**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

Khoa Kỹ Thuật – Công Nghệ

**BÁO CÁO CUỐI KỲ**

**HỌC MÁY**

Logo, company name

Description automatically generated

**XỬ LÝ DATA ÂM NHẠC DỰA TRÊN PHÂN CỤM K-MEANS**

**Tên người thực hiện:** Trần Phương Trường (1824801040041)

**Lớp:** D18HT01 **GVHD:** Dương Thị Kim Chi

# Bình Dương, tháng 12, 2021

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của thầy Dương Thị Kim Chi. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây.

Những screenshot, phân tich dự án và hình ảnh lấy từ dự án phần mềm tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong báo cáo còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung báo cáo của mình.** Trường Đại học Thủ Dầu Một không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*Bình Dương, ngày 12 tháng 12 năm 2021*

*Người thực hiện  
 (ký tên và ghi rõ họ tên)*

A white paper with writing on it

Description automatically generated with low confidence

**MỤC LỤC**

[Chương 1. TỔNG QUAN ĐỀ TÀI 1](#_Toc90486791)

[1 Mục tiêu đề tài. 1](#_Toc90486792)

[2 Câu hỏi nghiên cứu. 1](#_Toc90486793)

[Chương 2. CƠ SỞ LÝ THUYẾT 2](#_Toc90486794)

[1 Giới thiệu về Google Colab. 2](#_Toc90486795)

[2 Giới thiệu về PyCharm. 3](#_Toc90486796)

[3 Thuật toán K-Means. 4](#_Toc90486797)

[Chương 3. XỬ LÝ BỘ DỮ LIỆU 5](#_Toc90486798)

[1 Xây dựng thuật toán K-Means. 5](#_Toc90486799)

[2 Áp dụng với dữ liệu. 6](#_Toc90486800)

[Chương 4. KẾT QUẢ THỰC TẾ 8](#_Toc90486801)

[1 Kết quả đầu ra. 8](#_Toc90486802)

[2 Nhận xét. 10](#_Toc90486803)

[Chương 5. TỔNG KẾT 11](#_Toc90486804)

[1 Kết quả đạt được. 11](#_Toc90486805)

[2 Hướng phát triển. 11](#_Toc90486806)

[TÀI LIỆU THAM KHẢO 12](#_Toc90486807)

**DANH MỤC CÁC HÌNH**

[Hình 2.1. HTML5 2](#_Toc90486830)

[Hình 2.2. Một ví dụ của Google Colab 3](#_Toc90486831)

[Hình 2.3. PyCharm 3](#_Toc90486832)

[Hình 2.4. Một ví dụ về PyCharm 4](#_Toc90486833)

[Hình 4.1. Thuật toán đã đưa dữ liệu từ *spotifydata.csv* vào thành công 8](#_Toc90486834)

[Hình 4.2. Phân cụm K-means của X 8](#_Toc90486835)

[Hình 4.3. Các loại phân cụm của X 9](#_Toc90486836)

[Hình 4.4. Cost của X theo số thứ tự của phân cụm 9](#_Toc90486837)

[Hình 4.5. Sự phân bố của các phần tử trong mảng X 10](#_Toc90486838)

2. Chương 1. TỔNG QUAN ĐỀ TÀI
   1. Mục tiêu đề tài.

Việc phân tích nghiên cứu bộ dữ liệu luôn đóng vai trò quan trọng trong việc xây dựng các thuật toán AI cho việc tìm kiếm lẫn quản lý cơ sở dữ liệu của tất cả mọi thứ liên quan đến công nghệ thông tin. Những thông tin của dữ liệu đều sẽ được phân loại để có được tính tổ chức, góp phần phát triển những thuật toán, vượt qua các rào cản và thách thức khiến cho đời sống xã hội ngày càng tiến bộ hơn. Trong bối cảnh mà công nghệ càng ngày càng phát triển, cơ sở dữ liệu ngày càng trở nên đồ sộ và phức tạp hơn thì việc phân tích nghiên cứu bộ dữ liệu đang ngày càng trở thành xu thế mới nhằm giải quyết các bài toán phức tạp mà trong quá khứ khó có thể làm được.

Nhận thức được vai trò của việc quản lý cơ sở dữ liệu âm nhạc, em đã chọn đề tài xử lý data âm nhạc bằng thuật toán phân cụm K-Means.

* 1. Câu hỏi nghiên cứu.

Các trang web âm nhạc đều lưu trữ một lượng lớn các bài hát, vì thế, chúng đều phải đòi hỏi một thuật toán phân loại cũng như xác định tính chất các bài hát lại nhằm khiến cho trang web trở tiến có tổ chức cả về mặt nội dung lẫn hình thức, đồng thời thu thập các bài hát có chọn lọc để không dẫn đến tình trạng quá tải.

Ở bài toán này, chúng ta sẽ sử dụng hai thuật toán phân cụm đã nói ở trên để phân loại ra các trường cũng như những mối tương quan giữa các đối tượng với nhau, các đối tượng với các trường và các trường với nhau.

1. Chương 2. CƠ SỞ LÝ THUYẾT
   1. Giới thiệu về Google Colab.

Google Colaboratory là một công cụ môi trường miễn phí phát triển từ Project Jupyter, Colab chạy bằng điện toán đám mây và lưu trữ tại Google Drive. Colab ban đầu là một sản phẩm trong nội bộ Google; người ta nỗ lực biến Colab thành một mã nguồn mở và hoạt động đại chúng nhưng đã thất bại.

Logo

Description automatically generated

Hình 2.1. HTML5

Google Colab phát hành gần với thời điểm thành lập Project Jupyter (Đây là sản phẩm kết tinh trong quá trình phát triển Project Jupyter), tức là khoảng năm 2015. Colab là một công cụ cho học sinh sinh viên, những nhà nghiên cứu AI, những lập trình viên hay những nhà khoa học dữ liệu phát triển các chương trình phân tích, trực quan hóa, hay đánh giá các dữ liệu đầu vào nhằm đưa ra những kết quả có cơ sở về bộ dữ liệu trên và từ đó nhìn ra những mẫu và dự đoán chúng.

Google Colab bao gồm các thành phần:

* Cell: Nơi bạn nhập code và chạy chương trình. Cần phải kết nối với một môi trường để code mô phỏng, khai phá và phân tích dữ liệu.
* Những thành phần của Python:
  + Danh pháp hệ thống (System alias): Những toán tử có trong Python.
  + Magic: Những dòng code chạy những quy trình phức tạp để cho ra những kết quả đặc biệt.
  + Code tự động thực hiện (Automatic completion): Những dòng code được lập trình sẵn, có nhiệm vụ cố định và thực hiện ngay khi chạy.

Graphical user interface, text, application, email

Description automatically generated

Hình 2.2. Một ví dụ của Google Colab

## Google Colab giúp việc nghiên cứu bộ dữ liệu phát triển lên một tầm cao mới, giúp tạo ra những thuật toán hiệu quả trong việc phân loại, tìm kiếm và giúp ích cho việc phát triển đời sống xã hội, nhất là ở thời đại công nghệ này.

* 1. Giới thiệu về PyCharm.

PyCharm là một công cụ lập môi trường tích hợp dùng cho việc phát triển chương trình bằng Python. Công cụ sở hữu cho mình các bộ phận như thành phần phân tích dữ liệu và code, thành phần debug đồ họa, thành phần tester tích hợp bên trong, kết hợp với các hệ thống xử lý phiên bản và hỗ trợ cho việc xây dựng trang web cũng như phát triển khoa học dữ liệu.

Text, logo

Description automatically generated

Hình 2.3. PyCharm

PyCharm được ra mắt trong công chúng vào ngày 03/02/2010, được phát hành trên các nền tảng như Windows, macOS và Linux. PyCharm đã phát hành ra các phiên bản khác nhau, bản mới nhất là 2021.3 phát hành vào ngày 30/11/2021.

PyCharm được hỗ trợ những tính năng:

* Thành phần hỗ trợ lập trình và xử lý code, với khả năng tự động điền, đánh dấu lỗi cú pháp, kết hợp với công cụ đánh dấu lỗi lint và cách sửa lỗi.
* Thành phần định hướng dự án và code.
* Khả năng thay đổi thông tin, chi tiết sản phẩm.
* Được hỗ trợ bởi các trang web như Django, web2py, Flask,...
* Công cụ debug.
* Công cụ kiểm tra code.
* Thành phần cập nhật phiên bản sản phẩm.

A screenshot of a computer

Description automatically generated

Hình 2.4. Một ví dụ về PyCharm

Như Google Colab, PyCharm cũng đóng một vai trò quan trọng cho việc phát triển khoa học dữ liệu lên một tầm cao mới, giúp tạo ra những thuật toán hiệu quả trong việc phân loại, tìm kiếm và giúp ích cho việc phát triển đời sống xã hội.

* 1. Thuật toán K-Means.

Đây là một thuật toán phân tích đánh giá dữ liệu qua việc chọn một vài phần tử trong bộ dữ liệu làm trung tâm các cụm và phân bố các loại dữ liệu vào những cụm trên. Đây là bước tiền đề cho các thuật toán cao cấp hơn như PAM,... và bước tiến đầu trong việc nghiên cứu dữ liệu bằng phân cụm.

1. Chương 3. XỬ LÝ BỘ DỮ LIỆU
   1. Xây dựng thuật toán K-Means.

Bước 1: Gọi *matplotlib notebook* và lớp Kmeans từ tệp *k\_means.py*.

%matplotlib notebook

from k\_means import KMeans

­Bước 2: Gọi các thư viện cần thiết cho thuật toán.

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import cv2

from skimage import io

import time

from mpl\_toolkits.axes\_grid1 import make\_axes\_locatable

from sklearn.cluster import KMeans as Kmeans\_

from sklearn.metrics import mean\_squared\_error

import itertools

from sklearn.metrics import silhouette\_samples, silhouette\_score

Bước 3: Tạo hàm *elbow\_method* để vẽ đồ thị tương quan giữa cost của các phân cụm.

def elbow\_method(X, max\_k = 10):

    costs = []

    for k in range(2, max\_k):

        model = KMeans(n\_clusters=k, init\_method='var\_part')

        model.fit(X)

        costs.append(model.cost\_)

    plt.close()

    plt.plot(list(range(2, max\_k)), costs)

    plt.xlabel("# of clusters (K)")

    plt.ylabel("Cost")

Bước 4: Tạo hàm *draw\_scatterplot* vẽ đồ thị giữa hai trường khác nhau.

def draw\_scatterplot(x\_data, x\_label, y\_data, y\_label):

    fig = plt.figure(figsize=(5,5))

    ax = fig.add\_subplot(111)

    plt.xlim(0, 5)

    plt.ylim(0, 5)

    ax.set\_xlabel(x\_label)

    ax.set\_ylabel(y\_label)

    ax.scatter(x\_data, y\_data, s=30)

* 1. Áp dụng với dữ liệu.

Bước 1: Gọi data *spotifydata.csv* vào thuật toán.

data=pd.read\_csv("spotifydata.csv", index\_col="no")

data.head()

Bước 2: Tạo mảng đa chiều X từ hai thuộc tính *acousticness* và *danceability*.

X=data[['acousticness', 'danceability']]

Bước 3: Vẽ sơ đồ K-means mảng X.

plt.close()

plt.style.use('seaborn')

plt.scatter(X['acousticness'], X['danceability'])

plt.xlabel('acousticness')

plt.ylabel('danceability')

Bước 4: Tính giá trị model của X.

model=KMeans(max\_iter=300, tolerance=0.0001, n\_clusters=5, runs=100)

(clusters, data\_with\_clusters)=model.fit(X)

model.cost\_

Bước 5: Vẽ sơ đồ K-means đánh dấu các loại cụm trong X và phần tử trung tâm của chúng.

plt.close()

for i, cluster\_mean in enumerate(clusters):

  data\_cluster\_i=data\_with\_clusters[data\_with\_clusters[:, -1]==i]

  plt.scatter(data\_cluster\_i[:, 0], data\_cluster\_i[:, 1], label='Cluster '+str(i))

  plt.plot(cluster\_mean[0], cluster\_mean[1], label='Centroid '+str(i), marker='\*', markersize=15, markeredgecolor="k", markeredgewidth=1)

  plt.xlabel('acousticness')

  plt.ylabel('danceability')

  plt.style.use('seaborn')

  plt.legend()

Bước 6: Vẽ sơ đồ *elbow\_method* của X.

elbow\_method(X)

Bước 7: Gọi inline và vẽ *draw\_scatterplot* của X.

%matplotlib inline

draw\_scatterplot(X['acousticness'], 'Acousticness', X['danceability'], 'Danceability')

1. Chương 4. KẾT QUẢ THỰC TẾ
   1. Kết quả đầu ra.

Bảng đầu ra 1 cho thấy thuật toán đã cho dữ liệu vào thành công.

A picture containing table

Description automatically generated

Hình 4.1. Thuật toán đã đưa dữ liệu từ *spotifydata.csv* vào thành công

Bảng đầu ra 2 cho thấy mảng X được phân bố vào bên trái phía trên là chủ yếu.

Chart, scatter chart

Description automatically generated

Hình 4.2. Phân cụm K-means của X

Bảng đầu ra 3 cho thấy các phần tử trong mảng X được chia ra làm 5 phần và các trung tâm được phân bố như hình sau.

Chart, scatter chart

Description automatically generated

Hình 4.3. Các loại phân cụm của X

Bảng đầu ra 4 cho thấy cost của từng phân cụm ngày càng giảm và ít tập trung hơn.

Chart, line chart

Description automatically generated

Hình 4.4. Cost của X theo số thứ tự của phân cụm

Bảng đầu ra cuối cùng cho thầy các phần tử trong mảng X chỉ tập trung trong khoảng từ 0.0 đến 1.0.

Chart, histogram

Description automatically generated

Hình 4.5. Sự phân bố của các phần tử trong mảng X

* 1. Nhận xét.

Thuật toán K-means cho thấy mảng X được phân bố ra làm 5 loại cụm và các phần tử hầu như đều tập trung vào ba cụm đầu. Hầu hết các giá trị phần tử trong mảng X chỉ tập trung trong khoảng từ 0.0 đến 1.0 và các phân cụm đầu thường có cost cao hơn.

Dựa vào kết quả thu được ở trên, các bản nhạc trong data có độ “Acousticness” chủ yếu tập trung trong khoảng từ 0.0 đến 0.45 và “Danceability” chủ yếu tập trung từ 0.3 đến 0.85.

Đường link BCK Google Colab:

<https://colab.research.google.com/drive/1Tj4med1jGTsGqdSpxWRYFxZnuVQIfqcJ#scrollTo=6EadrBB2RTFq>.

Đường link github:

<https://github.com/andrewt18031999/HocMay_BCCK>.

1. Chương 5. TỔNG KẾT
   1. Kết quả đạt được.

* Cài đặt thành công thuật toán K-means.
* Kết quả chương trình hoạt động tốt và bảng kết quả ra chi tiết.
* Mặc dù vậy, chương trình có một bug đôi lúc khiến việc ra bảng thống kê phân cụm không ra hình.
  1. Hướng phát triển.

Đối với các nhà nghiên cứu công nghệ thông tin và nhà khoa học dữ liệu thì thuật toán K-means là bước đầu cho việc phát triển nghiên cứu dữ liệu theo hướng chuyên sâu hơn, đây cũng là tiền đề cho các thuật toán như PAM,... phát triển và tạo ra hướng đi mới, chân trời mới cho ngành nghiên cứu khoa học dữ liệu.

1. TÀI LIỆU THAM KHẢO

Dave Gershgorn, "Nerds rejoice: Google just released its internal tool to collaborate on AI", *Quartz*, October 27th, 2017.

Google Colab, "Open in Colab GitHub repository", *GitHub*, September 06th, 2018.

JuliaTPU, Julia on TPUs, *GitHub*, December 03rd, 2019.

"JetBrains PyCharm - New Python IDE by creators of IntelliJ", *news.ycombinator.com*, February 03rd, 2010.

Roberto Pesce, "Introducing PyCharm 2021.3!", *Jetbrains*, December 01st, 2021.

Ernst Haagsman, "Collaboration with Anaconda, Inc", *eWeek*, 4 April 2019.

Dmitry Jemerov, "PyCharm 3.0 Community Edition source code now available", *JetBrains*, 22 October 2013.

"Scientific & Data Science Tools - Features | PyCharm", *JetBrains*, January 02nd, 2020.

<https://github.com/tugrulhkarabulut/K-Means-Clustering>.

<https://github.com/JangirSumit/kmeans-clustering>.