2
Beetlebum - 2012 Remaster
Tender
Country House - 2012 Remaster
Parklife
Motorcycle Emptiness
When You Were Young
In [ ]:
In [ ]:
ENG
02:29 AM
23-04-2022


Home / Twitter
Watch Game Of T
Home Page - Sel
Untitled13 - Jupy
Untitled14 - Jupy
Music · Streamlit
eminem the real s
python - Convert
localhost:8501
music Recommender System
Enter the Name of song u like :
cigar
recommend
laid
in over my head
on the luna
she said
sure shot
liar, liar
ain't got no money
human performance
i will dare
ENG
01:38
28/04/2022


Home / Twitter
Watch Game Of T
Home Page - Sel
Untitled13 - Jupy
Untitled14 - Jupy
Music · Streamlit
eminem the real s
python - Convert
localhost:8501
music Recommender System
Enter the Name of song u like :
laid
recommend
born of frustration
laid
song 2 - 2012 remaster
beetlebum - 2012 remaster
tender
country house - 2012 remaster
parklife
alive
rock 'n' roll star - remastered
ENG
01:39
28/04/2022

DEPARTMENT OF COMPUTER ENGINEERING

SCTR's Pune Institute of Computer Technology
Dhankawadi, Pune
Maharashtra 411043

CERTIFICATE

This is to certify that the SPPU Curriculum-based internship report entitled
"Content Based Music Recommendation System
under Big Data and Text Analytics Group, PICT. "

Submitted by
Ashutosh Sanjay Thite
(Roll No. 31462)

has satisfactorily completed the curriculum-based internship under the guidance of *Dr. A. R. Deshpande* towards the partial fulfillment of third year Computer Engineering Semester VI, Academic Year 2021-22 of Savitribai Phule Pune University.

Dr. A. R. Deshpande
Internship Guide
PICT, Pune

Dr. G. V. Kale
Head
Department of Computer Engineering
PICT, Pune

Place: Pune
Date: 28 / 04 / 2022