

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NÔNG LÂM THÀNH PHỐ HỒ CHÍ MINH
KHOA MÔI TRƯỜNG VÀ TÀI NGUYÊN



KHÓA LUẬN TỐT NGHIỆP

ỨNG DỤNG HỌC MÁY TRONG XÂY DỰNG
HỆ THỐNG ĐỀ XUẤT PHIM HOẠT HÌNH

Họ và tên sinh viên: NGUYỄN THỊ HUỲNH NHƯ

Ngành: Hệ thống thông tin

Niên khóa: 2020 – 2024

Thành phố Hồ Chí Minh, tháng 07 năm 2024

**ỨNG DỤNG HỌC MÁY TRONG XÂY DỰNG
HỆ THỐNG ĐỀ XUẤT PHIM HOẠT HÌNH**

Tác giả

NGUYỄN THỊ HUỲNH NHƯ

Khóa luận được đệ trình để đáp ứng yêu cầu
cấp bằng Kỹ sư ngành Hệ thống thông tin

Giáo viên hướng dẫn:

TS. Nguyễn Đình Long

Thành phố Hồ Chí Minh, tháng 07 năm 2024

LỜI CẢM ƠN

Trước tiên, tôi xin chân thành cảm ơn thầy TS. Nguyễn Đình Long, Giảng viên thỉnh giảng tại Bộ Môn Hệ thống Thông tin, Khoa Tài Nguyên và Môi Trường, Trường Đại học Nông Lâm thành phố Hồ Chí Minh, người đã hướng dẫn tôi hoàn thành đề tài tốt nghiệp này. Cảm ơn thầy đã tận tình chỉ bảo, hỗ trợ và động viên tôi trong suốt thời gian qua.

Tôi xin gửi lời tri ân sâu sắc đến tất cả quý thầy cô Bộ môn Tài nguyên và GIS, Khoa Môi trường và Tài nguyên Trường Đại học Nông Lâm thành phố Hồ Chí Minh. Cảm ơn quý thầy cô về những kiến thức và giúp đỡ chân tình đã dành cho tôi trong bốn năm học tập tại trường.

Cuối cùng, con xin nói lời biết ơn sâu sắc đối với cha mẹ đã chăm sóc, nuôi dạy con thành người và luôn động viên tinh thần cho con để con yên tâm học tập.

Thành phố Hồ Chí Minh, tháng 07 năm 2024

Nguyễn Thị Huỳnh Như

Khoa Môi trường và Tài nguyên

Trường Đại học Nông Lâm thành phố Hồ Chí Minh

Số điện thoại: 0333087889

Email: nguyenthihuynhnhu271025@gmail.com

TÓM TẮT

Khóa luận tốt nghiệp “Ứng dụng học máy trong xây dựng hệ thống đề xuất phim hoạt hình” đã được thực hiện trong khoảng thời gian từ tháng 03 năm 2024 đến tháng 07 năm 2024. Mục tiêu của đề tài bao gồm: xây dựng web tích hợp hệ thống đề xuất phim hoạt hình, bằng việc biến đổi dữ liệu thứ cấp MyAnimeList của Kaggle được tải lên năm 2020 thành thông tin bằng thuật toán lọc cộng tác nhằm đưa ra các đề xuất phim hoạt hình kết hợp với việc xây dựng giao diện đề xuất phim hoạt hình bằng ngôn ngữ lập trình Python. Phương pháp tiếp cận của đề tài là thuật toán lọc cộng tác và thuật toán KNN của học máy. Theo đó, học máy có chức năng tạo mô hình đề xuất để đưa ra các đề xuất phim hoạt hình.

Kết quả đạt được của khóa luận là xây dựng mô hình huấn luyện dữ liệu đề xuất phim hoạt hình bằng thuật toán lọc cộng tác và thuật toán KNN, xây dựng web đơn giản bằng ngôn ngữ lập trình Python tích hợp hệ thống đề xuất phim hoạt hình bao gồm tên phim hoạt hình và hình ảnh phim hoạt hình.

MỤC LỤC

LỜI CẢM ƠN.....	i
TÓM TẮT.....	iii
MỤC LỤC	iv
DANH MỤC BẢNG	vi
DANH MỤC TỪ VIẾT TẮT	vii
DANH MỤC HÌNH ẢNH.....	viii
1.1. Tính cấp thiết của đề tài	1
1.2. Mục tiêu nghiên cứu	2
1.3. Đối tượng nghiên cứu	2
Chương 2: TỔNG QUAN TÀI LIỆU.....	3
2.1. Tổng quan đối tượng nghiên cứu	3
2.1.1. Khái niệm hệ thống đề xuất.....	3
2.1.2. Khái niệm phim hoạt hình	7
2.2. Tổng quan phương pháp nghiên cứu.....	7
2.3. Tình hình nghiên cứu	11
Chương 3: DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU	14
3.1. Dữ liệu nghiên cứu.....	14
3.2. Phương pháp nghiên cứu.....	17
3.2.1. Cách tiếp cận.....	17
3.2.2. Phương pháp nghiên cứu	18
3.2.3. Phần mềm và thư viện	19
Chương 4: KẾT QUẢ VÀ THẢO LUẬN.....	21
4.1. Mô hình đề xuất phim hoạt hình	21
4.1.1. Thu thập dữ liệu	21

4.1.2. Chuẩn bị dữ liệu.....	22
4.1.3. Chọn mô hình	22
4.1.4. Điều chỉnh thông số	23
4.1.5. Kết quả đề xuất	23
4.2. Ứng dụng web đề xuất phim hoạt hình	24
4.2.1. Thiết kế chức năng.....	24
4.2.2. Thiết kế giao diện	25
CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ.....	28
5.1. Kết luận	28
5.2. Ý nghĩa	28
5.3. Kiến nghị.....	28
TÀI LIỆU THAM KHẢO	29
Phụ lục 1: GIẢI THÍCH LẬP TRÌNH	32

DANH MỤC BẢNG

Bảng 2.1: Ma trận tương quan giữa người dùng và phim hoạt hình	8
--	---

DANH MỤC TỪ VIẾT TẮT

ALS	Alternating Least Squares	Phương pháp bình phương ít luân phiên
AI	Artificial Intelligence	Trí tuệ nhân tạo
BSD	Berkeley Software Distribution	Một loại giấy phép phần mềm miễn phí
CF	Collaborative Filtering	Hệ thống lọc cộng tác
EDA	Exploratory Data Analysis	Phân tích khám phá dữ liệu
GNN	Graph Neural Networks	Mạng nơ-ron đồ thị
GUI	Graphical User Interface	Giao diện đồ họa người dùng
KNN	K-nearest neighbors	Thuật toán K láng giềng gần nhất
ML	Machine Learning	Học máy
MPAA	Motion Picture Association of America	Hiệp hội Điện ảnh Hoa Kỳ
ODE	Ordinary Differential Equations	Phương trình vi phân thông thường
RS	Recommendation System	Hệ thống đề xuất
UNICEF	United Nations International Children's Emergency Fund	Quỹ Nhi đồng Liên Hợp Quốc
URL	Uniform Resource Locator	Địa chỉ web
Web-app	Web Application	Ứng dụng Web

DANH MỤC HÌNH ẢNH

Hình 2.1: Mối quan hệ giữa trí tuệ nhân tạo với hệ thống đề xuất.....	4
Hình 2.2: Các phân loại hệ thống đề xuất	5
Hình 2.3: Lệnh tạo ma trận.....	8
Hình 2.4: Kết quả tạo ma trận	8
Hình 2.5: Lệnh chuyển đổi NaN thành 0.....	9
Hình 2.6: Kết quả chuyển đổi.....	9
Hình 2.7: Lệnh tạo hàng không gian nén	9
Hình 2.8: Thuật toán KNN	10
Hình 2.9: Lệnh gọi thuật toán.....	11
Hình 2.10: Lệnh truyền hàng không gian nén	11
Hình 2.11: Kết quả lệnh gọi thuật toán KNN	11
Hình 3.1: Dữ liệu thứ cấp MyAnimeList	14
Hình 3.2: Thống kê tệp animes.csv	15
Hình 3.3: Thống kê tệp user.csv	15
Hình 3.4: Thống kê tệp user_watches.csv	16
Hình 3.5: Hệ thống phân loại phim MPAA	16
Hình 3.6: Phân loại theo giới tính của Users.....	17
Hình 3.7: Quy trình truy vấn và kết quả đề xuất được trả về	17
Hình 3.8: Hướng tiếp cận của thuật toán lọc cộng tác	18
Hình 3.9: Sơ đồ phương pháp nghiên cứu.....	19
Hình 4.1: Kết quả đọc tệp animes.csv	21
Hình 4.2: Kết quả đọc tệp user.csv	21
Hình 4.3: Kết quả đọc tệp user_watches.csv	22

Hình 4.4: Giá trị duy nhất và trung bình của anime_id và user_id trong tệp user_watches.csv	22
Hình 4.5: Khởi tạo hàm đề xuất	23
Hình 4.6: Kết quả trả về cho từ khóa tìm kiếm là “Minami-ke”	23
Hình 4.7: Đề xuất tên phim hoạt hình	24
Hình 4.8: Đề xuất hình ảnh của phim hoạt hình.....	24
Hình 4.9: Định cấu hình cài đặt mặc định của trang	25
Hình 4.10: Chèn hình ảnh.....	25
Hình 4.11: Tiêu đề của web-app.....	25
Hình 4.12: Thanh tìm kiếm.....	26
Hình 4.13: Giao diện web-app đề xuất phim hoạt hình.....	26

Chương 1: MỞ ĐẦU

1.1. Tính cấp thiết của đề tài

Theo DataReportal – Global Digital Insights tổng số người dùng Internet toàn cầu mới nhất tính đến tháng 1 năm 2023 ở mức 5,16 tỷ (Kemp, 2023). Lượng dữ liệu được tạo ra hằng ngày là vô cùng lớn. Việc hàng tỷ người dùng Internet toàn cầu không thể tiếp cận được với các tiện ích, sản phẩm,... là điều có thể xảy ra. Khi lượng thông tin có sẵn ngày càng tăng, nhu cầu về các công cụ giúp người dùng lọc và tìm kiếm thông tin phù hợp cũng trở nên cấp thiết hơn. Hệ thống đề xuất đóng vai trò quan trọng trong việc giải quyết vấn đề về quá tải thông tin bằng cách phân tích hành vi và sở thích của người dùng để đưa ra các đề xuất cá nhân hóa.

Hệ thống đề xuất không phải là một ý tưởng mới nhưng hiệu quả trong việc đưa ra các đề xuất dựa trên hành vi của người dùng, giảm công sức tìm kiếm và quá tải thông tin. Các hệ thống đề xuất có nguồn gốc từ "Usenet", một hệ thống thảo luận phân tán trên toàn thế giới bắt nguồn từ Đại học Duke vào cuối những năm 1970. Usenet hoạt động ở định dạng máy khách/máy chủ, cho phép đầu vào của người dùng được phân loại thành các "nhóm tin" cụ thể. Trong Usenet, các bài đăng của người dùng được phân loại thành các nhóm tin này, sau đó được chia thành các danh mục con nếu cần (Smith, 2012). Trong thập kỷ tiếp theo, hệ thống đề xuất với vai trò như một dạng của hệ hỗ trợ ra quyết định, cung cấp giải pháp mang tính cá nhân hóa mà không phải trải qua quá trình tìm kiếm phức tạp. Hệ thống đề xuất học từ người dùng và đề xuất các sản phẩm tốt nhất trong các kết quả đề xuất thích hợp.

Theo một nghiên cứu của McKinsey, hệ thống đề xuất giúp các nhà phát triển trên nền tảng Internet như: tăng 40% lượt cài ứng dụng trên Google Play, 60% thời gian xem trên YouTube, 35% mua hàng trên Amazon, 75% lượt xem phim trên Netflix (Ian MacKenzie, et al., 2013). Thị trường phim hoạt hình toàn cầu đang thu hút 100 triệu người trên toàn thế giới xem phim hoạt với sự yêu thích dành cho các thể loại, phong cách và nội dung đa dạng (Wise, 2023). Hệ thống đề xuất cần đưa ra các nội dung phù hợp với người dùng nhằm nâng cao trải nghiệm khi tìm kiếm.

Trong lĩnh vực Machine Learning (Học máy) có nhiều thuật toán có khả năng đưa ra đề xuất hiệu quả, tối ưu. Collaborative Filtering (Lọc cộng tác) là một kỹ thuật

cơ bản được ứng dụng trong hệ thống đề xuất, mô hình này không yêu cầu kiến thức về miền vì tính năng được nhúng tự động. Có thể hỗ trợ người dùng tìm thấy những mối quan tâm mới, học máy có thể không biết người dùng quan tâm đến một mục đích nhất định nhưng vẫn có thể đưa ra đề xuất vì những người dùng khác cũng quan tâm đến mục đó. Hệ thống đề xuất chỉ cần ma trận phản hồi để đào tạo một mô hình phân cấp ma trận và không cần các tính năng theo ngữ cảnh (Google for Developers, 2024).

Xuất phát từ các lý do trên, đề tài “Ứng dụng học máy trong xây dựng hệ thống đề xuất phim hoạt hình” đã được thực hiện.

1.2. Mục tiêu nghiên cứu

Mục tiêu chung là: Xây dựng web tích hợp hệ thống đề xuất và phim hoạt hình.

Mục tiêu cụ thể là: Bằng việc biến đổi dữ liệu thứ cấp thành thông tin bằng thuật toán lọc cộng tác nhằm đưa ra các đề xuất phim hoạt hình.

1.3. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài: hệ thống đề xuất phim hoạt hình.

Thời gian nghiên cứu của đề tài: từ tháng 03 năm 2024 đến tháng 07 năm 2024.

Chương 2: TỔNG QUAN TÀI LIỆU

2.1. Tổng quan đối tượng nghiên cứu

2.1.1. Khái niệm hệ thống đề xuất

Artificial Intelligence (Trí tuệ nhân tạo) là một nhánh của khoa học máy tính. Nó liên quan đến việc phát triển máy tính các chương trình để hoàn thành các nhiệm vụ đòi hỏi trí thông minh của con người. Các thuật toán trí tuệ nhân tạo có thể giải quyết việc học, nhận thức, giải quyết vấn đề, hiểu ngôn ngữ và hoặc lý luận logic (Saleh, 2019). Trí tuệ nhân tạo được chia là 4 loại (Park Thaichon, et al., 2022): Reactive Machines, Limited Memory, Theory of Mind, Self-Awareness. Reactive Machines dựa trên dữ liệu để đưa ra quyết định. Limited Memory sử dụng dữ liệu quá khứ để máy hiểu và đưa ra quyết định. Theory of Mind là dạng học máy, huấn luyện cho máy có cảm xúc như con người. Self-Awareness là dạng học máy, tự động thao tác, tự động ra quyết định dựa trên môi trường.

Machine Learning (Học máy) là một nhánh của trí tuệ nhân tạo và khoa học dữ liệu, tập trung sử dụng thuật toán dựa trên dữ liệu và mô tả lại cách con người có thể hiểu và đưa ra quyết định. Nó chủ yếu dựa trên kiến thức của con người về thuật toán để ra quyết định một cách hiệu quả, cung cấp cho hệ thống khả năng tự động học hỏi và cải thiện từ kinh nghiệm mà không cần lập trình rõ ràng (Alam, 2023). Học máy được chia là 3 loại cơ bản:

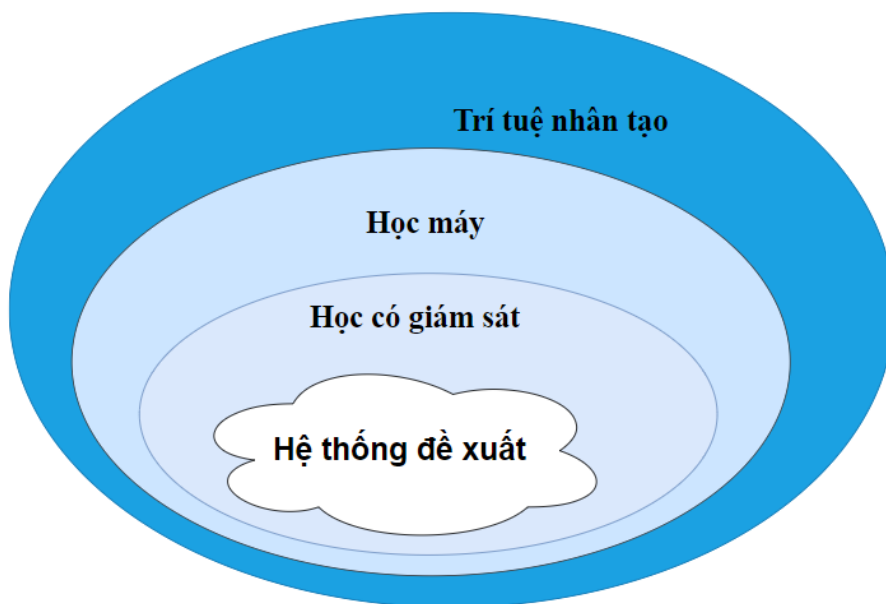
Supervised Learning (Học có giám sát) là cách học trong đó chúng ta dạy hoặc huấn luyện máy sử dụng dữ liệu được gắn nhãn rõ ràng, nghĩa là một số dữ liệu đã được gắn thẻ với các câu trả lời đúng. Sau đó, máy được cung cấp một bộ ví dụ (dữ liệu) mới để thuật toán học có giám sát phân tích dữ liệu huấn luyện (bộ ví dụ huấn luyện) và đưa ra kết quả chính xác từ dữ liệu được dán nhãn (Alam, 2023).

Unsupervised Learning (Học không có giám sát) là một kỹ thuật của học máy, trong đó người ta không cần giám sát mô hình. Thay vào đó, bạn cần cho phép mô hình tự hoạt động để khám phá thông tin. Nó chủ yếu xử lý các dữ liệu chưa được gắn nhãn. Các thuật toán học không giám sát cho phép một người thực hiện các nhiệm vụ xử lý phức tạp hơn so với học có giám sát. Mặc dù vậy, học không giám sát có thể khó dự đoán hơn so với các phương pháp học tự nhiên khác (Alam, 2023).

Reinforcement Learning (Học tăng cường) sử dụng các khái niệm về tác nhân, môi trường, trạng thái, hành động và phần thưởng (Alam, 2023).

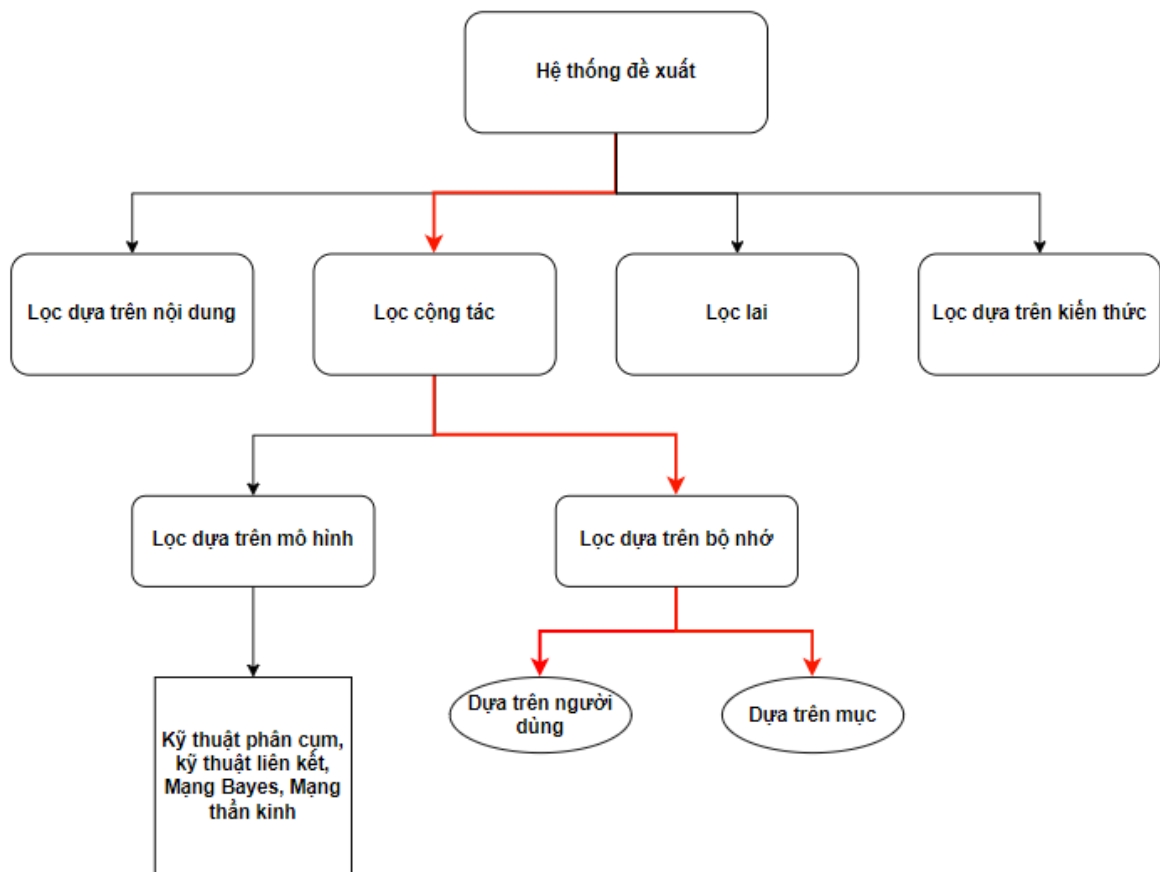
Recommendation System (Hệ thống đề xuất) là một tập hợp con của học máy sử dụng dữ liệu để giúp người dùng tìm thấy sản phẩm và nội dung. Các trang web và dịch vụ phát trực tuyến sử dụng hệ thống đề xuất để tạo các trang và nội dung “dành cho bạn” hoặc “bạn cũng có thể thích” (Shetty, 2023).

Mối quan hệ của Artificial Intelligence (Trí tuệ nhân tạo) với Recommendation System (Hệ thống đề xuất) được thể hiện như Hình 2.1.



Hình 2.1: Mối quan hệ giữa trí tuệ nhân tạo với hệ thống đề xuất

Hệ thống đề xuất là một hệ thống lọc thông tin hỗ trợ người dùng trong một tình huống ra quyết định nhất định bằng cách thu hẹp tập hợp các tùy chọn có thể và ưu tiên các yếu tố của nó trong một ngữ cảnh cụ thể. Việc sắp xếp thứ tự ưu tiên có thể dựa trên sở thích được thể hiện rõ ràng hoặc ẩn ý của người dùng và cũng dựa trên hành vi trước đó của người dùng có sở thích tương tự (Apáthy, 2019).



Hình 2.2: Các phân loại hệ thống đề xuất

(Farhin Mansur, et al., 2017), (Burke, 2000)

Hệ thống đề xuất có thể được phân loại rộng rãi thành một số danh mục tùy thuộc vào thông tin họ sử dụng để giới thiệu sản phẩm (Farhin Mansur, et al., 2017) có thể phân loại như Hình 2.2..

Content-based Filtering (Lọc dựa trên nội dung): là dựa trên nội dung hoạt động với dữ liệu mà người dùng cung cấp, rõ ràng (xếp hạng) hoặc ngầm (nhấp vào một liên kết). Dựa trên dữ liệu đó, một hồ sơ người dùng được tạo, sau đó được sử dụng để tạo đề xuất cho người dùng. Khi người dùng cung cấp nhiều hơn đầu vào hoặc xếp hạng hoặc thực hiện hành động trên đề xuất, động cơ ngày càng trở nên nhiều hơn hiệu quả và chính xác (Farhin Mansur, et al., 2017).

Collaborative Filtering (Lọc cộng tác): trong đó các đề xuất được đưa ra bởi những người khác có cùng sở thích trong quá khứ, nhưng những người đã trải nghiệm qua một mặt hàng hoặc sản phẩm chưa được người dùng hiện tại biết đến. Hệ thống lọc cộng tác yêu cầu người dùng thể hiện ý kiến về sản phẩm hoặc mặt hàng. Họ thu thập ý

kiến của người dùng và đề xuất các mặt hàng dựa trên ý kiến của mọi người giống nhau (Farhin Mansur, et al., 2017).

Model Based Filtering (Lọc dựa trên mô hình): Các hệ thống đề xuất dựa trên mô hình sử dụng một số các bộ dữ liệu nhỏ được gọi là mô hình. Mô hình này là thiết kế sử dụng trích xuất một số thông tin từ không lồ cơ sở dữ liệu liên quan đến tham số hoặc thuộc tính cụ thể và sử dụng mô hình này mọi lúc mà không cần sử dụng lớn cơ sở dữ liệu (Farhin Mansur, et al., 2017). Có thể đề cập đến như: Clustering Technique (Kỹ thuật phân cụm), Association Technique (Kỹ thuật liên kết), Bayesian Network (Mạng Bayes), Neural Network (Mạng thần kinh).

Memory Based Filtering (Lọc dựa trên bộ nhớ): là một kỹ thuật lọc truyền thống gồm User Based và Item Based. User Based (Lọc dựa trên người dùng) là tính toán sự tương đồng giữa những người dùng bằng cách so sánh xếp hạng của họ trên cùng một mặt hàng hoặc sản phẩm, sau đó nó sẽ tính toán xếp hạng dự đoán cho một mặt hàng hoặc sản phẩm theo hoạt động người dùng là xếp hạng trung bình có trọng số của mặt hàng hoặc sản phẩm bởi người dùng tương tự như người dùng đang hoạt động trong đó trọng lượng là điểm tương đồng của những người dùng này với mục tiêu mặt hàng hoặc sản phẩm (Farhin Mansur, et al., 2017). Item Based (Lọc dựa trên mục): tạo ra một mô hình tương tự về mặt hàng bằng cách truy xuất tất cả các mục được đánh giá bởi một người dùng đang hoạt động từ ma trận mục người dùng, nó xác định mức độ tương tự của các mục được truy xuất với mục tiêu, sau đó nó chọn các mục tương tự nhất và sự tương đồng phù hợp của chúng cũng được xác định (Farhin Mansur, et al., 2017).

Hybrid Filtering (Lọc lai): là sự kết hợp lọc cộng tác và lọc dựa trên nội dung, có thể được thực hiện trong một số cách thức: bằng cách lọc và cộng tác dựa trên nội dung lọc các dự đoán riêng biệt và sau đó kết hợp chúng; bằng cách thêm các chức năng dựa trên nội dung vào cộng tác kỹ thuật lọc (và ngược lại); hoặc bằng cách thống nhất kỹ thuật thành một mô hình (Farhin Mansur, et al., 2017).

Knowledge-based (Lọc dựa trên kiến thức): dựa trên phân tích tài liệu được thực hiện bằng cách sử dụng các hoạt động tìm kiếm, đánh giá và thảo luận về những đóng góp đã được xem xét (Mathias Uta, et al., 2024).

2.1.2. Khái niệm phim hoạt hình

Phim hoạt hình hay phim hoạt họa là những bức vẽ vui nhộn bình luận một cách châm biếm về các sự kiện hiện tại. Phim hoạt hình là một bộ phim ngắn được vẽ bằng tay hoặc do máy tính tạo ra, kể một câu chuyện bằng cách sử dụng một loạt hình vẽ để tạo ảo giác về chuyển động. Ban đầu, phim hoạt hình dùng để chỉ những bức vẽ có kích thước đầy đủ được sử dụng để chuyển tải các thiết kế, nhưng vào thế kỷ 19, thuật ngữ này mang ý nghĩa hiện đại là những bức vẽ hài hước hoặc châm biếm. Phim hoạt hình đa dạng từ mô tả hài hước nhẹ nhàng đến mô tả gay gắt về các sự kiện hiện tại và cuộc sống của con người (Yatti, 2024).

Trong tiếng Nhật, thuật ngữ anime được dùng để chỉ tất cả các tác phẩm hoạt hình, bất kể phong cách hay nguồn gốc. Nhiều từ điển tiếng Anh thường định nghĩa anime là “ một phong cách hoạt hình Nhật Bản “ hoặc là “một phong cách hoạt hình có nguồn gốc từ Nhật Bản” (Aeschliman, 2024).

2.2. Tổng quan phương pháp nghiên cứu

Thuật toán lọc cộng tác

Thuật toán Collaborative Filtering (Thuật toán lọc cộng tác) những hệ thống này cố gắng dự đoán “ratings” mà người dùng sẽ đưa ra cho một bộ phim hoạt hình dựa trên các chỉ số đánh giá, xếp hạng trước đây và sở thích của những người dùng khác (Debashis Das, et al., 2017).

Ma trận tương quan giữa người dùng và phim hoạt hình được xây dựng dựa trên điểm đánh giá của người dùng dành cho phim hoạt hình trên thang điểm 10 như Bảng 2.1.

Phim hoạt hình						
Người dùng	i1	i2	i3	i4	i5	i6
u1	8	10	2	5	4	2
u2	5	3	5	?	0	8
u3	10	8	?	5	10	7

u4	7	9	1	10	?	8
u5	7	?	4	8	7	3
u6	3	0	4	5	4	5

Bảng 2.1: Ma trận tương quan giữa người dùng và phim hoạt hình

Dựa vào bảng 2.1 có thể thấy người dùng số 4 và số 5 cùng đánh giá phim hoạt hình số 1 là 7 điểm thì có thể dự đoán rằng hai người dùng này có cùng sở thích và mức độ quan tâm gần giống nhau. Sự phân cụm này dựa vào những thông tin không nhất quán và xác định sự tương đồng của nhóm người dùng là điều mô hình lọc cộng tác hướng đến.

Tạo một bảng tổng hợp với các cột `user_id`, `title` và `score` để tiến hành xây dựng ma trận tương quan giữa người dùng và phim hoạt hình.

```
anime_pivot = final_rating.pivot_table(columns='user_id', index='title', values='score')
anime_pivot
```

Hình 2.1: Lệnh tạo ma trận

	user_id	36	228	384	436	448	488	527	540	592	614	...	7119604	7128939	7142335	7148055	7149397	7152703	7175775	7181454	7213650	7227104
title																						
"0"		NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
"Bungaku Shoujo" Kyou no Oyatsu: Hatsukoi		NaN	6.0	NaN	NaN	NaN	NaN	NaN	7.0	7.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
"Bungaku Shoujo" Memoire		NaN	7.0	NaN	NaN	NaN	NaN	NaN	8.0	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
"Bungaku Shoujo" Movie		NaN	7.0	NaN	8.0	NaN	NaN	0.0	8.0	0.0	NaN	...	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
"Eiji"		NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	
xxxHOLIC Movie: Manatsu no Yoru no Yume		NaN	7.0	10.0	7.0	7.0	0.0	NaN	7.0	0.0	0.0	...	NaN	0.0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
xxxHOLIC Rou		NaN	7.0	9.0	NaN	0.0	NaN	NaN	NaN	0.0	NaN	...	NaN	0.0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
xxxHOLIC Shunmuki		NaN	9.0	10.0	NaN	0.0	NaN	NaN	7.0	NaN	NaN	...	NaN	0.0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
eIDLIVE		NaN	NaN	0.0	7.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.0	NaN	4.0	0.0	NaN	NaN	NaN	NaN	7.0
○		NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

12322 rows × 8840 columns

Hình 2.2: Kết quả tạo ma trận

Thay thế các giá trị NaN thành 0, khi chuyển tập dữ liệu này sang thuật toán để mô hình hiểu và có thể tính toán khoảng cách giữa tất cả các cột.

```
anime_pivot.fillna(0, inplace=True)
anime_pivot
```

Hình 2.3: Lệnh chuyển đổi NaN thành 0

	user_id	36	228	384	436	448	488	527	540	592	614	...	7119604	7128939	7142335	7148055	7149397	7152703	7175775	7181454	7213650	7227104
	title																					
	"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	0.0	0.0	0.0
	"Bungaku Shoujo" Kyou no Oyatsu: Hatsukoi	0.0	6.0	0.0	0.0	0.0	0.0	0.0	7.0	7.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	"Bungaku Shoujo" Memoire	0.0	7.0	0.0	0.0	0.0	0.0	0.0	8.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	"Bungaku Shoujo" Movie	0.0	7.0	0.0	8.0	0.0	0.0	0.0	8.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	"Eiji"	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	0.0	0.0

	xxxHOLIC Movie: Manatsu no Yoru no Yume	0.0	7.0	10.0	7.0	7.0	0.0	0.0	7.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	xxxHOLIC Rou	0.0	7.0	9.0	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
	xxxHOLIC Shunmuki	0.0	9.0	10.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	eIDLIVE	0.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	7.0
	○	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	0.0	0.0	0.0

12322 rows × 8840 columns

Hình 2.4: Kết quả chuyển đổi

Các giá trị NaN thành 0 điều này đồng nghĩa với việc dữ liệu 0 không chứa giá trị gì cả và thuật toán sẽ cố gắng tính hết tất cả các giá trị từ cột này sang cột khác nên sẽ mất rất nhiều thời gian. Sử dụng hệ thống dữ liệu CSR, tạo một `csr_matrix` hàng thưa thớt được chỉ xem xét những giá trị có đại lượng và tự động bỏ qua các ô có giá trị bằng 0.

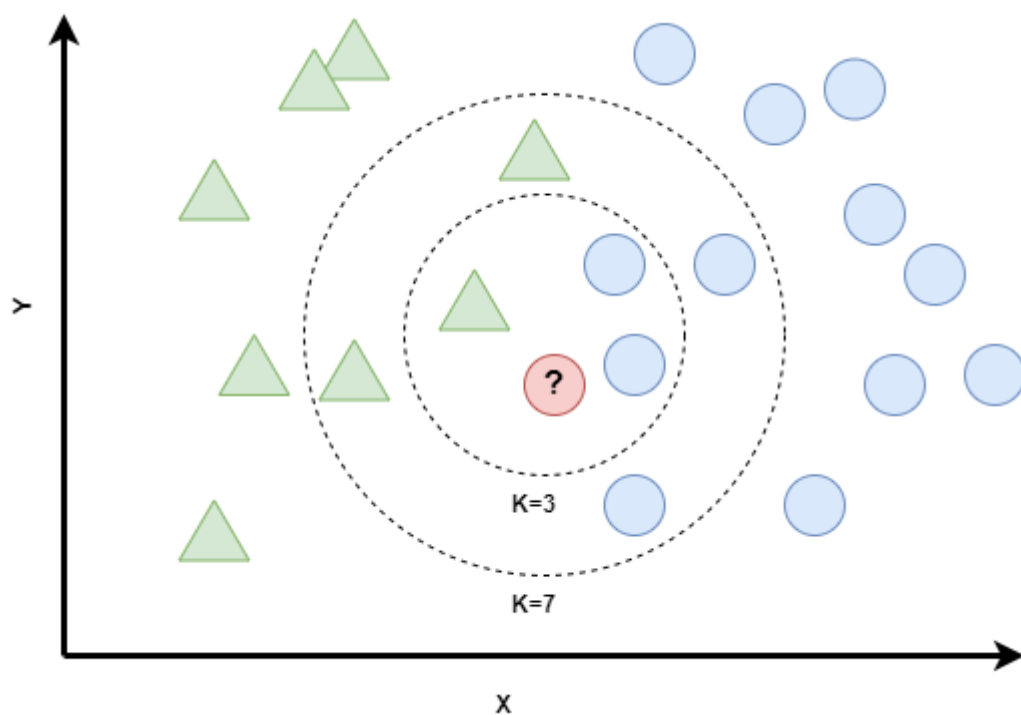
```
anime_sparse = csr_matrix(anime_pivot)
```

Hình 2.7: Lệnh tạo hàng không gian nén

Thuật toán lọc cộng tác có thể học hỏi và thích với các hành vi của người dùng giúp các đề xuất luôn phù hợp với sở thích người dùng. Thuật toán này không phụ thuộc vào thông tin mô tả phim hoạt hình mà chỉ dựa trên hành vi của người dùng tương tự, vì dựa vào hành vi thực tế của người dùng nên kết quả đề xuất có độ tin cậy cao và có thể đề xuất phim hoạt hình phù hợp với hành vi người dùng một cách chính xác và hiệu quả. Được ứng dụng trong nhiều lĩnh vực khác nhau như thương mại điện tử, phim ảnh, âm nhạc. Đóng vai trò quan trọng trong việc cá nhân hóa trải nghiệm người dùng và nâng cao hiệu quả kinh doanh. Những điều trên phù hợp với mục tiêu của đề tài nghiên cứu.

K-nearest neighbors (KNN)

Thuật toán K-nearest neighbors (KNN - Thuật toán K láng giềng gần nhất) là một kĩ thuật học có giám sát dùng để phân loại quan sát mới bằng cách tìm điểm tương đồng giữa quan sát mới này với dữ liệu sẵn có. Đó là một thuật toán Lazy Learning (Lười học), nghĩa là không có quá trình huấn luyện và được cung cấp dữ liệu thử nghiệm, nó chỉ cần dự đoán trạng thái của dữ liệu thử nghiệm dựa trên trạng thái của nó. k hàng xóm gần nhất trong dữ liệu huấn luyện. Những thách thức chính đối với người hàng xóm gần nhất k (KNN) bao gồm: k lựa chọn giá trị và lựa chọn hàng xóm, trong đó lựa chọn hàng xóm bao gồm đo khoảng cách và tìm kiếm hàng xóm (Jiaye Li, et al., 2023). Thuật toán KNN được thể hiện như Hình 2.5.



Hình 2.5: Thuật toán KNN

Cho x, y là hai trục mô phỏng giá trị, tiến hành chọn giá trị k là số lượng hàng xóm gần nhất sẽ được sử dụng để đưa ra dự đoán. Chọn k tùy theo bài toán muốn giải quyết và có thể thay đổi linh hoạt cho từng bài toán cụ thể. Tính khoảng cách giữa điểm hình tròn màu đỏ và tất cả các điểm trong tập huấn luyện (hình tam giác màu xanh lá và hình tròn màu xanh lam). Chọn k hàng xóm gần nhất dựa trên khoảng cách đã tính toán. Gán nhãn của lớp đa số cho điểm dữ liệu mới. Lặp lại các bước trên cho tất cả các điểm dữ liệu trong tập kiểm tra để xác định được k hàng xóm gần nhất là kết quả trả về.

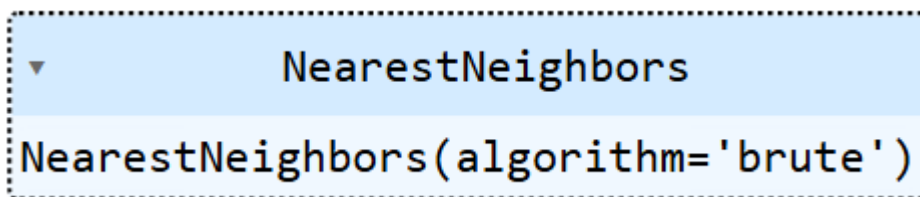
Xây dựng thuật toán KNN bằng cách tạo một biến “model” và truyền vào biến ma trận `csr_matrix`. Để dự đoán kết quả của một điểm dữ liệu mới: Tìm k điểm dữ liệu giống nhất (gần nhất) trong không gian nén để dự báo. Lấy kết quả trung bình (hồi quy) hoặc kiểu (phân loại) của k đó. Và k là tham số ảnh hưởng đến độ phức tạp của mô hình.

```
model = NearestNeighbors(algorithm='brute')
```

Hình 2.9: Lệnh gọi thuật toán

```
model.fit(anime_sparse)
```

Hình 2.10: Lệnh truyền hàng không gian nén



```
NearestNeighbors  
NearestNeighbors(algorithm='brute')
```

Hình 2.11: Kết quả lệnh gọi thuật toán KNN

Thuật toán KNN cho phép điều chỉnh hệ số k một cách dễ dàng giúp tối ưu hóa hiệu suất của thuật toán theo bài toán cụ thể. Không cần dữ liệu huấn luyện và chỉ tính toán khi có dữ liệu mới cần dự đoán. Cách hoạt động của thuật toán KNN khá đơn giản, dễ hiểu và không cần có kiến thức chuyên sâu về học máy. Việc triển khai thuật toán KNN không yêu cầu lập trình phức tạp. Nhưng hiệu quả mang lại cao trong việc phân loại và hồi quy, không yêu cầu mô hình thống kê phức tạp nên có thể thích ứng tốt với các tập dữ liệu. Thuật toán KNN được ứng dụng trong các bài toán khác nhau như phân loại hình ảnh, phân loại văn bản, xác định vị trí, đề xuất sản phẩm,... Vì vậy, thuật toán KNN hữu ích với đề tài nghiên cứu trong việc đề xuất phim hoạt hình.

2.3. Tình hình nghiên cứu

Hiện nay trên thế giới việc sử dụng học máy cụ thể là hệ thống đề xuất được thực hiện trong nhiều vấn đề: sàn thương mại điện tử (hỗ trợ bán hàng trực tuyến, đề xuất sản phẩm,...), giải trí (đề xuất phim ảnh, bài hát,...), giáo dục (đề xuất sách, các nguồn tài nguyên học tập, nghiên cứu,...),...

Đặc biệt học máy được áp dụng trong lĩnh vực giải trí cụ thể là việc đề xuất phim hoạt hình cũng có một sức hút nhất định trên thế giới. Một số nghiên cứu gần đây có thể được đề cập đến như:

Simbolon và cộng sự (2020) giới thiệu về một hệ thống đề xuất phim hoạt hình. Hệ thống sử dụng phương pháp lọc cộng tác để đề xuất phim hoạt hình cho người dùng dựa trên sở thích của họ. Hệ thống phân tích xếp hạng từ những người dùng trước đó để dự đoán phim hoạt hình mà người dùng sẽ thích. Điều này có thể hữu ích cho những người dùng cảm thấy choáng ngợp bởi số lượng phim hoạt hình không lồ có sẵn.

Athira Ramasundaran và cộng sự (2023) đề cập về một hệ thống đề xuất phim hoạt hình sử dụng thuật toán lọc cộng tác. Hệ thống này tìm kiếm phim hoạt hình để đề xuất cho người dùng dựa trên xếp hạng của những phim họ đã xem trước đó. Hệ thống sử dụng tập dữ liệu từ Kaggle bao gồm 73516 người dùng và 12294 phim hoạt hình. Lịch sử xem phim của người dùng được so sánh với lịch sử xem phim của những người dùng khác để tìm ra sở thích tương tự. Việc so sánh này được thực hiện bằng phương pháp bình phương ít luân phiên (ALS). Sau đó, hệ thống đề xuất phim hoạt hình dựa trên những kết quả so sánh này.

Jiyue Chen (2023) giới thiệu về một hệ thống đề xuất phim hoạt hình. Hệ thống giải quyết vấn đề khó khăn trong việc lựa chọn phim hoạt hình để xem với vô số lựa chọn hiện có. Hệ thống sử dụng học máy để đề xuất phim hoạt hình cho người dùng dựa trên sở thích của họ. Điều này có thể giúp người dùng khám phá những phim hoạt hình mới mà họ có thể thích. Hệ thống sử dụng nhiều kỹ thuật khác nhau, bao gồm lọc cộng tác, lọc dựa trên nội dung và phân tích khám phá dữ liệu (EDA).

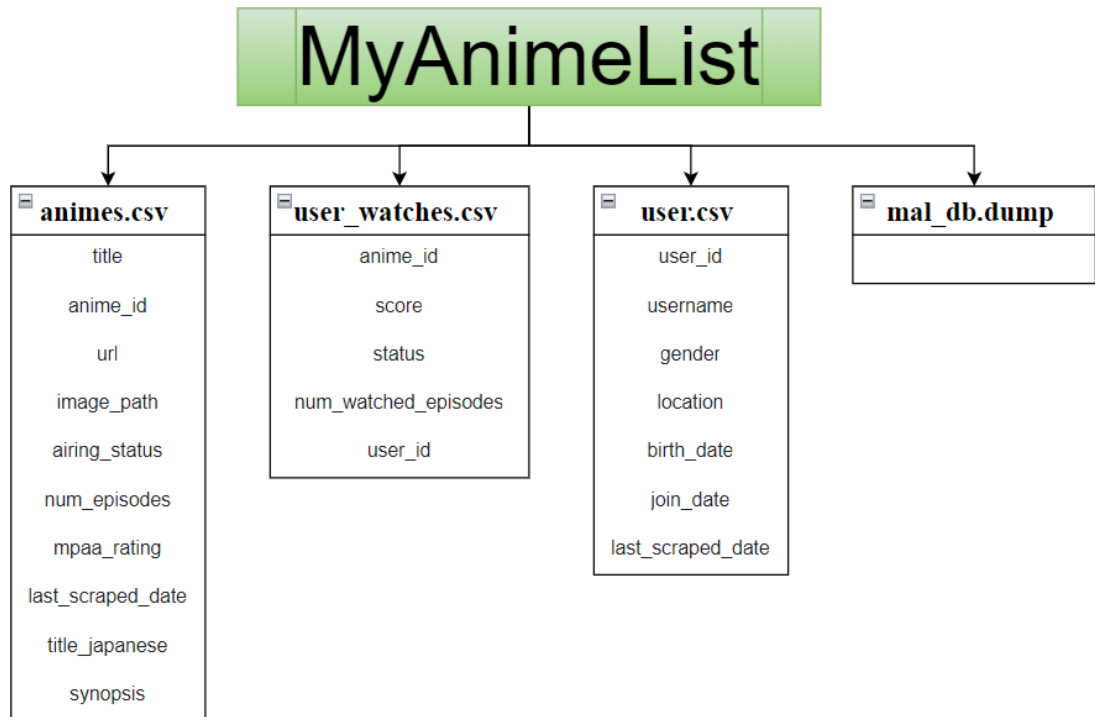
Shashidhar Reddy Javaji và Krutika Sarode (2023) đề cập về một hệ thống đề xuất phim hoạt hình mới kết hợp. Hệ thống sử dụng mạng nơ-ron đồ thị (GNN) với những biến đổi câu để đề xuất phim hoạt hình. Hệ thống xem xét cả tính năng của phim hoạt hình và tương tác của người dùng với các phim hoạt hình khác nhau. Cách tiếp cận này nắm bắt cả các tính năng liên cấp và nội cấp của dữ liệu phim hoạt hình. Mô hình cũng có thể dự đoán xếp hạng mà một người dùng cụ thể sẽ dành cho một phim hoạt hình.

Qua các nghiên cứu trên, có thể thấy hệ thống đề xuất phim hoạt hình được xây dựng thông qua nhiều mô hình khác nhau mang lại ưu điểm và nhược điểm riêng. Simbolon và cộng sự (2020) hệ thống sử dụng phương pháp lọc cộng tác đơn giản và dễ triển khai, nhưng hiệu quả có thể bị hạn chế do chỉ sử dụng xếp hạng từ người dùng. Athira Ramasundaran và cộng sự (2023) xây dựng một hệ thống tìm kiếm phim hoạt hình để đề xuất cho người dùng dựa trên xếp hạng của những phim họ đã xem trước đó. Nhưng hệ thống này có thể gặp khó khăn khi đối mặt với dữ liệu thưa thớt. Jiyue Chen (2023) hệ thống sử dụng nhiều kỹ thuật khác nhau cho phép đề xuất phim hoạt hình đa dạng hơn, nhưng có thể tốn nhiều thời gian và tài nguyên tính toán để triển khai. Shashidhar Reddy Javaji và Krutika Sarode (2023) sử dụng mạng nơ-ron đồ thị là hệ thống tiên tiến nhất, có thể đề xuất phim hoạt hình chính xác và dự đoán xếp hạng của người dùng, nhưng đòi hỏi kiến thức chuyên môn về học máy và mạng nơ-ron đồ thị. Vì vậy, việc lựa chọn hệ thống đề xuất phim hoạt hình phù hợp là một quyết định quan trọng. Các yếu tố như nhu cầu của người dùng, đặc điểm của dữ liệu phim hoạt hình và khả năng triển khai, cần được xem xét trước khi lựa chọn hệ thống phù hợp.

Chương 3: DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Dữ liệu nghiên cứu

Bộ dữ liệu thứ cấp MyAnimeList của Kaggle được tải lên năm 2020 bởi QUANG-VINH DO được thể hiện như Hình 3.1:



Hình 3.1: Dữ liệu thứ cấp MyAnimeList

Kaggle là một cộng đồng khoa học dữ liệu với các công cụ và tài nguyên mạnh mẽ, phần lớn các dữ liệu ở đây được chia sẻ miễn phí. Tập dữ liệu với dung lượng khoảng 943 MB chứa tập hợp các anime, người dùng và xếp hạng được lấy từ myanimelist.net. Tập dữ liệu này chứa 4 tệp.

animes.csv: gồm 17058 phim hoạt hình chứa thông tin phim hoạt hình gồm tiêu đề, anime_id, trạng thái phát sóng, số tập và tóm tắt.

user.csv: gồm 302674 người dùng chứa thông tin người dùng như tên người dùng, giới tính, vị trí, ngày sinh và ngày tham gia.

user_watches.csv: gồm 68235827 xếp hạng danh sách anime của người dùng cùng với điểm số và trạng thái xem.

mal_db.dump - Tập kết xuất của cơ sở dữ liệu postgresql chứa tất cả 3 thông tin csv ở trên cùng với Primary Key/ Foreign Key thích hợp, các ràng buộc và chỉ mục khác.

Thống kê một số chỉ số cơ bản cho các cột có giá trị là dạng số.

Thống kê tệp `animes.csv` được thể hiện như Hình 3.2. Cột `anime_id` có 17058 hàng (phim hoạt hình được gán số mặc định). Cột `num_episodes` có số tập phim hoạt hình nhỏ nhất là 1 tập và số tập phim hoạt hình lớn nhất là 3057 tập, trung bình các phim hoạt hình có 12 tập và độ lệch chuẩn là 48 tập, bách phân vị thứ nhất là 1 tập, bách phân vị thứ hai là 2 tập và bách phân vị thứ ba là 12 tập.

Out[6]:

	anime_id	airing_status	num_episodes
count	17058.00	17058.0	16583.00
mean	20799.79	2.0	11.57
std	14576.79	0.2	47.92
min	1.00	1.0	1.00
25%	5664.25	2.0	1.00
50%	21772.00	2.0	2.00
75%	35058.75	2.0	12.00
max	42754.00	3.0	3057.00

Hình 3.2: Thống kê tệp `animes.csv`

Thống kê tệp `user.csv` được thể hiện như Hình 3.3. Cột `user_id` có 302674 tài khoản người dùng (được gán số mặc định).

Out[2]:

	user_id
count	302674.00
mean	2081417.54
std	2118054.64
min	1.00
25%	232244.75
50%	1187588.00
75%	3883448.50
max	7273308.00

Hình 3.3: Thống kê tệp `user.csv`

Thống kê tệp `user_watches.csv` được thể hiện như Hình 3.4. Cột `anime_id` có 68235827 hàng. Cột `score` có giá trị thấp nhất là 0 điểm và có giá trị lớn nhất là 10 điểm, giá trị trung bình là 4,42 điểm và độ lệch chuẩn là 3,94 điểm. Có bách phân vị thứ nhất là 0 điểm, bách phân vị thứ hai là 6 điểm và bách phân vị thứ ba là 8 điểm. Cột `num_watched_episodes` số lượng tập đã xem có giá trị nhỏ nhất là 0 tập và có giá trị lớn

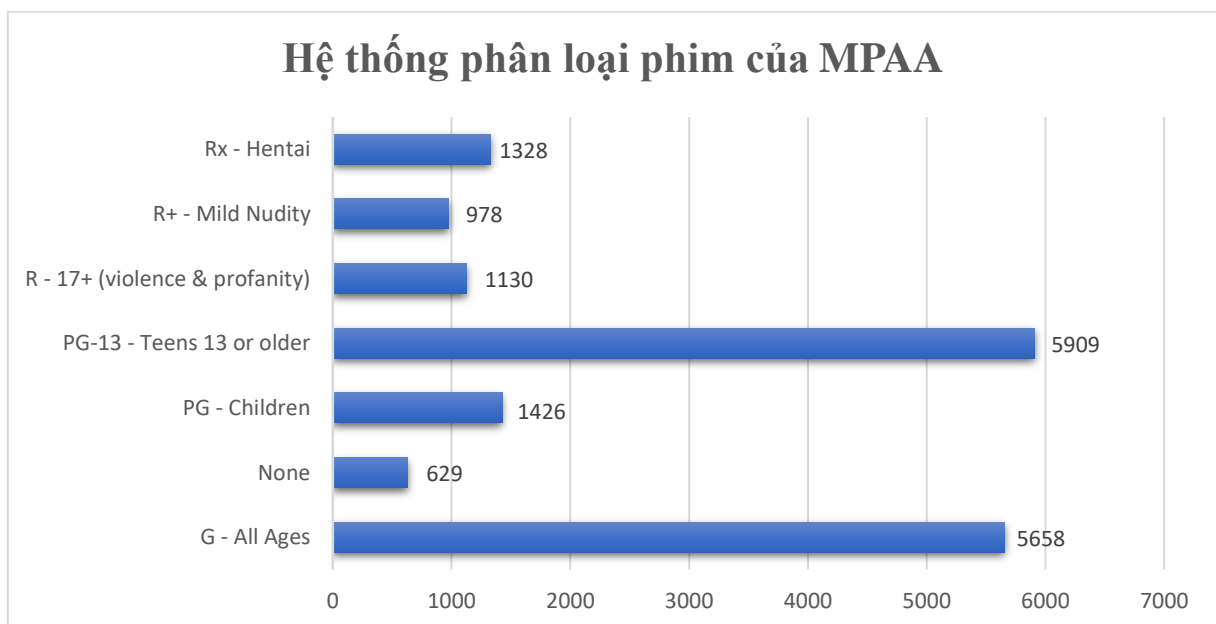
nhất là 65535 tập, giá trị trung bình là 14,53 tập và độ lệch chuẩn là 395,81 tập. Có bách phân vị thứ nhất là 0 tập, bách phân vị thứ hai là 4 tập và bách phân vị thứ ba là 13 tập.

Out[12]:

	anime_id	score	status	num_watched_episodes	user_id
count	68235827.00	68235827.00	68235827.00	68235827.00	68235827.00
mean	13378.03	4.42	2.98	14.53	2199570.02
std	12595.30	3.94	1.72	359.81	2077056.68
min	1.00	0.00	0.00	0.00	1.00
25%	1954.00	0.00	2.00	0.00	288885.00
50%	9513.00	6.00	2.00	4.00	1407861.00
75%	23251.00	8.00	4.00	13.00	4044265.00
max	42754.00	10.00	55.00	65535.00	7273306.00

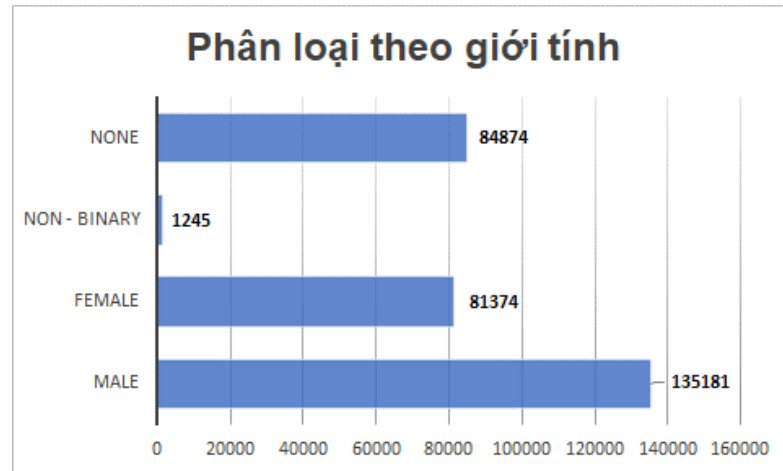
Hình 3.4: Thống kê tệp user_watches.csv

Thống kê số lượng phim hoạt hình được đánh nhãn phân loại theo độ tuổi dựa vào hệ thống phân loại phim MPAA được thể hiện như Hình 3.5. Qua đó, kết quả phân loại dựa vào bộ dữ liệu animes.csv có 17058 anime_id cho thấy có 5909 nhãn PG-13 – Teens 13 or older chiếm số lượng lớn nhất trong bộ dữ liệu. Tiếp theo, có 5658 nhãn G – All Ages, theo sau là 1426 nhãn PG – Children, sau đó là 1328 nhãn Rx – Hentai, tiếp đến là 1130 nhãn R – 17+ (violence & profanity) và 978 nhãn R+ – Mild Nudity. Cuối cùng là có 629 anime_id không được đánh nhãn chiếm một phần nhỏ không đáng kể trong bộ dữ liệu.



Hình 3.5: Hệ thống phân loại phim MPAA

-Thống kê phân loại giới tính của Users được thể hiện như Hình 3.6. Trong 302674 user_id của bộ dữ liệu user.csv có 135181 nhãn Male được đánh chiếm số lượng lớn nhất trong bộ dữ liệu, kế tiếp có 81374 nhãn Female và 1245 nhãn Non – Binary chiếm số lượng nhỏ nhất. Bên cạnh đó còn có 84874 user_id không được đánh nhãn chiếm một phần lớn trong bộ dữ liệu.

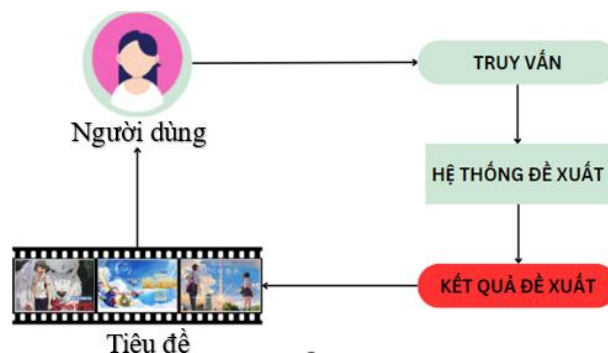


Hình 3.6: Phân loại theo giới tính của Users

3.2. Phương pháp nghiên cứu

3.2.1. Cách tiếp cận

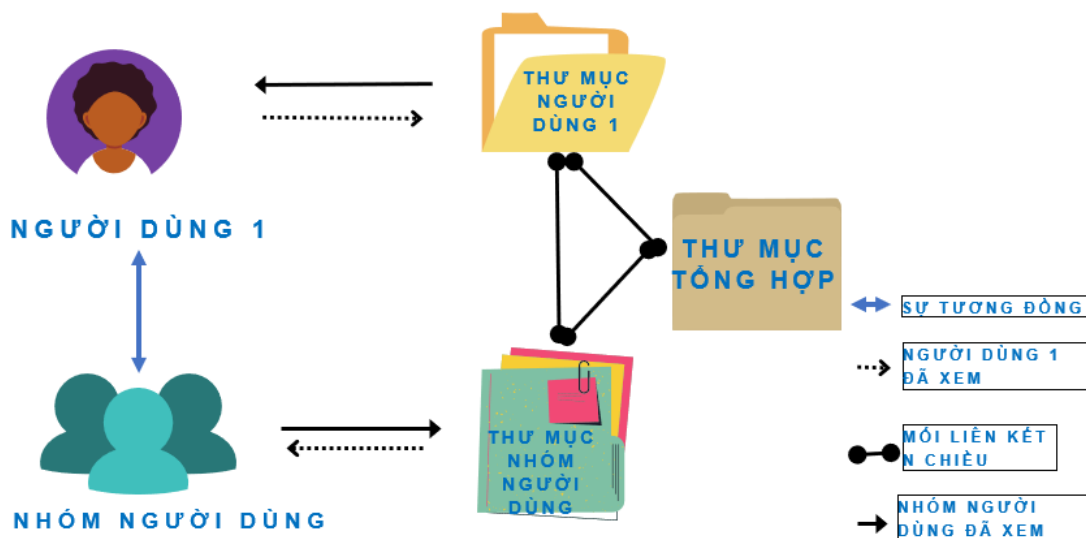
Quy trình truy vấn và kết quả đề xuất được trả về của đề tài được thể hiện như Hình 3.7. Người dùng thực hiện truy vấn tên một bộ phim hoạt hình bất kỳ thông qua hệ thống đề xuất sẽ trả về các kết quả đề xuất là tên và hình ảnh phim hoạt hình có sự tương đồng với nhau về mặt khoảng cách được tính toán bằng mô hình đề xuất.



Hình 3.7: Quy trình truy vấn và kết quả đề xuất được trả về

Tiếp cận dựa trên Collaborative Filtering- CF (lọc cộng tác): trả lời câu hỏi có ai có sở thích tương tự mình không? Khi người dùng đánh giá phim hoạt hình thì cơ sở dữ liệu sẽ lưu trữ lại và tương tự với nhóm người dùng được thể hiện như Hình 3.8. Dựa

vào hành vi của người dùng hệ thống sẽ phân cụm những người dùng có hành vi tương đồng.



Hình 3.8: Hướng tiếp cận của thuật toán lọc cộng tác

3.2.2. Phương pháp nghiên cứu

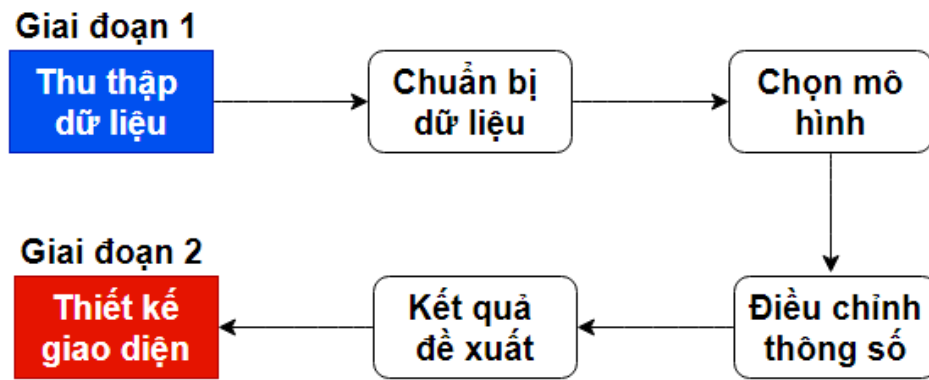
Việc thực hiện đề tài được chia thành 2 giai đoạn cụ thể được thể hiện như Hình 3.9.

Giai đoạn 1: Xây dựng mô hình đề xuất phim hoạt hình

Trong giai đoạn 1 lựa chọn bộ dữ liệu phù hợp với nghiên cứu, tuyển chọn các đặc trưng cơ bản từ bộ dữ liệu thứ cấp. Tiến hành phân tích các thông tin của bộ dữ liệu và thể hiện trực quan hóa bằng các dạng bản đồ phân tích trực quan. Chọn mô hình, điều chỉnh thông số và đưa ra đề xuất phim hoạt hình.

Giai đoạn 2: Xây dựng ứng dụng web đề xuất phim hoạt hình

Tiến hành phân tích xây dựng chức năng đề xuất phim hoạt hình và thiết kế giao diện đơn giản để trình bày kết quả đề xuất.



Hình 3.9: Sơ đồ phương pháp nghiên cứu

3.2.3. Phần mềm và thư viện

Đề tài được thực hiện trên các phần mềm sau:

Anaconda Navigator là giao diện người dùng đồ họa (GUI) trên máy tính để bàn, là một bản phân phối mã nguồn mở phổ biến của ngôn ngữ lập trình Python và trình quản lý môi trường, đơn giản hóa việc quản lý và triển khai gói (Anaconda, Inc., 2024). Sử dụng Anaconda Navigator để thiết lập các thư viện Python.

Phần mềm Python 3.12 là một phiên bản ổn định được phát hành vào tháng 10 năm 2023, thỏa thuận cấp phép bởi Python Software Foundation (PSF) và Individual or Organization ("Licensee"). Python là một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng, do Guido van Rossum tạo ra và lần đầu ra mắt vào năm 1991. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Các thư viện Python được sử dụng trong đề tài gồm:

Thư viện Numpy: là gói cơ bản dành cho tính toán khoa học với bằng Python (Travis E. Oliphant, et al., 2024).

Thư viện Pandas: là gói Python cung cấp các cấu trúc dữ liệu nhanh, linh hoạt và biểu cảm được thiết kế để giúp làm việc với dữ liệu "quan hệ" hoặc "được gắn nhãn" vừa dễ dàng vừa trực quan. Nhằm mục đích trở thành khối xây dựng cấp cao cơ bản để thực hiện phân tích dữ liệu trong thế giới thực, thực tế bằng Python (The Pandas Development Team, 2024).

Thư viện Streamlit: cho phép bạn chuyển đổi tập lệnh Python thành ứng dụng web tương tác. Xây dựng trang tổng quan, tạo báo cáo hoặc tạo ứng dụng trò chuyện.

Sau khi tạo ứng dụng, có thể sử dụng nền tảng đám mây cộng đồng của nhà phát triển để triển khai, quản lý và chia sẻ ứng dụng của cá nhân (Snowflake Inc, 2024).

Thư viện Scikit-learn (Sklearn): là một mô-đun Python dành cho máy học được xây dựng dựa trên SciPy và được phân phối theo giấy phép BSD 3 khoản. Dự án được David Cournapeau bắt đầu vào năm 2007 như một dự án Google Summer of Code và kể từ đó, nhiều tình nguyện viên đã đóng góp (Scikit-learn developers, 2024).

Thư viện Pickle: là mô-đun triển khai các giao thức nhị phân để tuần tự hóa và hủy tuần tự hóa cấu trúc đối tượng Python. Là quá trình trong đó hệ thống phân cấp đối tượng Python được chuyển đổi thành luồng byte và “giải nén” là hoạt động nghịch đảo, theo đó luồng byte (từ tệp nhị phân hoặc đối tượng giống byte) được chuyển đổi trở lại thành hệ thống phân cấp đối tượng (Sphinx 7.3.7., 2024).

Thư viện SciPy: là một phần mềm mã nguồn mở dành cho toán học, khoa học và kỹ thuật. Bao gồm các mô-đun để thống kê, tối ưu hóa, tích hợp, đại số tuyến tính, biến đổi Fourier, xử lý tín hiệu và hình ảnh, bộ giải ODE,... (SciPy Developers, 2024).

Phần mềm Jupyter Notebook là một ứng dụng soạn thảo sổ ghi chép đơn giản hóa và là một phần của Project Jupyter, một dự án lớn tập trung vào mục tiêu cung cấp các công cụ (và tiêu chuẩn) cho điện toán tương tác với sổ ghi chép tính toán... (Jupyter, 2024). Tiến hành huấn luyện mô hình đề xuất trên Jupyter Notebook.

Phần mềm Visual Studio Code là trình chỉnh sửa mã nguồn nhẹ nhưng mạnh mẽ chạy trên máy tính để bàn của bạn và có sẵn cho Windows, macOS và Linux. Sự hỗ trợ tích hợp cho JavaScript, TypeScript và Node.js, đồng thời có hệ sinh thái tiện ích mở rộng phong phú cho các ngôn ngữ và thời gian chạy khác (chẳng hạn như C++, C#, Java, Python, PHP, Go, .NET) (Visual Studio Code , 2024). Thực hiện việc chạy ứng dụng web để xuất phim hoạt hình.

Phần mềm Microsoft Excel 2019 giúp ghi lại, trình bày các thông tin với dạng bảng và nhiều tiện ích khác. Excel 2019 cho Windows nâng cấp từ một phiên bản Excel trước đó (Microsoft, 2018). Thực hiện các thao tác thống kê cơ bản với dữ liệu gốc.

Chương 4: KẾT QUẢ VÀ THẢO LUẬN

4.1. Mô hình đề xuất phim hoạt hình

4.1.1. Thu thập dữ liệu

Kết quả đọc 10 hàng đầu tiên của tệp `animes.csv` được thể hiện như Hình 4.1. Tệp `animes.csv` có 17058 hàng và 10 cột (`title` (tiêu đề), `anime_id`, `url`, `image_path` (đường dẫn hình ảnh), `airing_status` (trạng thái phát sóng), `num_episodes` (số lượng tập), `mpaa_rating`, `last_scraped_date`, `title_japanese` (tiêu đề bằng tiếng nhật), `synopsis` (tóm tắt)).

Out[2]:	title	anime_id	url	image_path	airing_status	num_episodes	mpaa_rating	last_scraped_date	title_japanese	synopsis	title_english
0	Hanasaku Iroha	9289	https://myanimelist.net/anime/9289/Hanasaku_Iroha	https://cdn.myanimelist.net/images/anime/3/289...		2	PG-13 - Teens 13 or older	2020-09-13 11:23:05.509374	花咲くいろは	Ohana Matsumae is an energetic and wild teenag...	Hanasaku Iroha: Blossoms for Tomorrow
1	Uchuu Kyoudai	12431	https://myanimelist.net/anime/12431/Uchuu_Kyoudai	https://cdn.myanimelist.net/images/anime/7/375...		2	PG-13 - Teens 13 or older	2020-09-13 12:58:18.633841	宇宙兄弟	On a fateful summer night in 2006, Mutta Namba...	Space Brothers
2	Yozakura Quartet: Hana no Uta	18497	https://myanimelist.net/anime/18497/Yozakura_Q...	https://cdn.myanimelist.net/images/anime/5/525...		2	PG-13 - Teens 13 or older	2020-09-13 14:39:45.15066	夜桜四重奏 〜ハナノウタ〜	Hundreds of years ago, the borders between the...	NaN
3	Glass no Kamen Desu ga to Z Specials	34336	https://myanimelist.net/anime/34336/Glass_no_K...	https://cdn.myanimelist.net/images/anime/13/82...		2	PG-13 - Teens 13 or older	2020-09-14 01:53:48.51646	ガラスの仮面ですがZ	Two unsaid episodes of Glass no Kamen Desu ga...	NaN
4	Re:Zero kara Hajimeru Isekai Seikatsu 2nd Season	39587	https://myanimelist.net/anime/39587/Re_Zero_k...	https://cdn.myanimelist.net/images/anime/1444/...		1	R - 17+ (violence & profanity)	2020-09-14 06:45:31.052373	Re:ゼロから始める異世界生活 2nd season	Even after dying countless times, Subaru final...	Re:ZERO -Starting Life in Another World- Season 2
5	Sword Art Online: Alicization - War of Underworld...	40540	https://myanimelist.net/anime/40540/Sword_Art_...	https://cdn.myanimelist.net/images/anime/1438/...		1	PG-13 - Teens 13 or older	2020-09-14 07:30:27.945843	ソードアート・オンライン アリシゼーション War of Underworld	Second part of Sword Art Online: Alicization ~...	Sword Art Online Alicization War of Underworld...
6	Tanu no Magical Holiday	25545	https://myanimelist.net/anime/25545/Tanu_no_Ma...	https://cdn.myanimelist.net/images/anime/13/65...		2	G - All Ages	2020-09-13 16:44:33.037659	たぬのマジカルホリデー		NaN
7	KimiKiss Pure Rouge	2927	https://myanimelist.net/anime/2927/KimiKiss_Pu...	https://cdn.myanimelist.net/images/anime/3/199...		2	PG-13 - Teens 13 or older	2020-09-13 06:48:05.084739	キミ Kiss pure rouge	Kouchi Sanada and Kazuki Aihara's childhood f...	KimiKiss: Pure Rouge
8	Tokyo Magnitude 8.0	6211	https://myanimelist.net/anime/6211/Tokyo_Magni...	https://cdn.myanimelist.net/images/anime/10/13...		2	PG-13 - Teens 13 or older	2020-09-13 09:35:06.970559	東京マグニチュード8.0	Middle school student Mirai Onozawa is disast...	Tokyo Magnitude 8.0
9	Jankenman	7405	https://myanimelist.net/anime/7405/Jankenman	https://cdn.myanimelist.net/images/anime/2/178...		2	PG - Children	2020-09-13 10:19:02.552156	ジャンケンマン	There is a small village called Janken Town. T...	NaN

Hình 4.1: Kết quả đọc tệp `animes.csv`

Kết quả đọc 10 hàng đầu tiên của tệp `user.csv` được thể hiện như Hình 4.2. Tệp `user.csv` có 302674 hàng và 7 cột (`user_id`, `username` (tên người dùng), `gender` (giới tính), `location`(vị trí), `birth_date` (ngày sinh), `join_date` (ngày đăng ký), `last_scraped_date`).

Out[8]:	user_id	username	gender	location	birth_date	join_date	last_scraped_date
0	1	Xinil	Male	California	1985-03-04 00:00:00	2004-11-05 00:00:00	2020-09-02 15:20:23.893907
1	3	Aokaado	Male	Oslo, Norway	1988-11-11 00:00:00	2004-11-11 00:00:00	2020-09-02 15:20:23.893907
2	4	Crystal	Female	Melbourne, Australia	1989-01-10 00:00:00	2004-11-13 00:00:00	2020-09-02 15:20:23.893907
3	20	vondur	Male	Bergen, Norway	1988-01-25 00:00:00	2005-01-05 00:00:00	2020-09-02 15:20:23.893907
4	23	Amuro	NaN	Canada	1988-02-22 00:00:00	2005-01-23 00:00:00	2020-09-02 15:20:23.893907
5	37	megan	Female	San Diego, California	1987-06-18 00:00:00	2005-02-12 00:00:00	2020-09-02 15:20:23.893907
6	44	Koreth	Male	NaN	NaN	2005-02-21 00:00:00	2020-09-02 15:20:23.893907
7	47	kei-clone	Male	New York, NY	1988-01-01 00:00:00	2005-03-09 00:00:00	2020-09-02 15:20:23.893907
8	53	Ladholymen	NaN	NaN	NaN	2005-03-19 00:00:00	2020-09-02 15:20:23.893907
9	66	Hiromi	Male	Calgary, AB	1990-02-09 00:00:00	2005-04-01 00:00:00	2020-09-02 15:20:23.893907

Hình 4.2: Kết quả đọc tệp `user.csv`

Kết quả đọc 10 hàng đầu tiên của tệp `user_watches.csv` được thể hiện như Hình 4.3. Tệp `user.csv` có 68235827 hàng và 5 cột (`anime_id`, `score` (điểm), `status` (trạng thái), `num_watched_episodes`(số tập đã xem), `user_id`).

Out[10]:

	anime_id	score	status	num_watched_episodes	user_id
0	21	9	1	626	1
1	11061	9	2	30	1
2	8676	0	1	4	3
3	9919	0	1	2	3
4	5081	0	1	8	3
5	109	0	1	1	3
6	2986	6	1	14	3
7	269	9	1	201	3
8	31043	0	1	3	3
9	28223	0	1	7	3

Hình 4.3: Kết quả đọc tệp `user_watches.csv`

4.1.2. Chuẩn bị dữ liệu

Xác định giá trị duy nhất và trung bình được làm tròn đến chữ số thập phân thứ 2 của các cột `anime_id` và `user_id` trong tệp `user_watches.csv` được thể hiện như Hình 4.4.

```
Number of ratings: 68235827
Number of unique anime_id's: 17058
Number of unique users: 259137
Average ratings per user: 263.32
Average ratings per anime: 4000.22
```

Hình 4.4: Giá trị duy nhất và trung bình của `anime_id` và `user_id` trong tệp `user_watches.csv`

Số lượng xếp hạng: 68235827, số lượng id phim duy nhất: 17058, số lượng người dùng duy nhất: 259137, xếp hạng trung bình trên mỗi người dùng: 263,32, xếp hạng trung bình mỗi phim: 4000,22.

4.1.3. Chọn mô hình

Hệ thống đề xuất đề tài nghiên cứu sử dụng mô hình lọc cộng tác và thuật toán KNN được đề cập ở mục 2.2.

KNN là một thuật toán Lazy Learning (Lười học), nghĩa là không có quá trình huấn luyện và được cung cấp dữ liệu thử nghiệm, nó chỉ cần dự đoán trạng thái của dữ

liệu thử nghiệm dựa trên trạng thái của nó (Jiaye Li, et al., 2023). Thuật toán KNN cho phép điều chỉnh hệ số k một cách dễ dàng giúp tối ưu hóa hiệu suất của thuật toán theo bài toán cụ thể. Không cần dữ liệu huấn luyện và chỉ tính toán khi có dữ liệu mới cần dự đoán.

4.1.4. Điều chỉnh thông số

Khởi tạo hàm `recommend_anime` với tham số là biến `anime_name` (tương ứng với giá trị cột `title`). Từ khóa `anime_name` được tìm kiếm thông qua hàm `recommend_anime` có chức năng truy vấn vào ma trận hàng thừa nên `anime_pivot`, sau đó mảng một chiều ánh xạ kết quả đề xuất. Chọn k cố định là 11 tương ứng trả về 10 kết quả phim hoạt hình. Khởi tạo hàm được thể hiện như Hình 4.5

```
In [74]: def recommend_anime(anime_name):
         anime_id = np.where(anime_pivot.index == anime_name)[0][0]
         distance, suggestion = model.kneighbors(anime_pivot.iloc[anime_id,:].values.reshape(1,-1), n_neighbors=11 )

         for i in range(len(suggestion)):
             anims = anime_pivot.index[suggestion[i]]
             for j in anims:
                 if j == anime_name:
                     print(f"You searched '{anime_name}'\n")
                     print("The suggestion anims are: \n")
                 else:
                     print(j)
```

Hình 4.5: Khởi tạo hàm đề xuất

4.1.5. Kết quả đề xuất

Với k=11, chọn tên một bộ phim hoạt hình của cột `title` để kiểm tra khả năng đề xuất của phương thức `recommmed_anime` được tạo ra trước đó. Và kết quả trả về là 10 kết quả đề xuất cho tên một bộ phim hoạt hình đã được chọn. Kết quả trả về cho từ khóa tìm kiếm là “Minami-ke” được thể hiện như Hình 4.6.

```
In [79]: recommend_anime("Minami-ke")

You searched 'Minami-ke'

The suggestion anims are:

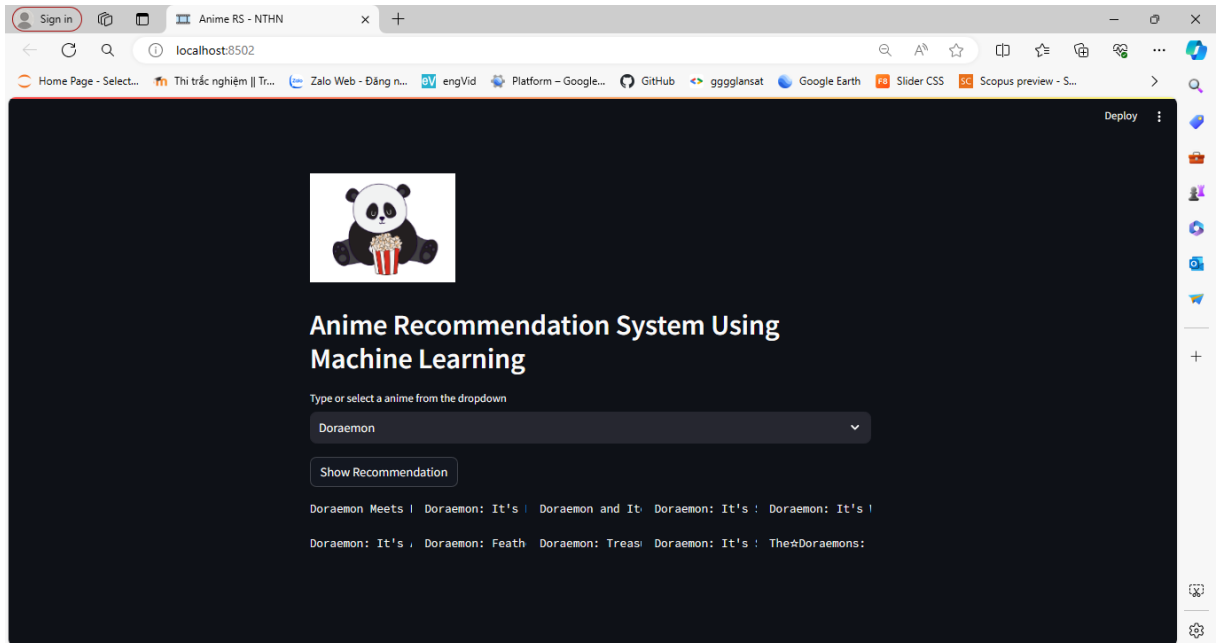
Minami-ke Okawari
Minami-ke Okaeri
Minami-ke Tadaima
Minami-ke Betsubara
Minami-ke Omatase
Minami-ke Natsuyasumi
Kannagi
Hyakko
Ichigo Mashimaro
Mitsudomoe
```

Hình 4.6: Kết quả trả về cho từ khóa tìm kiếm là “Minami-ke”

4.2. Ứng dụng web đề xuất phim hoạt hình

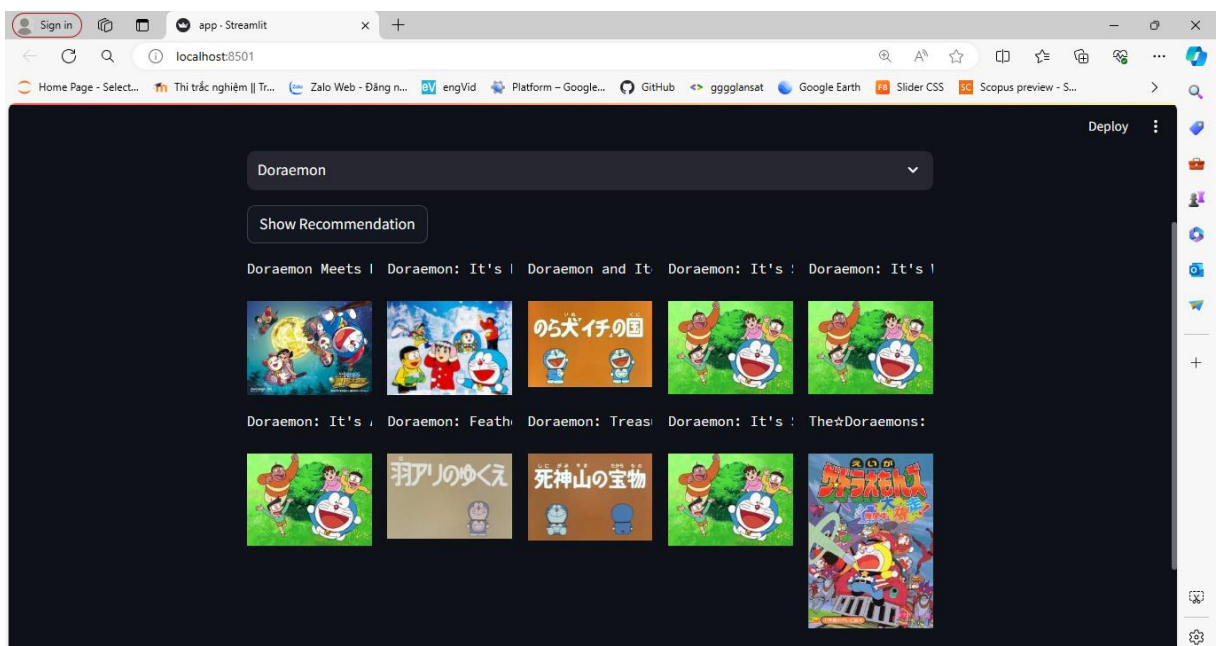
4.2.1. Thiết kế chức năng

Thực hiện truy vấn bằng tên một phim hoạt hình và kết quả đề xuất tên 10 phim hoạt hình được thể hiện như Hình 4.7.



Hình 4.7: Đề xuất tên phim hoạt hình

Kết quả truy vấn trả về là đề xuất tên phim hoạt hình và bổ sung thêm phần hình ảnh của phim hoạt hình được thể hiện như Hình 4.8.



Hình 4.8: Đề xuất hình ảnh của phim hoạt hình

4.2.2. Thiết kế giao diện

Định cấu hình cài đặt mặc định của trang gồm tiêu đề trang và biểu tượng trang được thể hiện như Hình 4.9.



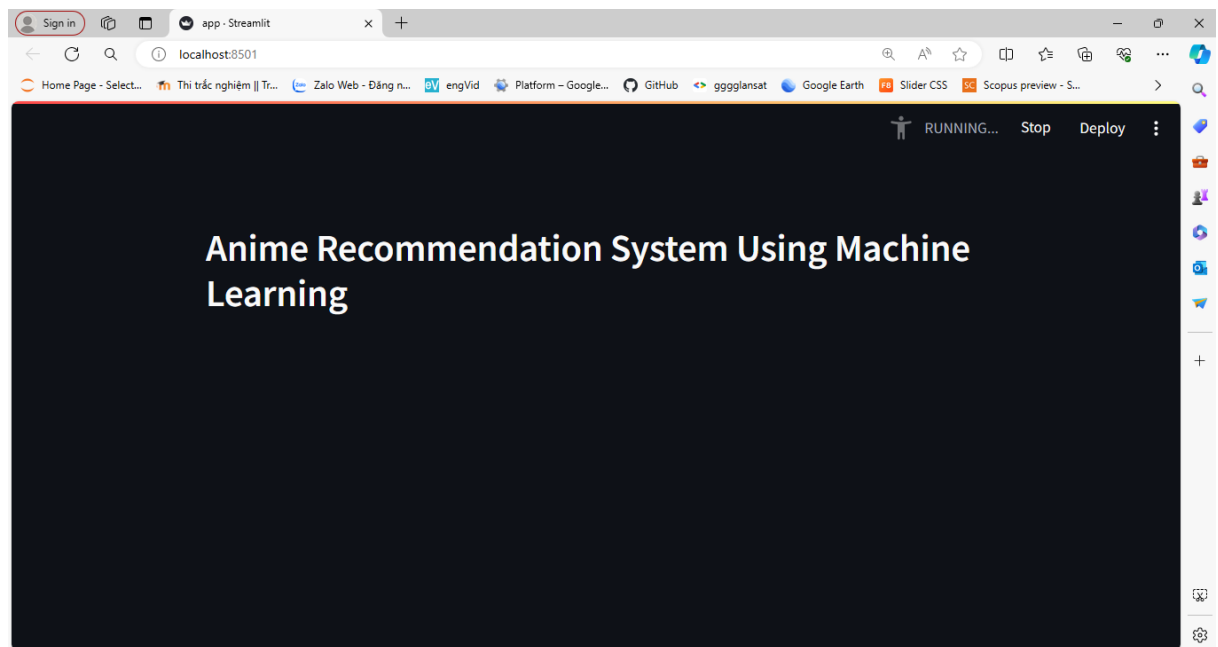
Hình 4.9: Định cấu hình cài đặt mặc định của trang

Chèn hình ảnh được thể hiện như Hình 4.10.



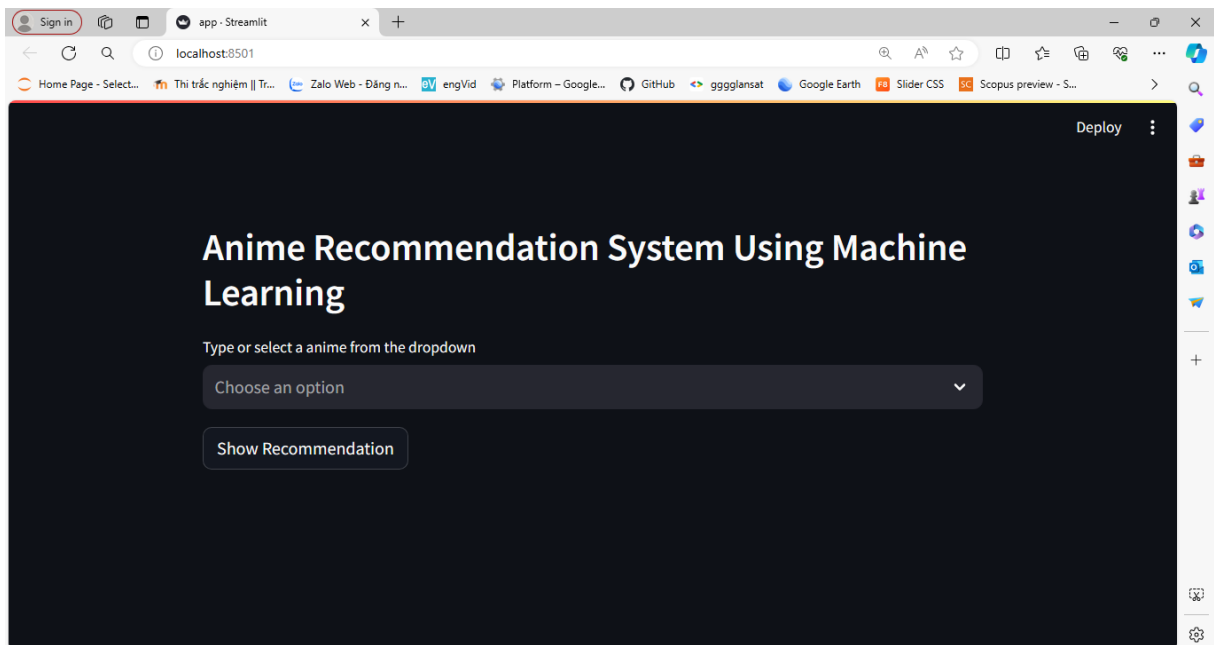
Hình 4.10: Chèn hình ảnh

Tiêu đề của web-app được thể hiện như Hình 4.11.



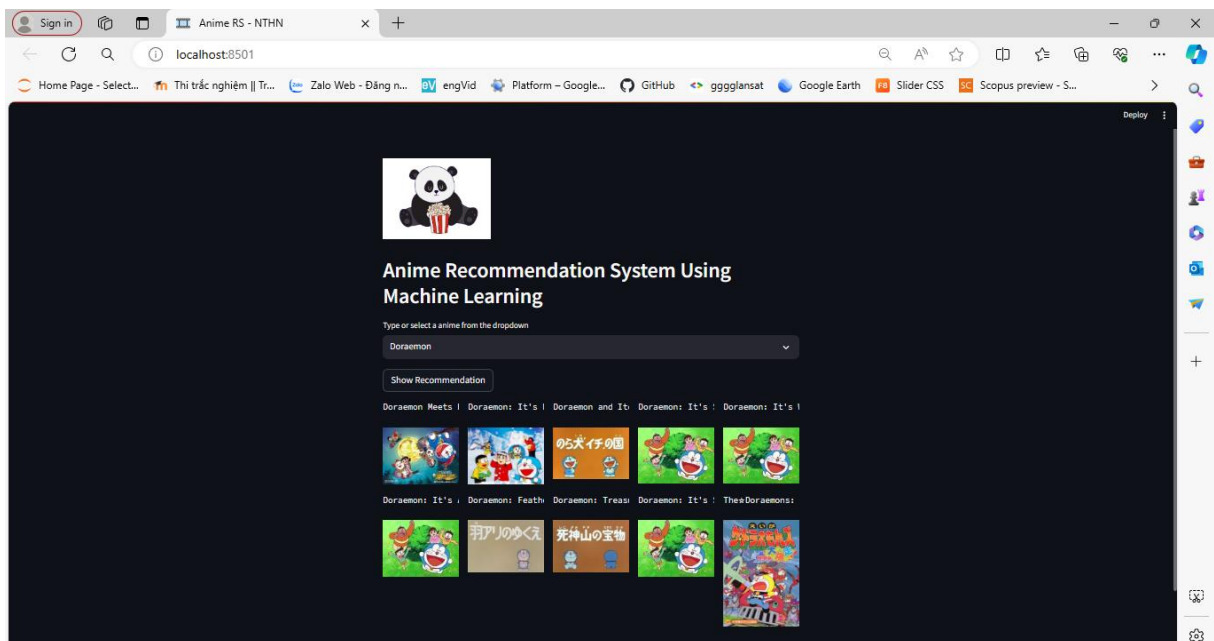
Hình 4.11: Tiêu đề của web-app

Thiết lập thanh tìm kiếm, tại đây có thể bằng cách nhập tên phim hoạt hình hoặc chọn phim hoạt hình trong danh sách kéo xuống được thể hiện như Hình 4.12. Sau đó nhấn nút “Enter” trên bàn phím máy tính hoặc nhấp đúp chuột vào nút “Show Recommendation” để hiển thị kết quả đề xuất phim hoạt hình.



Hình 4.12: Thanh tìm kiếm

Giao diện web-app tích hợp hệ thống đề xuất thông tin phim hoạt hình sử dụng học máy được thể hiện như Hình 4.13.



Hình 4.13: Giao diện web-app đề xuất phim hoạt hình

4.3. Thảo luận

Nghiên cứu sử dụng dữ liệu của myanimelist.net mang tính cập nhật đến năm 2020. Mở rộng so với các nghiên cứu liên quan thông qua việc triển khai xây dựng ứng dụng web đề xuất phim hoạt hình với chức năng đề xuất phim hoạt hình dựa trên thuật

toán lọc cộng tác và thuật toán KNN. Kết quả nghiên cứu hỗ trợ việc ra quyết định của người dùng và khai thác tài nguyên trên nền tảng Internet có hiệu quả hơn giảm quá tải thông tin. Ứng dụng trí tuệ nhân tạo trong việc đưa ra đề xuất phim hoạt hình.

CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ

5.1. Kết luận

Nghiên cứu đã đạt được các kết quả chính như sau:

Xây dựng mô hình huấn luyện dữ liệu thứ cấp đề xuất phim hoạt hình bằng thuật toán lọc cộng tác và thuật toán KNN.

Xây dựng ứng dụng web tích hợp hệ thống đề xuất phim hoạt hình gồm tên phim hoạt hình và hình ảnh phim hoạt hình.

5.2. Ý nghĩa

Ý nghĩa khoa học: Kết quả của đề tài cung cấp cơ sở để đánh giá việc áp dụng Artificial Intelligence (Trí tuệ nhân tạo) và khả năng khai phá thông tin vào dữ liệu phim hoạt hình nói riêng cũng như dữ liệu trên nền tảng Internet nói chung trong việc hỗ trợ ra quyết định của người dùng.

Ý nghĩa thực tiễn: Kết quả nghiên cứu của đề tài hỗ trợ người dùng xem phim hoạt hình có hành vi tương tự với người dùng khác bằng phương pháp đề xuất phù hợp với hành vi của người dùng dựa trên hành vi trước đó đã được thu thập. Kết quả đề xuất mang lại lợi ích cho người dùng trong việc tìm kiếm có hiệu quả hơn và không quá tải thông tin phim hoạt hình.

5.3. Kiến nghị

Vấn đề của hệ thống đề xuất đã được đề cập và ứng dụng trong nhiều lĩnh vực khác nhau. Với kết quả hệ thống đề xuất phim hoạt hình mà đề tài đã thực hiện được thì quá trình mở rộng đề tài sẽ thêm nhiều yếu tố như mô tả phim hoạt hình, phân loại theo độ tuổi, giới tính, khu vực, thời lượng phim hoạt hình, bình luận,... từ đó có thể tăng tiêu chí để đưa ra đề xuất phù hợp với người dùng hơn và tăng trải nghiệm người dùng.

Bên cạnh đó, nghiên cứu đề xuất một số hướng phát triển tiếp theo như sau:

Về dữ liệu: cập nhật thêm dữ liệu mới tại myanimelist.net.

Về giải thuật: huấn luyện các mô hình khác như: Content-based Filtering, Hybrid Filtering, Knowledge-based.

TÀI LIỆU THAM KHẢO

Tài liệu tham khảo từ hệ thống internet

- Aeschliman, L. (2024). What is *Anime*? . Được truy lục từ Internet Archive Wayback Machine:
<https://web.archive.org/web/20071107150423/http://www.bellaonline.com/articles/art4260.asp> . [Truy cập ngày 23/04/2024].
- Anaconda, Inc. (2024). Anaconda Documentation. Được truy lục từ Anaconda:
<https://docs.anaconda.com/> [Truy cập ngày 23/06/2024].
- Apáthy, S. (2019). *History of recommender systems*. Được truy lục từ Onespire:
<https://onespire.net/news/history-of-recommender-systems/>
- Burke, R. (2000). *Knowledge-based recommender systems*. Được truy lục từ Encyclopedia of library and information systems, 69(Supplement 32), 175-186.
- Debashis Das, et al. (2017). A Survey on Recommendation System. *International Journal of Computer*. Được truy lục từ
<https://www.researchgate.net/publication/313787463>
- DO, Q.-V. (2020). *MyAnimeList*. Được truy lục từ Kaggle:
<https://www.kaggle.com/datasets/qvinhdo/myanimelist> . [Dữ liệu]
- Farhin Mansur, et al. (2017, 3). A review on recommender systems. *2017 4th International Conference on Innovations in Information, Embedded and Communication Systems*. doi:10.1109/ICIIECS.2017.8276182
- Jupyter. (2024). *Jupyter Notebook Documentation*. Được truy lục từ Jupyter:
<https://jupyter-notebook.readthedocs.io/en/latest/> . [Truy cập ngày 24/04/2024]
- Google for Developers (2024). *Collaborative Filtering Advantages & Disadvantages* . Được truy lục từ Google: <https://developers.google.com/machine-learning/recommendation/collaborative/summary> . [Truy cập ngày 16/05/2024]
- Ian MacKenzie, et al. (2013). How retailers can keep up with consumers. *McKinsey & Company*.

- Kemp, S. (2023). *Digital 2023: global overview report*. Được truy lục từ <https://datareportal.com/reports/digital-2023-global-overview-report> . [Truy cập ngày 15/05/2024]
- Mathias Uta, et al. (2024). *Knowledge-based recommender system: overview and research directions*. doi:10.3389/fdata.2024.1304439
- Microsoft. (2018). *Có gì mới trong Excel 2019 cho Windows*. Được truy lục từ Microsoft: <https://support.microsoft.com/vi-vn/office/c%C3%B3-g%C3%AC-m%E1%BB%9Bi-trong-excel-2019-for-windows-5a201203-1155-4055-82a5-82bf0994631f>
- Park Thaichon, et al. (2022). *Artificial Intelligence for Marketing Management*. LonDon. doi:<https://doi.org/10.4324/9781003280392>
- Scikit-learn developers. (2024). scikit-learn 1.4.2. Được truy lục từ PyPi: <https://pypi.org/project/scikit-learn/>. [Truy cập ngày 24/04/2024].
- SciPy Developers. (2024). scipy 1.13.0. Được truy lục từ PyPi: <https://pypi.org/project/scipy/> . [Truy cập ngày 24/04/2024].
- Shetty, B. (2023). *Recommender Systems: What Are They And How Do They Work?* . Được truy lục từ Built In: <https://builtin.com/data-science/recommender-systems>
- Sihar Simbolon, et al. (2020). Collaborative Recommendation System in Users of Anime Films. *Journal of Physics: Conference Series*. Được truy lục từ https://www.researchgate.net/publication/342690182_Collaborative_Recommendation_System_in_Users_of_Anime_Films
- Smith, J. (2012). *Recommendation system*. Được truy lục từ Citizendium: https://citizendium.org/wiki/Recommendation_system.
- Snowflake Inc. (2024). streamlit 1.33.0. Được truy lục từ Pypi: <https://pypi.org/project/streamlit/>. [Truy cập ngày 24/04/2024].
- Sphinx 7.3.7. (2024). *pickle — Python object serialization*. Được truy lục từ Python: <https://docs.python.org/3/library/pickle.html> . [Truy cập ngày 24/04/2024].

The Pandas Development Team. (2024). pandas 2.2.2. Được truy lục từ PyPi:
<https://pypi.org/project/pandas/> . [Truy cập ngày 24/04/2024].

Travis E. Oliphant, et al. (2024). numpy 1.26.4. Được truy lục từ PyPi:
<https://pypi.org/project/numpy/> . [Truy cập ngày 24/04/2024].

Visual Studio Code . (2024). *VS Code Day 2024*. Được truy lục từ Visual Studio Code
: <https://code.visualstudio.com/docs>. [Truy cập ngày 24/04/2024].

Wise, J. (2023). *How many people watch anime in 2024?* Retrieved from Earthweb:
<https://earthweb.com/blog/how-many-people-watch-anime/#>

Yatti, N. (2024). *What Are Cartoons*. Được truy lục từ Scrib:
<https://fr.scribd.com/document/242477934/What-Are-Cartoons> . [Truy cập
ngày 15/05/2024]

Tài liệu tham khảo từ tạp chí

Alam, A. (2023). What is Machine Learning? *Zenodo*, version v1 . Được truy lục từ
Zenodo: doi:10.5281/zenodo.8231580

Athira Ramasundaran. S. R , et al. (2023). Anime Recommendation System.
International Journal for Multidisciplinary Research, pp.2582-2160. Được truy
lục từ <https://www.ijfmr.com/papers/2023/2/2372.pdf>

Chen, J. (2023). The Investigation on Anime-Themed Recommendation System.
Highlights in Science, Engineering and Technology, vol.81, 2024. Được truy
lục từ <https://drpress.org/ojs/index.php/HSET/article/view/15926>

Jiaye Li, et al. (2023). Quantum KNN Classification With K Value Selection and
Neighbor Selection. *IEEE Transactions on Computer-Aided Design of
Integrated Circuits and Systems*, vol.43, no.5, pp.1332-1345, 2024.
doi:10.1109/TCAD.2023.3345251

Shashidhar Reddy Javaji, Krutika Sarode. (2023). Hybrid Recommendation System
using Graph Neural Network and BERT Embeddings. *Computer Science ,
Information Retrieval*, version 1. Được truy lục từ ArXiv:
<https://arxiv.org/abs/2310.04878>

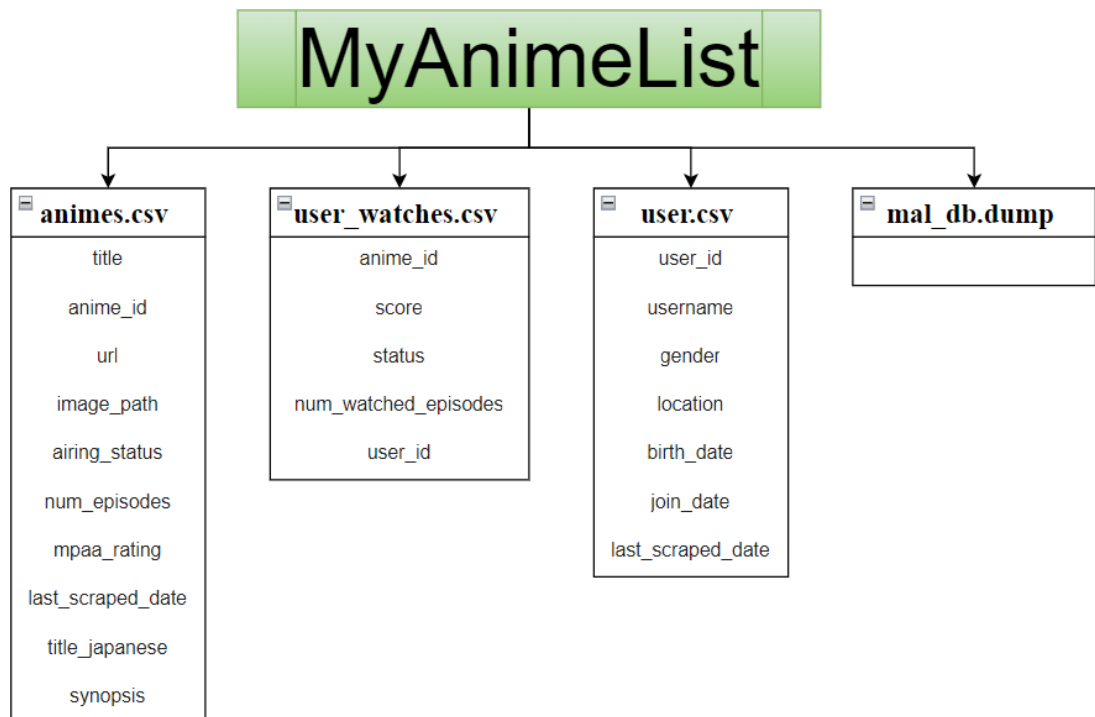
PHỤ LỤC

Phụ lục 1: GIẢI THÍCH LẬP TRÌNH

1. Procedure Artificial Intelligence

1.1. Collecting Data (Thu thập dữ liệu)

Bộ dữ liệu thứ cấp MyAnimeList của Kaggle được tải lên năm 2020 bởi QUANG-VINH DO tại <https://www.kaggle.com/datasets/qvinhdo/myanimelist> được thể hiện như Hình 1:



Hình 1: Dữ liệu thứ cấp MyAnimeList

Tập dữ liệu với dung lượng khoảng 943 MB chứa tập hợp các anime, người dùng và xếp hạng được lấy từ MyAnimeList.net. Tập dữ liệu này chứa 4 tệp.

- animes.csv: gồm 17058 phim hoạt hình chứa thông tin về tiêu đề, anime_id, trạng thái phát sóng, số tập và tóm tắt.
- user.csv: gồm 302674 người dùng chứa thông tin người dùng đơn giản như tên người dùng, giới tính, vị trí, ngày sinh và ngày tham gia.

- user_watches.csv: gồm 68235827 xếp hạng danh sách anime của người dùng cùng với điểm số và trạng thái xem.
- mal_db.dump - Tập kết xuất của cơ sở dữ liệu postgresql chứa tất cả 3 thông tin csv ở trên cùng với Primary Key/ Foreign Key thích hợp, các ràng buộc và chỉ mục khác.

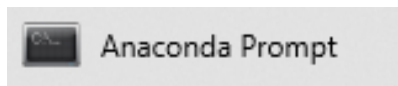
1.2. Preparing Data (Chuẩn bị dữ liệu)

Bước 1: Cài đặt phần mềm

Các phần mềm đề tài nghiên cứu sử dụng được giới thiệu tại mục 3.1.2. Phần mềm và thư viện

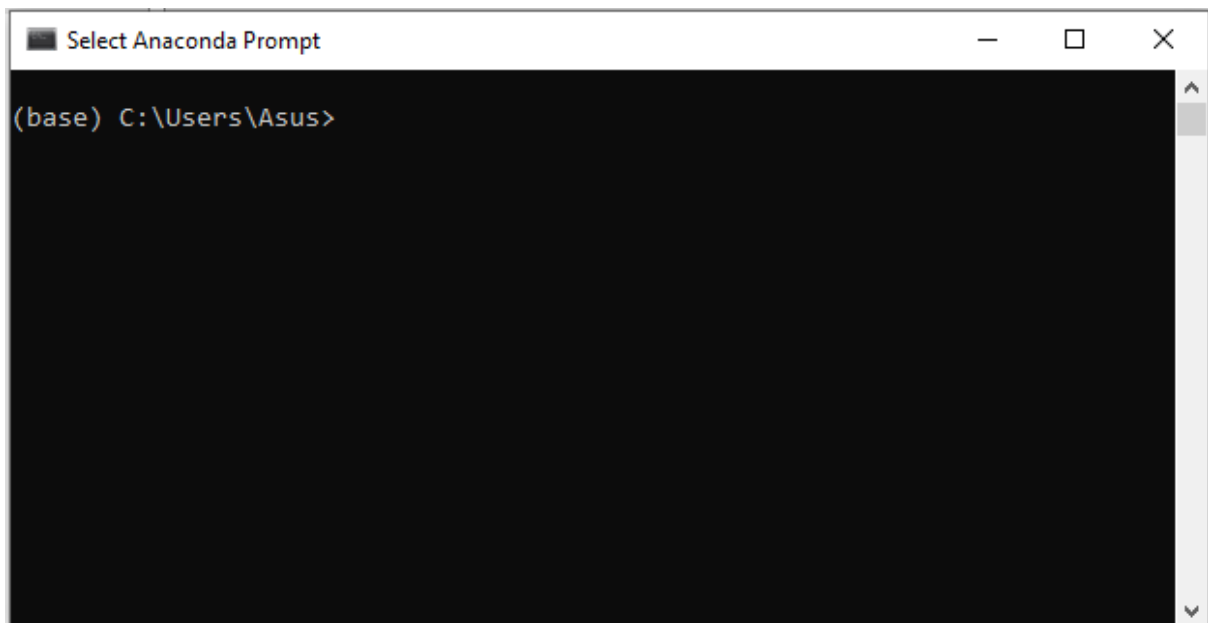
Bước 2: Cài đặt viện Python 3.12

Mở Anaconda Prompt



Hình 2: Phần mềm Anaconda Prompt

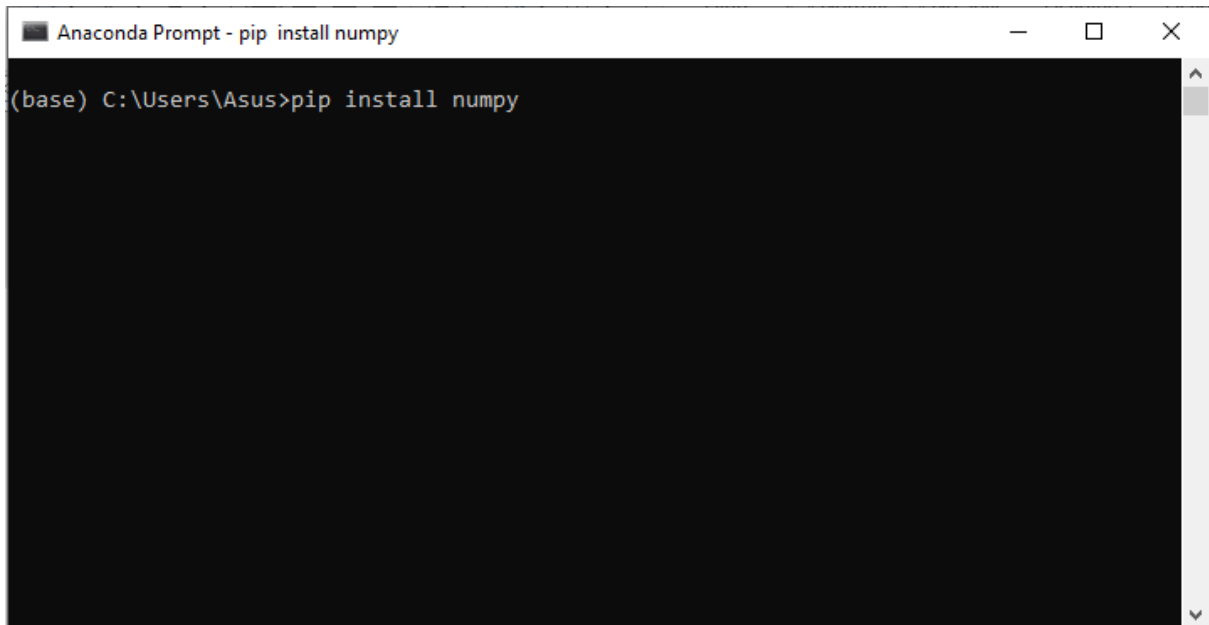
Xuất hiện cửa sổ lệnh Select Anaconda Prompt



Hình 3: Cửa sổ lệnh của phần mềm Anaconda Prompt

Cài đặt các thư viện của Python

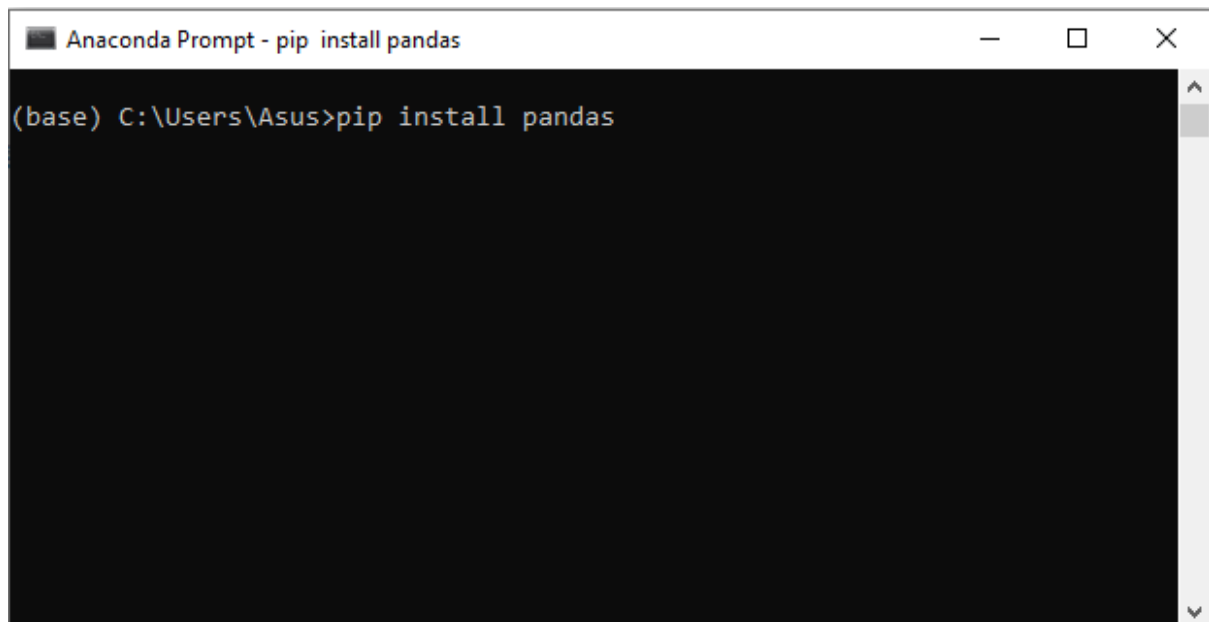
- Thư viện Numpy

A screenshot of an Anaconda Prompt window. The title bar reads "Anaconda Prompt - pip install numpy". The command prompt shows the text "(base) C:\Users\Asus>pip install numpy". The rest of the window is black, indicating the command has been executed and the output is not visible in this view.

```
Anaconda Prompt - pip install numpy
(base) C:\Users\Asus>pip install numpy
```

Hình 4: Cài đặt thư viện Numpy

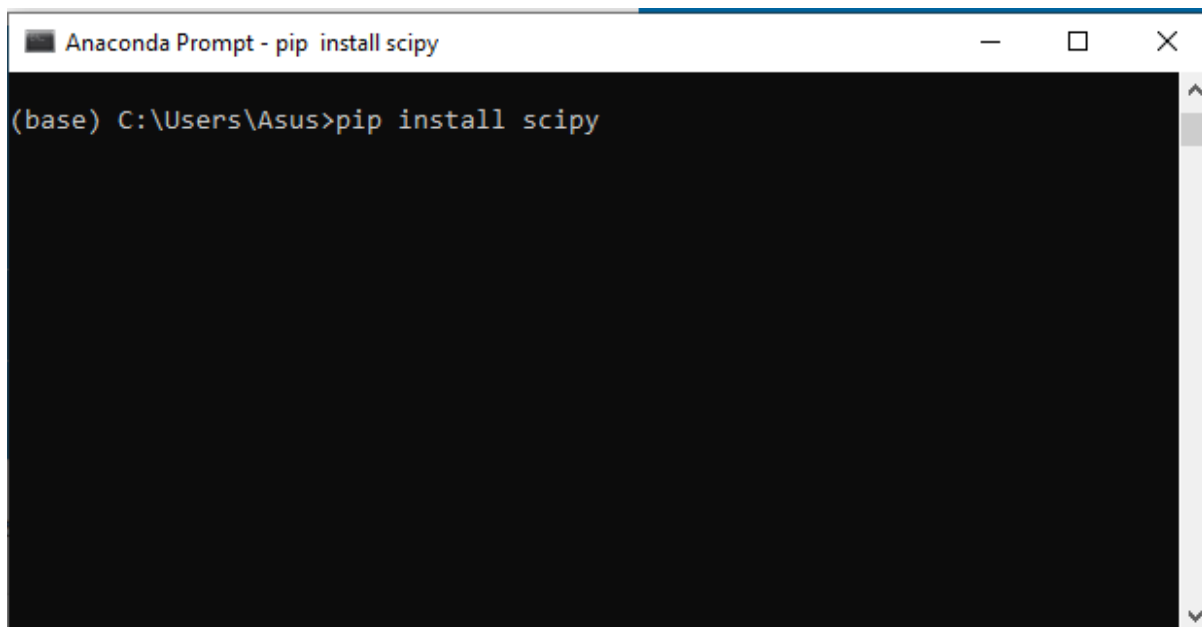
- Thư viện pandas

A screenshot of an Anaconda Prompt window. The title bar reads "Anaconda Prompt - pip install pandas". The command prompt shows the text "(base) C:\Users\Asus>pip install pandas". The rest of the window is black, indicating the command has been executed and the output is not visible in this view.

```
Anaconda Prompt - pip install pandas
(base) C:\Users\Asus>pip install pandas
```

Hình 5: Cài đặt thư viện Pandas

- Thư viện Scipy

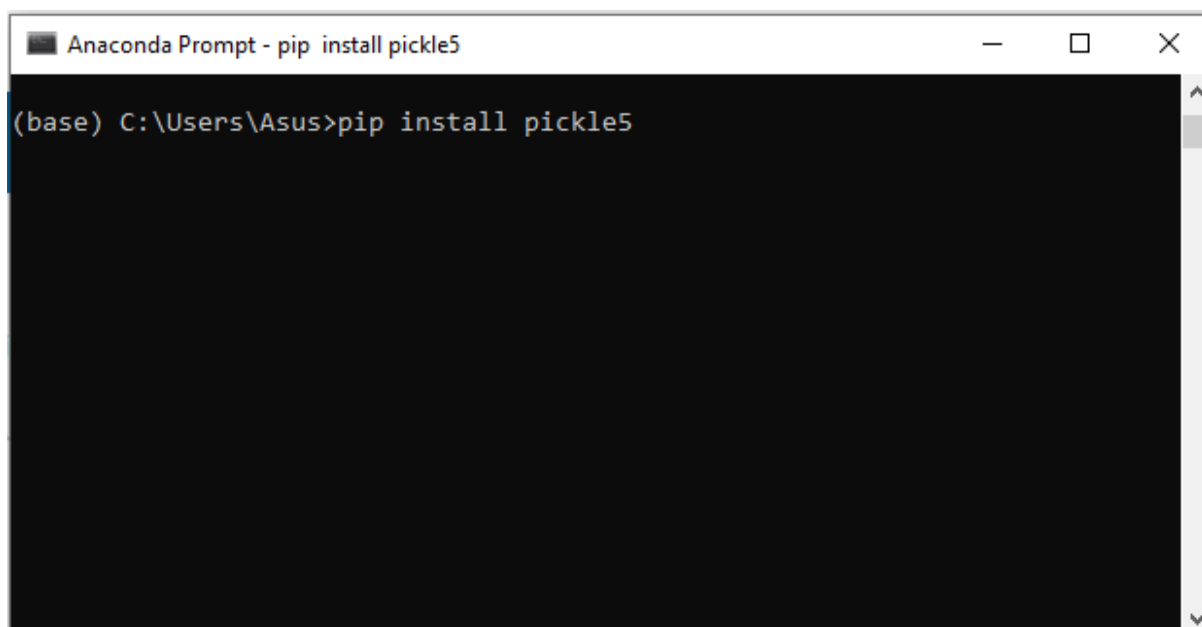
A screenshot of an Anaconda Prompt window. The title bar reads "Anaconda Prompt - pip install scipy". The command prompt shows the text "(base) C:\Users\Asus>pip install scipy". The rest of the window is empty, indicating the command has been entered but not yet executed or the output is not visible.

```
Anaconda Prompt - pip install scipy

(base) C:\Users\Asus>pip install scipy
```

Hình 6: Cài đặt thư viện Scipy

- Thư viện Pickle

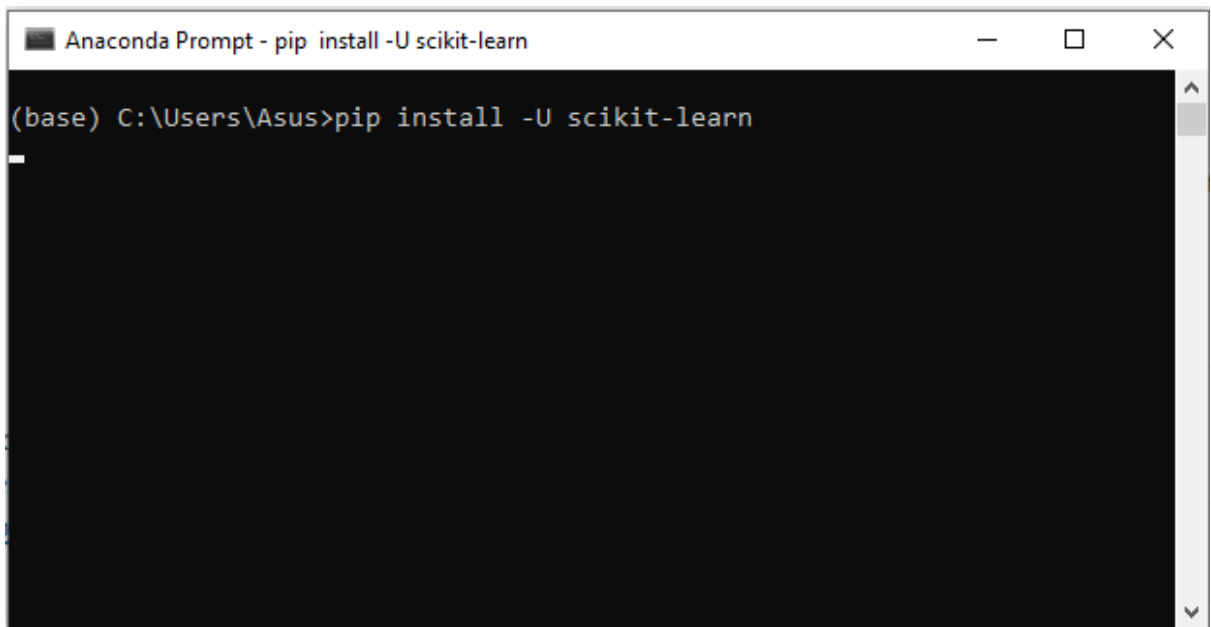
A screenshot of an Anaconda Prompt window. The title bar reads "Anaconda Prompt - pip install pickle5". The command prompt shows the text "(base) C:\Users\Asus>pip install pickle5". The rest of the window is empty, indicating the command has been entered but not yet executed or the output is not visible.

```
Anaconda Prompt - pip install pickle5

(base) C:\Users\Asus>pip install pickle5
```

Hình 7: Cài đặt thư viện Pickle5

- Thư viện Scikit-learn (Sklearn)

A screenshot of an Anaconda Prompt window. The title bar reads "Anaconda Prompt - pip install -U scikit-learn". The command prompt shows the command `(base) C:\Users\Asus>pip install -U scikit-learn` entered. The cursor is at the end of the command line.

```

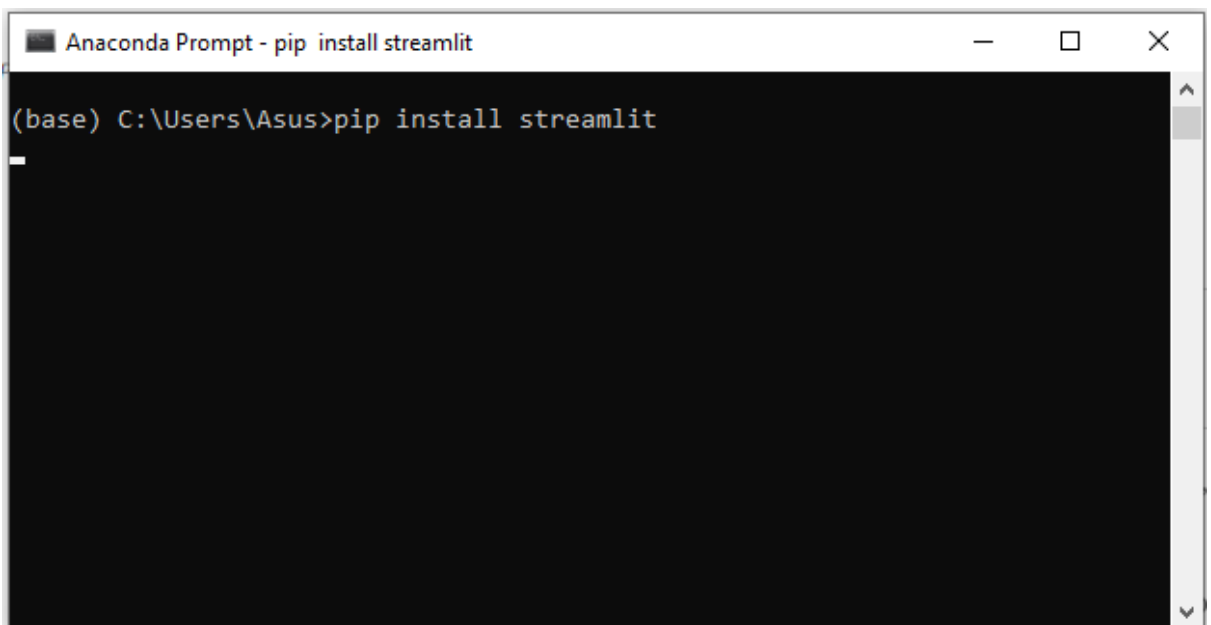
Anaconda Prompt - pip install -U scikit-learn

(base) C:\Users\Asus>pip install -U scikit-learn

```

Hình 8: Cài đặt thư viện Scikit-learn

- Thư viện Streamlit

A screenshot of an Anaconda Prompt window. The title bar reads "Anaconda Prompt - pip install streamlit". The command prompt shows the command `(base) C:\Users\Asus>pip install streamlit` entered. The cursor is at the end of the command line.

```

Anaconda Prompt - pip install streamlit

(base) C:\Users\Asus>pip install streamlit

```

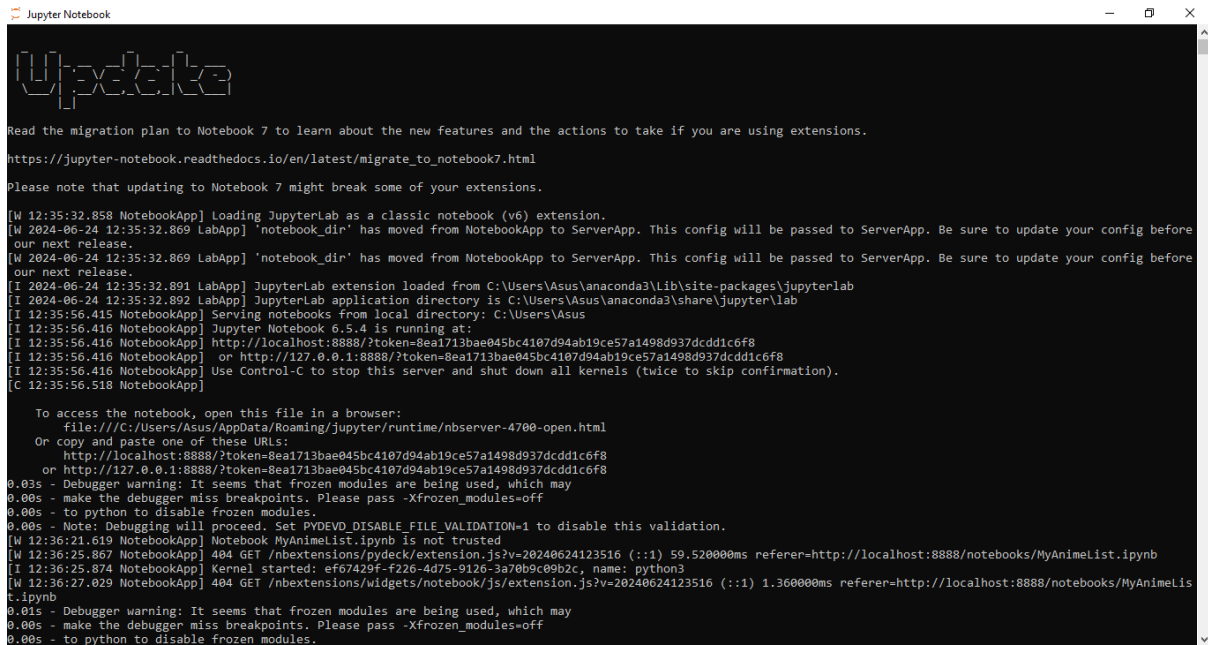
Hình 9: Cài đặt thư viên Streamlit

Bước 3: Phân tích dữ liệu

Quy trình phân tích dữ liệu gồm các bước được thực hiện tại mục 3.1.1.2. Phân tích và thống kê dữ liệu.

Bước 4: Chuẩn bị dữ liệu

❖ Mở phần mềm Jupyter Notebook



```
Jupyter Notebook

Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions.
https://jupyter-notebook.readthedocs.io/en/latest/migrate_to_notebook7.html

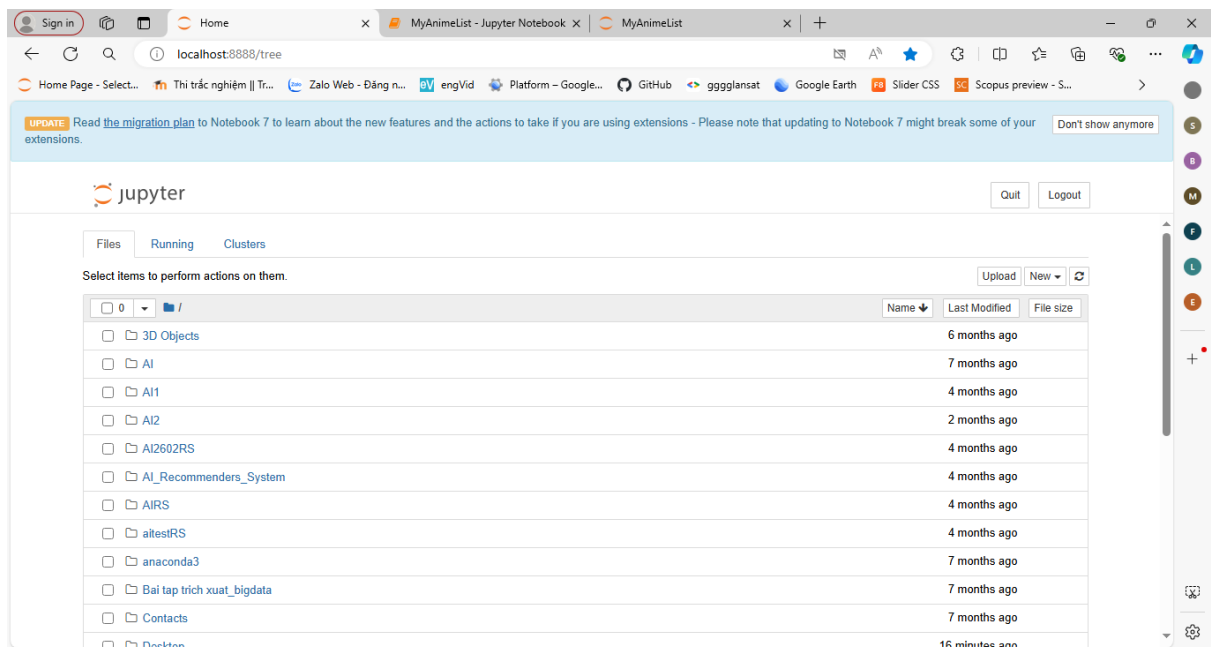
Please note that updating to Notebook 7 might break some of your extensions.

[W 12:35:32.858 NotebookApp] Loading JupyterLab as a classic notebook (v6) extension.
[W 2024-06-24 12:35:32.869 LabApp] 'notebook_dir' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2024-06-24 12:35:32.869 LabApp] 'notebook_dir' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[I 2024-06-24 12:35:32.891 LabApp] JupyterLab extension loaded from C:\Users\Asus\anaconda3\Lib\site-packages\jupyterlab
[I 2024-06-24 12:35:32.892 LabApp] JupyterLab application directory is C:\Users\Asus\anaconda3\share\jupyter\lab
[I 12:35:56.415 NotebookApp] Serving notebooks from local directory: C:\Users\Asus
[I 12:35:56.416 NotebookApp] Jupyter Notebook 6.5.4 is running at:
[I 12:35:56.416 NotebookApp] http://localhost:8888/?token=8ea1713bae045bc4107d94ab19ce57a1498d937dcd1c6f8
[I 12:35:56.416 NotebookApp] or http://127.0.0.1:8888/?token=8ea1713bae045bc4107d94ab19ce57a1498d937dcd1c6f8
[I 12:35:56.416 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 12:35:56.518 NotebookApp]

To access the notebook, open this file in a browser:
file:///C:/Users/Asus/AppData/Roaming/jupyter/runtime/nbserver-4700-open.html
Or copy and paste one of these URLs:
http://localhost:8888/?token=8ea1713bae045bc4107d94ab19ce57a1498d937dcd1c6f8
or http://127.0.0.1:8888/?token=8ea1713bae045bc4107d94ab19ce57a1498d937dcd1c6f8
0.03s - Debugger warning: It seems that frozen modules are being used, which may
0.00s - make the debugger miss breakpoints. Please pass -Xfrozen_modules=off
0.00s - to python to disable frozen modules.
0.00s - Note: Debugging will proceed. Set PYDEVD_DISABLE_FILE_VALIDATION=1 to disable this validation.
[W 12:36:21.619 NotebookApp] Notebook MyAnimeList.ipynb is not trusted
[W 12:36:25.867 NotebookApp] 404 GET /nbextensions/pydeck/extension.js?v=20240624123516 (::1) 59.520000ms referer=http://localhost:8888/notebooks/MyAnimeList.ipynb
[I 12:36:25.874 NotebookApp] Kernel started: ef67429f-f226-4d75-9126-3a70b9c09b2c, name: python3
[W 12:36:27.029 NotebookApp] 404 GET /nbextensions/widgets/notebook/js/extension.js?v=20240624123516 (::1) 1.360000ms referer=http://localhost:8888/notebooks/MyAnimeList.ipynb
0.01s - Debugger warning: It seems that frozen modules are being used, which may
0.00s - make the debugger miss breakpoints. Please pass -Xfrozen_modules=off
0.00s - to python to disable frozen modules.
```

Hình 10: Giao diện phần mềm Jupyter Notebook

Sau khi mở phần mềm, tạo một thư mục mới tên “AI2”.



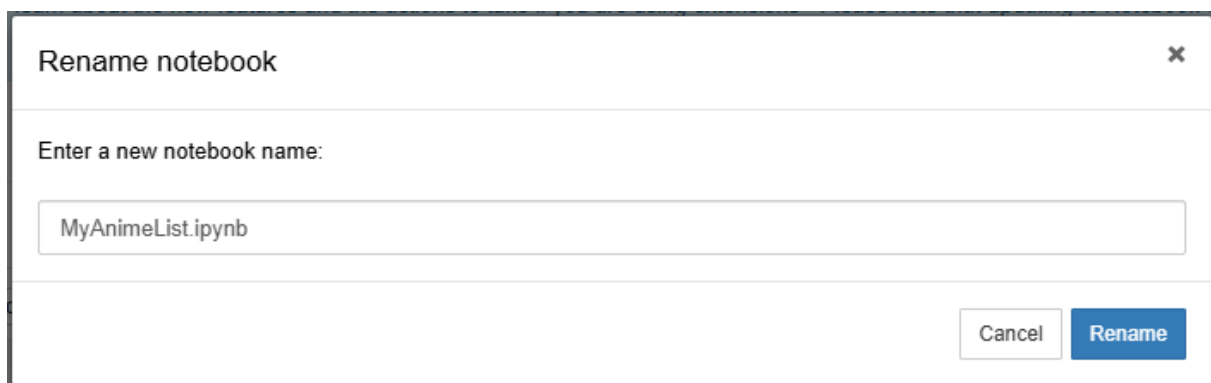
Hình 11: Giao diện web của phần mềm Jupyter Notebook

Mở thư mục AI2 và tải các tệp chứa thông tin của dữ liệu vào thư mục AI2.



Hình 12: Tải dữ liệu vào thư mục AI2

Tạo thêm một tệp Python 3 có định dạng .ipynb và đặt tên là MyAnimeList.ipynb.



Hình 13: Đặt tên cho tệp Python 3

Mở file MyAnimeList.ipynb.

❖ *Gọi thư viện*

- Thư viện Numpy

```
import numpy as np
```

Hình 14: Lệnh gọi thư viện Numpy

- Thư viện Pandas

```
import pandas as pd
```

Hình 15: Lệnh gọi thư viện Pandas

- Thư viện Scipy

```
from scipy.sparse import csr_matrix
```

Hình 16: Lệnh gọi thư viện Scipy

- Thư viện Pickle

```
import pickle
```

Hình 17: Lệnh gọi thư viện Pickle

- Thư viện Scikit-learn (Sklearn)

```
from sklearn.neighbors import NearestNeighbors
```

Hình 18: Lệnh gọi thư viện Sklearn

- Thư viện Streamlit

```
import streamlit as st
```

Hình 19: Lệnh gọi thư viện Streamlit

❖ Đọc dữ liệu

- animes.csv

```
animes = pd.read_csv('animes.csv')
animes.head(10)
```

Hình 20: Lệnh đọc tệp animes.csv

	title	anime id	url	image path	airing status	num episodes	mpaa rating	last scraped date	title japanese	synopsis	title english
0	Hanasaku Iroha	9289	https://myanimelist.net/anime/9289/Hanasaku_Iroha	https://cdn.myanimelist.net/images/anime/3/289...	2	26.0	PG-13 - Teens 13 or older	2020-09-13 11:23:05.509374	花咲くいろは	Ohana Matsuura is an energetic and wild teenag...	Hanasaku Iroha: Blossoms for Tomorrow
1	Uchuu Kyoudai	12431	https://myanimelist.net/anime/12431/Uchuu_Kyoudai	https://cdn.myanimelist.net/images/anime/7/375...	2	99.0	PG-13 - Teens 13 or older	2020-09-13 12:58:18.033841	宇宙兄弟	On a fateful summer night in 2006, Mutta Nanba...	Space Brothers
2	Yozakura Quartet: Hana no Uta	18497	https://myanimelist.net/anime/18497/Yozakura_Q...	https://cdn.myanimelist.net/images/anime/5/525...	2	13.0	PG-13 - Teens 13 or older	2020-09-14 14:39:45.15066	夜桜四重奏 - ハナノタ〜	Hundreds of years ago, the borders between the...	NaN
3	Glass no Kamen Desu ga to Z Specials	34336	https://myanimelist.net/anime/34336/Glass_no_K...	https://cdn.myanimelist.net/images/anime/13/82...	2	2.0	PG-13 - Teens 13 or older	2020-09-14 01:53:48.51646	ガラスの仮面ですが、とZ	Two unaired episodes of Glass no Kamen Desu ga...	NaN
4	Re:Zero kara Hajimeru Isekai Seikatsu 2nd Season	39587	https://myanimelist.net/anime/39587/Re_Zero_ka...	https://cdn.myanimelist.net/images/anime/1444/...	1	13.0	R - 17+ (violence & profanity)	2020-09-14 06:45:31.052373	Re:ゼロから始める異世界生活 2nd season	Even after dying countless times, Subaru final...	Re:ZERO - Starting Life in Another World - Season 2
5	Sword Art Online: Alicization - War of Underw...	40540	https://myanimelist.net/anime/40540/Sword_Art_...	https://cdn.myanimelist.net/images/anime/14/38/...	1	11.0	PG-13 - Teens 13 or older	2020-09-14 07:30:27.945843	ソードアート・オンライン アリシゼーション War of Underworld	Second part of Sword Art Online: Alicization ...	Sword Art Online Alicization War of Underworld...
6	Taru no Magical Holiday	25545	https://myanimelist.net/anime/25545/Taru_no_Ma...	https://cdn.myanimelist.net/images/anime/13/65...	2	1.0	G - All Ages	2020-09-13 16:44:33.037659	たぬのマジカルホリデー		NaN
7	Kimikiss Pure Rouge	2927	https://myanimelist.net/anime/2927/KimiKiss_Pu...	https://cdn.myanimelist.net/images/anime/7/199...	2	24.0	PG-13 - Teens 13 or older	2020-09-13 06:48:05.084739	キミキス pure rouge	Kouchi Sanada and Kazuki Aihara's childhood f...	Kimikiss: Pure Rouge
8	Tokyo Magnitude 8.0	6211	https://myanimelist.net/anime/6211/Tokyo_Magni...	https://cdn.myanimelist.net/images/anime/10/13...	2	11.0	PG-13 - Teens 13 or older	2020-09-13 09:35:06.970559	東京マグニチュード8.0	Middle school student Mirai Onozawa is disast...	Tokyo Magnitude 8.0
9	Jankenman	7405	https://myanimelist.net/anime/7405/Jankenman	https://cdn.myanimelist.net/images/anime/2/178...	2	51.0	PG - Children	2020-09-13 10:19:02.552156	ジャンケンマン	There is a small village called Janken Town. T...	NaN

Hình 21: Kết quả đọc tệp animes.csv

- users.csv

```
import pandas as pd
users = pd.read_csv('users.csv')
users.head(10)
```

Hình 22: Lệnh đọc tệp users.csv

	user_id	username	gender	location	birth_date	join_date	last_scraped_date
0	1	Xinil	Male	California	1985-03-04 00:00:00	2004-11-05 00:00:00	2020-09-02 15:20:23.893907
1	3	Aokaado	Male	Oslo, Norway	1988-11-11 00:00:00	2004-11-11 00:00:00	2020-09-02 15:20:23.893907
2	4	Crystal	Female	Melbourne, Australia	1989-01-10 00:00:00	2004-11-13 00:00:00	2020-09-02 15:20:23.893907
3	20	vondur	Male	Bergen, Norway	1988-01-25 00:00:00	2005-01-05 00:00:00	2020-09-02 15:20:23.893907
4	23	Amuro	NaN	Canada	1988-02-22 00:00:00	2005-01-23 00:00:00	2020-09-02 15:20:23.893907
5	37	megan	Female	San Diego, California	1987-06-18 00:00:00	2005-02-12 00:00:00	2020-09-02 15:20:23.893907
6	44	Koreth	Male	NaN	NaN	2005-02-21 00:00:00	2020-09-02 15:20:23.893907
7	47	kei-clone	Male	New York, NY	1988-01-01 00:00:00	2005-03-09 00:00:00	2020-09-02 15:20:23.893907
8	53	Ladholymán	NaN	NaN	NaN	2005-03-19 00:00:00	2020-09-02 15:20:23.893907
9	66	Hiromi	Male	Calgary, AB	1990-02-09 00:00:00	2005-04-01 00:00:00	2020-09-02 15:20:23.893907

Hình 23: Kết quả đọc tệp users.csv

- user_watches.csv

```
rating = pd.read_csv('user_watches.csv')
rating.head(10)
```

Hình 24: Lệnh đọc tệp user_watches.csv

	anime_id	score	status	num_watched_episodes	user_id
0	21	9	1	626	1
1	11061	9	2	30	1
2	8676	0	1	4	3
3	9919	0	1	2	3
4	5081	0	1	8	3
5	109	0	1	1	3
6	2986	6	1	14	3
7	269	9	1	201	3
8	31043	0	1	3	3
9	28223	0	1	7	3

Hình 25: Kết quả đọc tệp `user_watches.csv`

❖ Kích thước tệp

- `animes.csv`

```
animes.shape
```

Hình 26: Lệnh xác định kích thước của tệp `animes.csv`

(17058, 11)

Hình 27: Kết quả kích thước của tệp `animes.csv`

- `users.csv`

```
users.shape
```

Hình 28: Lệnh xác định kích thước của tệp `users.csv`

(302674, 7)

Hình 29: Kết quả kích thước của tệp `users.csv`

- user_watches.csv

```
rating.shape
```

Hình 30: Lệnh xác định kích thước của tệp user_watches.csv

```
(68235827, 5)
```

Hình 31: Kết quả kích thước của tệp user_watches.csv

❖ Thống kê

- animes.csv

```
round(animes.describe(),2)
```

Hình 32: Lệnh thống kê số cơ bản của tệp animes.csv làm tròn đến chữ số thập phân thứ 2

	anime_id	airing_status	num_episodes
count	17058.00	17058.0	16583.00
mean	20799.79	2.0	11.57
std	14576.79	0.2	47.92
min	1.00	1.0	1.00
25%	5664.25	2.0	1.00
50%	21772.00	2.0	2.00
75%	35058.75	2.0	12.00
max	42754.00	3.0	3057.00

Hình 33: Kết quả thống kê số cơ bản của tệp animes.csv

- users.csv

```
round(users.describe(),2)
```

Hình 34: Lệnh thống kê số cơ bản của tệp users.csv làm tròn đến chữ số thập phân thứ

	user_id
count	302674.00
mean	2061417.54
std	2118054.64
min	1.00
25%	232244.75
50%	1167586.00
75%	3883448.50
max	7273306.00

Hình 35: Kết quả thống kê số cơ bản của tệp users.csv

- user_watches.csv

```
round(rating.describe(),2)
```

Hình 36 : Lệnh thống kê số cơ bản của tệp user_watches.csv làm tròn đến chữ số thập phân thứ 2

	anime_id	score	status	num_watched_episodes	user_id
count	68235827.00	68235827.00	68235827.00	68235827.00	68235827.00
mean	13378.03	4.42	2.98	14.53	2199570.02
std	12595.30	3.94	1.72	359.81	2077056.68
min	1.00	0.00	0.00	0.00	1.00
25%	1954.00	0.00	2.00	0.00	288885.00
50%	9513.00	6.00	2.00	4.00	1407861.00
75%	23251.00	8.00	4.00	13.00	4044265.00
max	42754.00	10.00	55.00	65535.00	7273306.00

Hình 37: Kết quả thống kê số cơ bản của tệp user_watches.csv

1.3. Choosing model (Chọn mô hình)

Mô hình đề tài nghiên cứu được giới thiệu tại mục 2.3. Tổng quan phương pháp nghiên cứu.

Value_count() trả về đối tượng chứa số lượng giá trị duy nhất. Ở đây có thể xác định được người dùng đã xếp hạng cho bao nhiêu phim hoạt hình, ví dụ cho user_id 4053169 đã xếp hạng cho 17058 phim hoạt hình và user_id 323300 đã xếp hạng cho 1 phim hoạt hình.

```
rating['user_id'].value_counts()
```

Hình 38: Lệnh Value_count()

```
user_id
4053169    17058
4676838    16583
477008     16557
3979333    14428
1381655    14200
...
271793         1
271752         1
30642          1
271689         1
323300         1
Name: count, Length: 259137, dtype: int64
```

Hình 39: Kết quả Value_count()

Kích thước Value_count() của cột user_id trong user_watches.csv.

```
rating['user_id'].value_counts().shape
```

Hình 40: Lệnh xác định kích thước Value_count() của cột user_id trong user_watches.csv

```
(259137,)
```

Hình 41: Kết quả kích thước Value_count() của cột user_id trong user_watches.csv

Kiểm tra giá trị duy nhất của cột user_id trong user_watches.csv.

```
rating['user_id'].unique().shape
```

Hình 42: Lệnh xác kích thước `unique()` của cột `user_id` trong `user_watches.csv`

```
(259137,)
```

Hình 43: Kết quả kích thước `unique()` của cột `user_id` trong `user_watches.csv`

Lọc để loại các giá trị nhỏ hơn 1000. Tiến hành lọc bỏ đi các `user_id` của người dùng đã xếp hạng cho ít hơn 10000 phim hoạt hình nhằm tăng độ tin cậy cho dữ liệu.

```
x = rating['user_id'].value_counts()>1000  
x[x].shape
```

Hình 44: Lệnh lọc các giá trị lớn hơn 1000

Và kết quả lọc có 8840 `user_id` đã xếp hạng cho hơn 1000 phim hoạt hình.

```
(8840,)
```

Hình 45: Kết quả lọc

Tìm `user_id` của người dùng.

```
y = x[x].index  
y
```

Hình 46: Lệnh tìm `user_id` của người dùng

```
Index([4053169, 4676838, 477008, 3979333, 1381655, 4561255, 441467, 1636745,  
       1493371, 6011354,  
       ...  
       17335, 3498629, 561795, 4484242, 1211967, 319777, 304295, 321360,  
       3084345, 5939940],  
      dtype='int64', name='user_id', length=8840)
```

Hình 47: Kết quả `user_id` của người dùng

Sắp xếp các xếp hạng và chuyển lại xếp hạng theo `user_id`.

```
rating = rating[rating['user_id'].isin(y)]  
rating.head(10)
```

Hình 48: Lệnh sắp xếp các xếp hạng theo `user_id` và trình bày 10 dòng đầu tiên

	anime_id	score	status	num_watched_episodes	user_id
597	29893	6	1	1	36
598	40056	6	1	6	36
599	40956	6	1	7	36
600	41120	6	1	5	36
601	28537	5	1	1	36
602	40074	5	1	5	36
603	40942	5	1	11	36
604	39730	6	1	7	36
605	2598	5	1	28	36
606	40196	5	1	5	36

Hình 49: Kết quả sắp xếp các xếp hạng theo user_id

Kích thước của sắp xếp các xếp hạng theo user_id.

```
rating.shape
```

Hình 50: Lệnh xác định kích thước các xếp hạng theo user_id

```
(13331895, 5)
```

Hình 51: Kết quả kích thước các xếp hạng theo user_id

Gộp tệp “animes.csv” và “user_watches.csv” theo cột anime_id.

```
rating_animes = rating.merge(anime, on = 'anime_id')
rating_animes.head(10)
```

Hình 52: Lệnh gộp và trình bày 10 dòng đầu tiên

	anime_id	score	status	num_watched_episodes	user_id	title	image_path	synopsis
0	29893	6	1	1	36	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
1	29893	1	2	3	228	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
2	29893	8	2	3	540	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
3	29893	9	2	3	813	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
4	29893	0	6	0	922	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
5	29893	0	3	0	2704	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
6	29893	10	2	3	2976	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
7	29893	0	6	0	3447	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
8	29893	0	6	0	3703	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...
9	29893	10	2	3	3807	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...

Hình 53: Kết quả gộp

Kích thước sau khi gộp.

```
rating_animes.shape
```

Hình 54: Lệnh xác định kích thước sau khi gộp

```
(13331895, 8)
```

Hình 55: Kết quả kích thước sau khi gộp

Nhóm cột title và score. Số lượng các xếp hạng của phim hoạt hình.

```
number_rating = rating_animes.groupby('title')['score'].count().reset_index()
number_rating.head()
```

Hình 56: Lệnh nhóm cột title và score

	title	score
0	"0"	347
1	"Aesop" no Ohanashi yori: Ushi to Kaeru, Yokub...	27
2	"Bungaku Shoujo" Kyou no Oyatsu: Hatsukoi	1858
3	"Bungaku Shoujo" Memoire	2139
4	"Bungaku Shoujo" Movie	3357

Hình 57: Kết quả nhóm cột title và score

Thay đổi tên cột score thành num_of_score.

```
number_rating.rename(columns={'score': 'num_of_score'}, inplace=True)
number_rating.head()
```

Hình 58: Lệnh đổi tên cột score thành num_of_score

	title	num_of_score
0	"0"	347
1	"Aesop" no Ohanashi yori: Ushi to Kaeru, Yokub...	27
2	"Bungaku Shoujo" Kyou no Oyatsu: Hatsukoi	1858
3	"Bungaku Shoujo" Memoire	2139
4	"Bungaku Shoujo" Movie	3357

Hình 59: Kết quả đổi tên

Hợp nhất các cột theo cột title.

```
final_rating = rating_animes.merge(number_rating, on='title')
final_rating.head()
```

Hình 60: Lệnh gộp và trình bày 5 dòng đầu tiên

	anime_id	score	status	num_watched_episodes	user_id	title	image_path	synopsis	num_of_score
0	29893	6	1	1	36	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
1	29893	1	2	3	228	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
2	29893	8	2	3	540	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
3	29893	9	2	3	813	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
4	29893	0	6	0	922	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584

Hình 61: Kết quả gộp

Kích thước sau khi gộp.

```
final_rating.shape
```

Hình 62: Lệnh xác định kích thước sau khi gộp

```
(13331895, 9)
```

Hình 63: Kết quả sau khi gộp

Lọc loại bỏ các phim hoạt hình có nhỏ hơn hoặc bằng 50 xếp hạng.

```
final_rating = final_rating[final_rating['num_of_score'] >= 50]
final_rating.head()
```

Hình 64: Lệnh lọc các phim hoạt hình có hơn 50 xếp hạng

	anime_id	score	status	num_watched_episodes	user_id	title	image_path	synopsis	num_of_score
0	29893	6	1	1	36	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
1	29893	1	2	3	228	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
2	29893	8	2	3	540	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
3	29893	9	2	3	813	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
4	29893	0	6	0	922	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584

Hình 65: Kết quả các phim hoạt hình có hơn 50 xếp hạng

Kích thước sau khi lọc.

```
final_rating.shape
```

Hình 66: Lệnh xác định kích thước sau khi lọc

```
(13216557, 9)
```

Hình 67: Kết quả sau khi lọc

Tiêu đề của các cột.

```
final_rating.columns
```

Hình 68: Lệnh trả về các tiêu đề của các cột

```
Index(['anime_id', 'score', 'status', 'num_watched_episodes', 'user_id',  
      'title', 'image_path', 'synopsis', 'num_of_score'],  
      dtype='object')
```

Hình 69: Kết quả các tiêu đề cột

Điều chỉnh và loại bỏ một số cột không dùng.

```
final_rating = final_rating[['anime_id', 'score', 'user_id',
                             'title', 'image_path', 'synopsis', 'num_of_score']]
final_rating.head()
```

Hình 70: Lệnh điều chỉnh cột

	anime_id	score	user_id	title	image_path	synopsis	num_of_score
0	29893	6	36	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
1	29893	1	228	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
2	29893	8	540	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
3	29893	9	813	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584
4	29893	0	922	Aria the Avvenire	https://cdn.myanimelist.net/images/anime/10/74...	Now that Akari Mizunashi is a Prima Undine and...	1584

Hình 71: Kết quả điều chỉnh cột

Loại bỏ các giá trị trùng lặp ở cột user_id và title.

```
final_rating.drop_duplicates(['user_id', 'title'], inplace=True)
```

Hình 72: Lệnh loại bỏ các giá trị trùng lặp

Kích thước sau khi loại bỏ.

```
final_rating.shape
```

Hình 73: Lệnh xác định kích thước sau khi loại bỏ trùng lặp

```
(13216531, 7)
```

Hình 74: Kết quả kích thước sau khi loại bỏ trùng lặp

PivotTable

Utility Matrix:

User_id	36	228	540	...
Title				
Aria the Avvenire	6	1	8	...
Deca-Dence	6	0	NaN	...

11eyes: Momoiro Genmutan	6	10	6	...
Cowboy Bebop	9	9	10	...
...

Sử dụng PivotTable để nhanh chóng tóm tắt dữ liệu lớn. Ở đây việc xây dựng Utility Matrix là quan trọng để tập hợp tất cả điểm gồm các giá trị đã biết và chưa biết cần được dự đoán. Liệu người dùng 540 có thích “Deca-Dence” không? Dựa vào sự tương đồng về xếp hạng điểm số của người dùng 540 với người dùng 36 về phim hoạt hình “11eyes: Momoiro Genmutan” là 6. Nếu đề xuất “Deca-Dence” thì điểm có thể là bao nhiêu? Mô hình sự phân cụm, một số người dùng có xếp hạng tương đồng cho một số phim hoạt hình có thể có sở thích hoặc hành vi tương đồng nhau. Có dự đoán xếp hạng điểm số của người dùng 540 cho phim hoạt hình “Deca-Dence”.

Tạo một bảng tổng hợp với các cột user_id, title và score để xây dựng ma trận.

```
anime_pivot = final_rating.pivot_table(columns='user_id', index='title', values='score')
anime_pivot
```

Hình 75: Lệnh tạo ma trận

	user_id	36	228	384	436	448	488	527	540	592	614	...	7119604	7128939	7142335	7148055	7149397	7152703	7175775	7181454	7213650	7227104
title																						
"0"	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
"Bungaku Shoujo" Kyou no Oyatsu: Hatsukoi	NaN	6.0	NaN	NaN	NaN	NaN	NaN	NaN	7.0	7.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
"Bungaku Shoujo" Memoire	NaN	7.0	NaN	NaN	NaN	NaN	NaN	NaN	8.0	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
"Bungaku Shoujo" Movie	NaN	7.0	NaN	8.0	NaN	NaN	NaN	0.0	8.0	0.0	NaN	...	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
"Eiji"	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...
xxxHOLIC Movie: Manatsu no Yoru no Yume	NaN	7.0	10.0	7.0	7.0	0.0	NaN	7.0	0.0	0.0	NaN	0.0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
xxxHOLIC Rou	NaN	7.0	9.0	NaN	0.0	NaN	NaN	NaN	0.0	NaN	NaN	0.0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
xxxHOLIC Shunmuki	NaN	9.0	10.0	NaN	0.0	NaN	NaN	7.0	NaN	NaN	NaN	0.0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
eiDLIVE	NaN	NaN	0.0	7.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	NaN	4.0	0.0	NaN	NaN	NaN	NaN	7.0
○	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

12322 rows × 8840 columns

Hình 76: Kết quả tạo ma trận

Kiểm tra lại kích thước của ma trận.

```
anime_pivot.shape
```

Hình 77: Lệnh xác định kích thước ma trận

(12322, 8840)

Hình 78: Kết quả kích thước ma trận

Thay thế các giá trị NaN thành 0 để mô hình có thể tính toán khoảng cách giữa các cột.

```
anime_pivot.fillna(0, inplace=True)
anime_pivot
```

Hình 79: Lệnh chuyển đổi NaN thành 0

	user_id	36	228	384	436	448	488	527	540	592	614	...	7119604	7128939	7142335	7148055	7149397	7152703	7175775	7181454	7213650	7227104	
title																							
	"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	0.0	0.0	0.0	
"Bungaku Shoujo" Kyou no Oyatsu: Hatsukoi	0.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0	7.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
	"Bungaku Shoujo" Memoire	0.0	7.0	0.0	0.0	0.0	0.0	0.0	8.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
"Bungaku Shoujo" Movie	0.0	7.0	0.0	8.0	0.0	0.0	0.0	0.0	8.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
"Eiji"	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	0.0	0.0	0.0	
...	
xxxHOLIC Movie: Manatsu no Yoru no Yume	0.0	7.0	10.0	7.0	7.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
xxxHOLIC Rou	0.0	7.0	9.0	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	
xxxHOLIC Shunmuki	0.0	9.0	10.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
eIDLIVE	0.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	7.0	
○	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	0.0	0.0	0.0	

12322 rows × 8840 columns

Hình 80: Kết quả chuyển đổi

Sử dụng `csr_matrix` hàng thưa thớt được nén để cắt hàng nhanh, sản phẩm vector ma trận nhanh hơn.

```
anime_sparse = csr_matrix(anime_pivot)
```

Hình 81: Lệnh tạo hàng không gian nén

```
type(anime_sparse)
```

Hình 82: Lệnh kiểm tra loại kiểu nén

```
scipy.sparse._csr.csr_matrix
```

Hình 83: Kết quả loại kiểu nén

Xây dựng thuật toán KNN.

```
model = NearestNeighbors(algorithm='brute')
```

Hình 84: Lệnh gọi thuật toán

```
model.fit(anime_sparse)
```

Hình 85: Lệnh truyền hàng không gian nén

```
NearestNeighbors  
NearestNeighbors(algorithm='brute')
```

Hình 86: Kết quả lệnh gọi thuật toán KNN

1.4. Tuning Parameter (Điều chỉnh thông số)

Khởi tạo biến distance và suggestion. Xây dựng mô hình có chức năng tìm KNN và k=11 với giá trị trả về là 10 kết quả khoảng cách và đề xuất cho mục chỉ định là anime_id là 30.

```
distance, suggestion = model.kneighbors(anime_pivot.iloc[30,:].values.reshape(1,-1), n_neighbors=11)
```

Hình 87: Lệnh xây dựng mô hình

Khoảng cách 10 anime_id gần nhất với anime_id 30.

```
distance
```

Hình 88: Lệnh gọi biến distance

```
array([[ 0.          , 99.72963451, 99.84487969, 99.86490875,  
        99.91496384, 100.019998  , 100.039992  , 100.11493395,  
        100.1598722 , 100.1598722 , 100.23472452]])
```


Hình 89: Kết quả biến distance

Đề xuất 10 anime_id cho kết quả tìm kiếm với anime_id 30.

```
suggestion
```

Hình 90: Lệnh gọi suggestion

```
array([[ 30, 4084, 10989, 11807, 4161, 4985, 2000, 62, 3185,
        3132, 9106]], dtype=int64)
```

Hình 91: Kết quả biến suggestion

Lọc điểm xếp hạng (trên thang điểm 10) của user_id dành cho anime_id 1.

```
anime_pivot.iloc[1,:]
```

Hình 92: Lệnh lọc điểm xếp hạng của user_id dành cho anime_id 1

```
user_id
36      0.0
228     6.0
384     0.0
436     0.0
448     0.0
...
7152703 0.0
7175775 0.0
7181454 0.0
7213650 0.0
7227104 0.0
Name: "Bungaku Shoujo" Kyou no Oyatsu: Hatsukoi, Length: 8840, dtype: float64
```

Hình 93: Kết quả lọc điểm xếp hạng của user_id dành cho anime_id 1

Trình bày 10 đề xuất trên theo cột title.

```
for i in range(len(suggestion)):
    print(anime_pivot.index[suggestion[i]])
```

Hình 94: Lệnh đề xuất tên phim hoạt hình

```
Index(['009-1: R&B', 'Hengen Taima Yakou Karura Mau! Sendai Kokeshi Enka',
      'Toilet no Hanako-san', 'Xabungle Graffiti', 'Hidari no O'Clock!!',
      'Juuippiki no Neko to Ahoudori', 'D-1 Devastator',
      '21 Emon Uchuu ike! Hadashi no Princess',
      'Gakkou no Kowai Uwasa Shin: Hanako-san ga Kita!!',
      'Future GPX Cyber Formula Graffiti', 'Ruri-iro Princess'],
      dtype='object', name='title')
```

Hình 95: Kết quả đề xuất tên phim hoạt hình

Gọi tên sách thông qua anime_id. Thử với anime_id là 100.

```
anime_pivot.index[100]
```

Hình 96: Lệnh gọi tên phim hoạt hình

```
'40 Shuunen Da yo!! CoroCoro All-Star Shougakkou'
```

Hình 97: Kết quả gọi tên phim hoạt hình

```
anime_names = anime_pivot.index
```

Hình 98: Lệnh tạo biến tên phim hoạt hình

```
anime_names[100]
```

Hình 99: Lệnh gọi biến tên phim hoạt hình

```
'40 Shuunen Da yo!! CoroCoro All-Star Shougakkou'
```

Hình 100: Kết quả gọi biến tên phim hoạt hình

Gọi ngược lại từ title thì anime_id ở đâu.

```
np.where(anime_pivot.index == '40 Shuunen Da yo!! CoroCoro All-Star Shougakkou')[0][0]
```

Hình 101: Lệnh gọi anime_id bằng title

100

Hình 102: Kết quả anime_id trả về

```
ids = np.where(final_rating['title'] == '40 Shuunen Da yo!! CoroCoro All-Star Shougakkou')[0][0]
```

Hình 103: Lệnh tạo biến gọi anime_id bằng title

❖ Tìm ảnh phim hoạt hình

```
final_rating.iloc[ids]['image_path']
```

Hình 104: Lệnh gọi đường dẫn ảnh phim hoạt hình

```
12952339 https://cdn.myanimelist.net/images/anime/13/86...
12952340 https://cdn.myanimelist.net/images/anime/13/86...
12952341 https://cdn.myanimelist.net/images/anime/13/86...
12952342 https://cdn.myanimelist.net/images/anime/13/86...
12952343 https://cdn.myanimelist.net/images/anime/13/86...
...
12952404 https://cdn.myanimelist.net/images/anime/13/86...
12952405 https://cdn.myanimelist.net/images/anime/13/86...
12952406 https://cdn.myanimelist.net/images/anime/13/86...
12952407 https://cdn.myanimelist.net/images/anime/13/86...
12952408 https://cdn.myanimelist.net/images/anime/13/86...
Name: image_path, Length: 70, dtype: object
```

Hình 105: Kết quả đường dẫn ảnh phim hoạt hình

```
anime_name = []
for anime_id in suggestion:
    anime_name.append(anime_pivot.index[anime_id])
```

Hình 106: Lệnh đề xuất tên phim hoạt hình

```
ids_index = []
for name in anime_name[0]:
    ids = np.where(final_rating['title'] == name)[0][0]
    ids_index.append(ids)
```

Hình 107: Lệnh gọi anime_id bằng title

```
for idx in ids_index:
    url = final_rating.iloc[idx]['image_path']
    print(url)
```

Hình 108: Lệnh gọi đường dẫn ảnh phim hoạt hình

```
https://cdn.myanimelist.net/images/anime/6/23353.jpg
https://cdn.myanimelist.net/images/anime/1539/97315.jpg
https://cdn.myanimelist.net/images/anime/7/23159.jpg
https://cdn.myanimelist.net/images/anime/1555/90528.jpg
https://cdn.myanimelist.net/images/anime/10/24830.jpg
https://cdn.myanimelist.net/images/anime/7/13275.jpg
https://cdn.myanimelist.net/images/anime/7/55845.jpg
https://cdn.myanimelist.net/images/anime/13/13278.jpg
https://cdn.myanimelist.net/images/anime/6/26414.jpg
https://cdn.myanimelist.net/images/anime/3/31699.jpg
https://cdn.myanimelist.net/images/anime/11/25551.jpg
```

Hình 109: Kết quả gọi đường dẫn ảnh phim hoạt hình

❖ Hàm đề xuất

```
def recommend_anime(anime_name):
    anime_id = np.where(anime_pivot.index == anime_name)[0][0]
    distance, suggestion = model.kneighbors(anime_pivot.iloc[anime_id,:].values.reshape(1,-1), n_neighbors=11 )

    for i in range(len(suggestion)):
        animes = anime_pivot.index[suggestion[i]]
        for j in animes:
            if j == anime_name:
                print(f"You searched '{anime_name}'\n")
                print("The suggestion animes are: \n")
            else:
                print(j)
```

Hình 110: Lệnh tạo hàm đề xuất

1.5. Making Prediction (Đưa ra đề xuất)

❖ Đề xuất thử

```
recommend_anime("Aria the Avvenire")
```

Hình 111: Lệnh đề xuất cho “ Aria the Avvenire”

You searched 'Aria the Avvenire'

The suggestion animes are:

Aria the Origination: Sono Choppiri Himitsu no Basho ni...

Aria the OVA: Arietta

Aria the Natural: Sono Futatabi Deaeru Kiseki ni...

Aria the Origination Picture Drama

Tamayura: Sotsugyou Shashin Part 3 - Akogare

Tamayura: Sotsugyou Shashin Part 4 - Ashita

Tamayura: Sotsugyou Shashin Part 2 - Hibiki

Tamayura: Sotsugyou Shashin Part 1 - Kizashi

Hidamari Sketch x ☆☆☆ Recap

PriPara Movie: Mi~nna de Kagayake! Kirarin☆Star Live!

Hình 112: Kết quả đề xuất cho “Aria the Avvenire”

```
recommend_anime("Hanasaku Iroha")
```

Hình 113: Lệnh đề xuất cho “Hanasaku Iroha”

You searched 'Hanasaku Iroha'

The suggestion animes are:

Hanasaku Iroha: Home Sweet Home

Tari Tari

Tamako Market

Ano Natsu de Matteru

Denpa Onna to Seishun Otoko

Kamisama no Memochou

Ikoku Meiro no Croisée The Animation

A-Channel

So Ra No Wo To

Minami-ke

Hình 114: Kết quả đề xuất cho “Hanasaku Iroha”

```
recommend_anime("Minami-ke")
```

Hình 115: Lệnh đề xuất cho “Minami-ke”

```
You searched 'Minami-ke'

The suggestion animes are:

Minami-ke Okawari
Minami-ke Okaeri
Minami-ke Tadaima
Minami-ke Betsubara
Minami-ke Omatase
Minami-ke Natsuyasumi
Kannagi
Hyakko
Ichigo Mashimaro
Mitsudomoe
```

Hình 116: Kết quả đề xuất cho “Minami-ke”

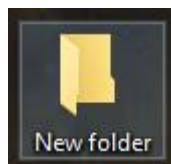
❖ **Xuất các tệp với định dạng .pkl**

```
pickle.dump(model,open('model.pkl','wb'))
pickle.dump(anime_names,open('anime_names.pkl','wb'))
pickle.dump(final_rating,open('final_rating.pkl','wb'))
pickle.dump(anime_pivot,open('anime_pivot.pkl','wb'))
```

Hình 117: Lệnh xuất mô hình

2. Design (Thiết kế)

Tạo một thư mục trống ngoài desktop đặt tên là “08032024”.

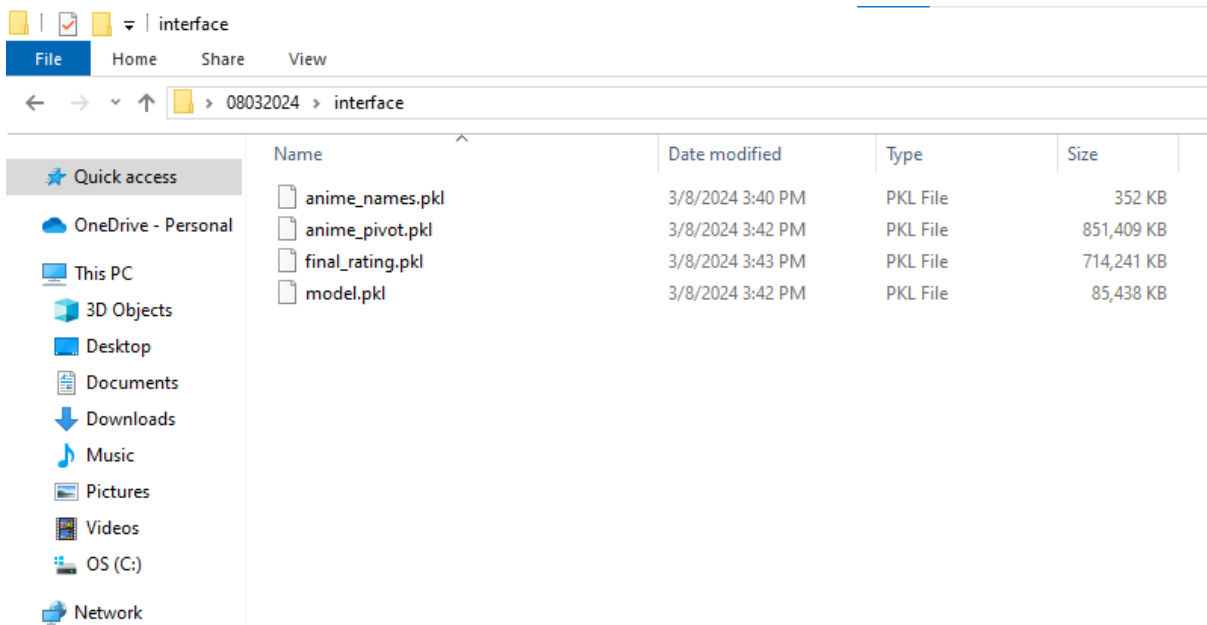


Hình 118: Tạo thư mục mới

Tiếp tục tạo một thư mục có tên là “interface” chứa 4 tệp .pkl vừa xuất.



Hình 119: Tạo thư mục “interface”



Hình 120: Mở thư mục “08032024”

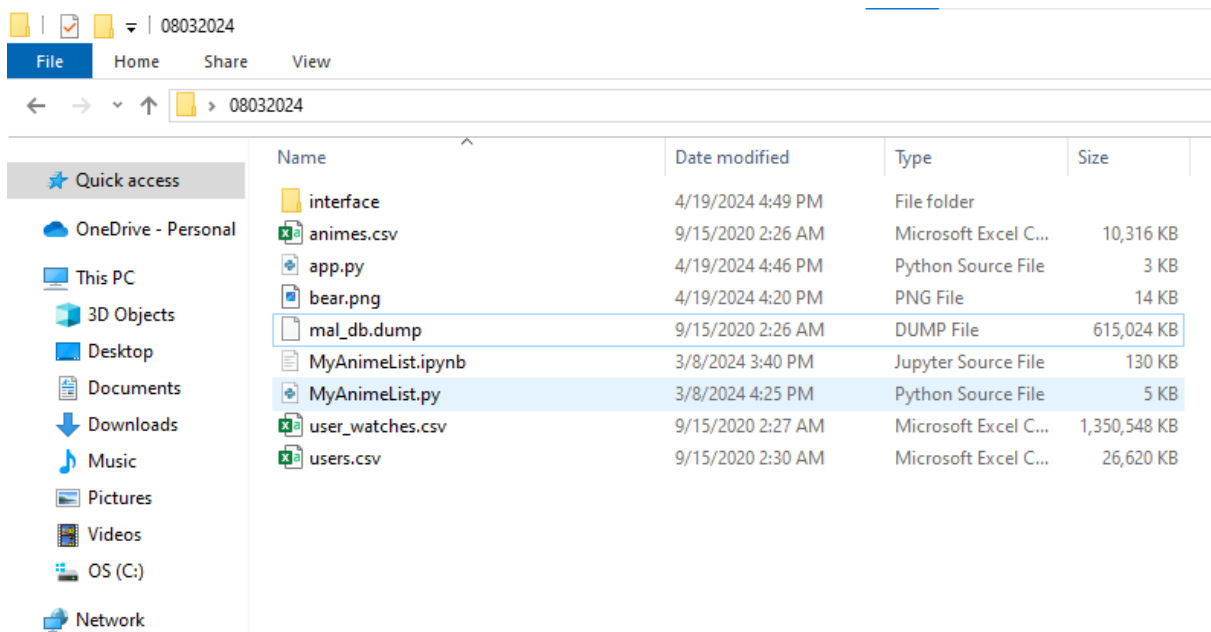
Tải tệp “MyAnimeList.ipynb” từ Jupyter Notebook vào thư mục “08032024”.

Tải tệp “MyAnimeList.py” từ Jupyter Notebook vào thư mục “08032024”.

Tải thêm 4 tệp của bộ dữ liệu MyAnimeList gồm “animes.csv”, “users.csv”, “user_watches.csv” và “mal_db.dump” vào thư mục “08032024”.

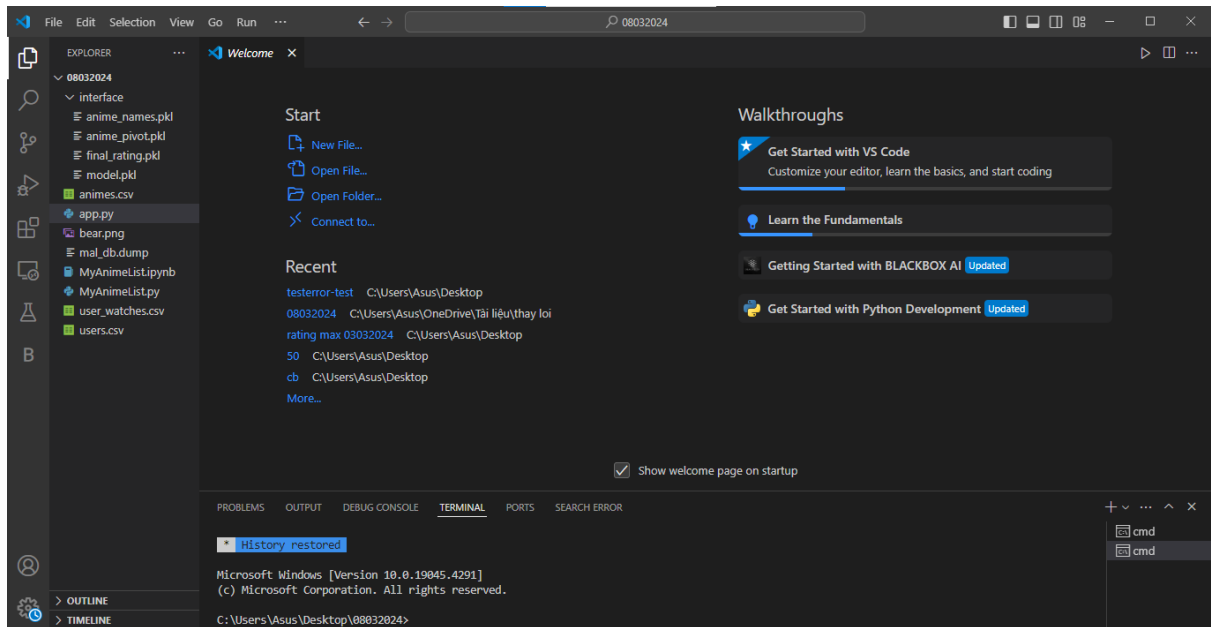
Tạo thêm một tệp “app.py” và một hình ảnh để là chèn trang trí.

Kết quả:



Hình 121: Mở thư mục “08032024”

Mở phần mềm Visual Studio Code. Mở thư mục “08032024”.

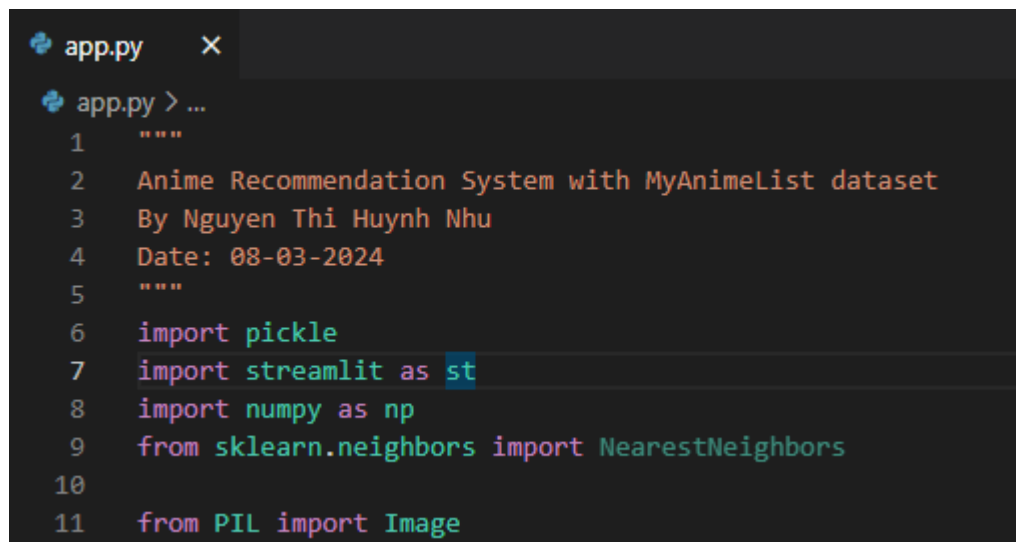


Hình 122: Mở phần mềm Visual Studio Code

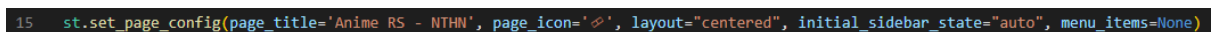
2.1. Web design (Thiết kế Web)

Mở tệp “app.py”

❖ Chuẩn bị thư viện



Hình 123: Mở tệp “app.py”



Hình 124: Lệnh thiết lập tiêu đề trang



Hình 125: Kết quả thiết lập tiêu đề

```
19 st.image(Image.open('C:/Users/Asus/Desktop/08032024/bear.png'))
```

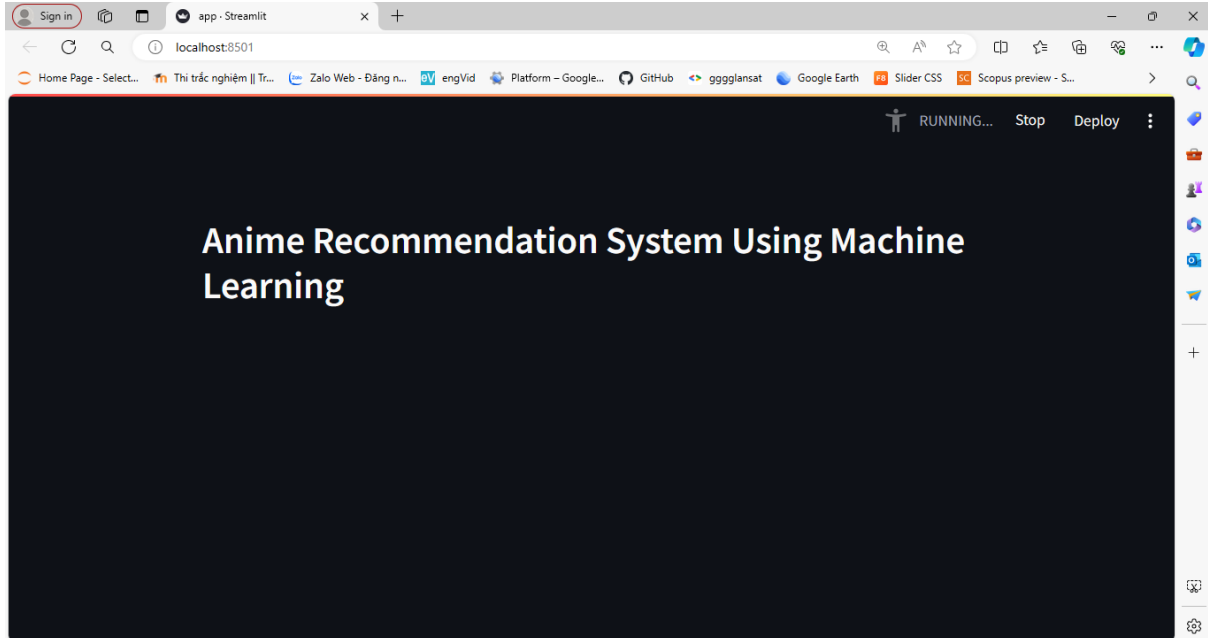
Hình 126: Lệnh chèn ảnh



Hình 127: Kết quả chèn ảnh

```
st.header('Anime Recommendation System Using Machine Learning')
model = pickle.load(open('interface/model.pkl','rb'))
anime_names = pickle.load(open('interface/anime_names.pkl','rb'))
final_rating = pickle.load(open('interface/final_rating.pkl','rb'))
anime_pivot = pickle.load(open('interface/anime_pivot.pkl','rb'))
```

Hình 128: Lệnh tạo tiêu đề web-app và gọi tệp chứa mô hình



Hình 129: Kết quả tiêu đề web-app

2.2. Function design (Thiết kế chức năng)

```

31 def fetch_poster(suggestion):
32     anime_name = []
33     ids_index = []
34     poster_url = []
35
36     for anime_id in suggestion:
37         anime_name.append(anime_pivot.index[anime_id])
38
39     for name in anime_name[0]:
40         ids = np.where(final_rating['title'] == name)[0][0]
41         ids_index.append(ids)
42
43     for idx in ids_index:
44         url = final_rating.iloc[idx]['image_path']
45         poster_url.append(url)
46
47     return poster_url

```

Hình 130: Lệnh đề xuất ảnh phim hoạt hình

```

54 def recommend_anime(anime_name):
55     animes_list = []
56     anime_id = np.where(anime_pivot.index == anime_name)[0][0]
57     distance, suggestion = model.kneighbors(anime_pivot.iloc[anime_id,:].values.reshape(1,-1), n_neighbors=11 )
58
59     poster_url = fetch_poster(suggestion)
60
61     for i in range(len(suggestion)):
62         animes = anime_pivot.index[suggestion[i]]
63         for j in animes:
64             animes_list.append(j)
65     return animes_list , poster_url

```

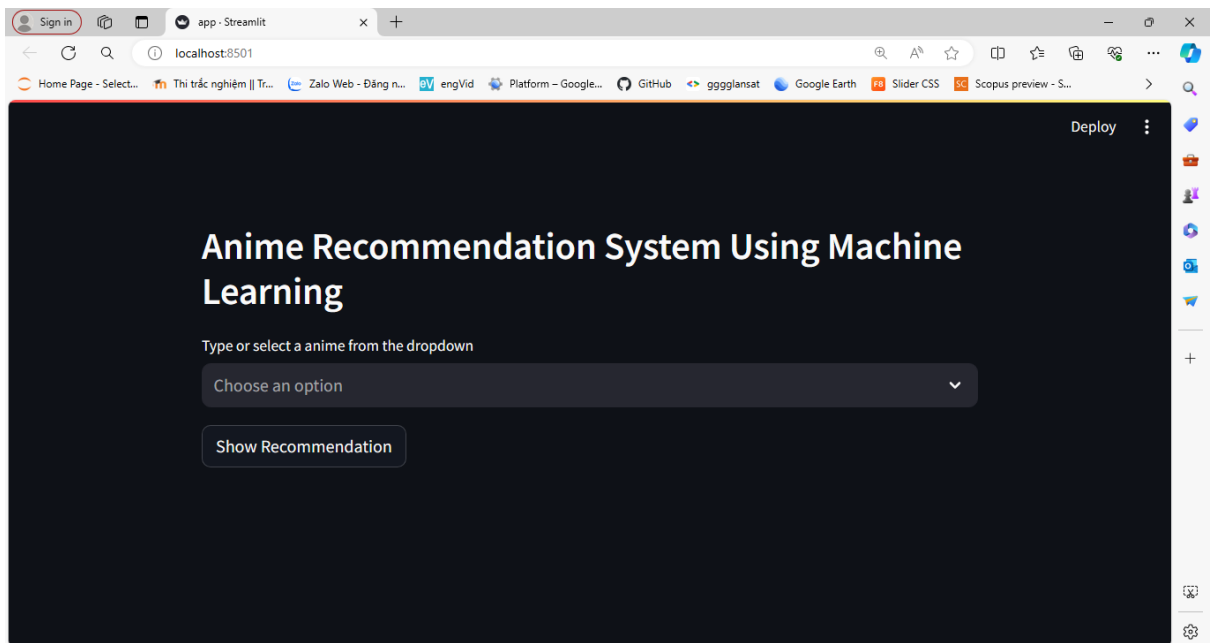
Hình 131: Lệnh đề xuất tên phim hoạt hình

```

69 selected_animes = st.selectbox(
70     "Type or select a anime from the dropdown",
71     anime_names
72 )

```

Hình 132: Lệnh thiết lập thanh tìm kiếm



Hình 133: Kết quả thành tìm kiếm

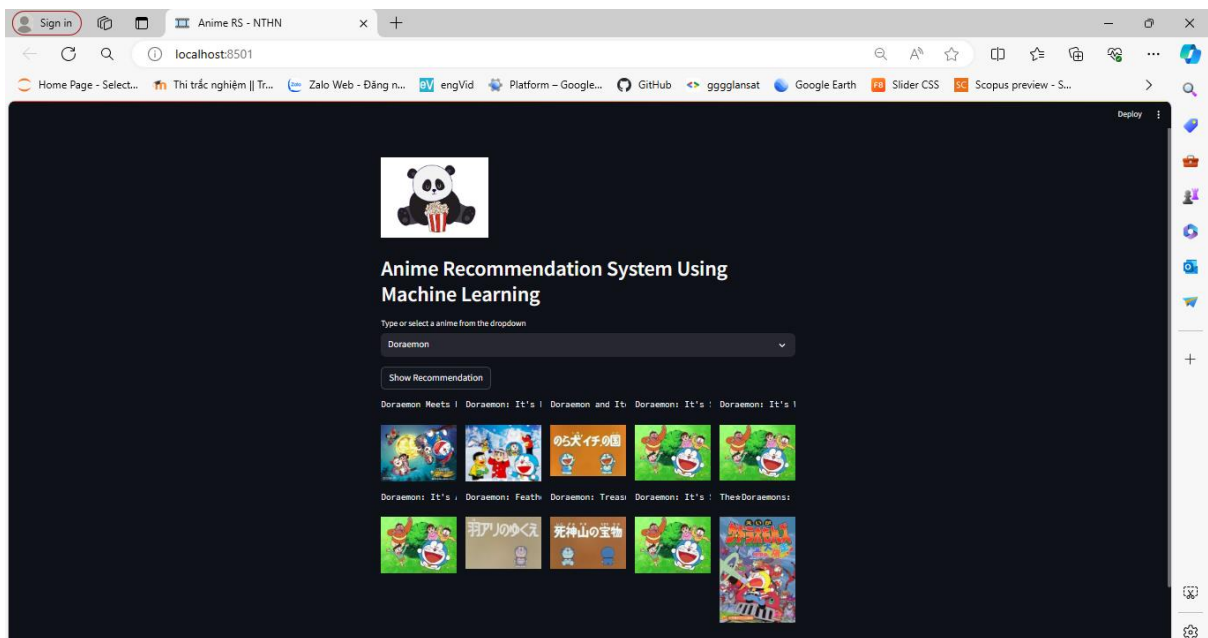
```
74 if st.button('Show Recommendation'):  
75     recommended_animes, poster_url = recommend_anime(selected_animes)  
76     col1, col2, col3, col4, col5 = st.columns(5)  
77     col6, col7, col8, col9, col10 = st.columns(5)  
78     with col1:  
79         st.text(recommended_animes[1])  
80         st.image(posters_url[1])  
81     with col2:  
82         st.text(recommended_animes[2])  
83         st.image(posters_url[2])  
84  
85     with col3:  
86         st.text(recommended_animes[3])  
87         st.image(posters_url[3])  
88     with col4:  
89         st.text(recommended_animes[4])  
90         st.image(posters_url[4])  
91     with col5:  
92         st.text(recommended_animes[5])  
93         st.image(posters_url[5])
```

```

94
95     with col6:
96         st.text(recommended_animes[6])
97         st.image(posters_url[6])
98     with col7:
99         st.text(recommended_animes[7])
100        st.image(posters_url[7])
101
102    with col8:
103        st.text(recommended_animes[8])
104        st.image(posters_url[8])
105    with col9:
106        st.text(recommended_animes[9])
107        st.image(posters_url[9])
108    with col10:
109        st.text(recommended_animes[10])
110        st.image(posters_url[10])

```

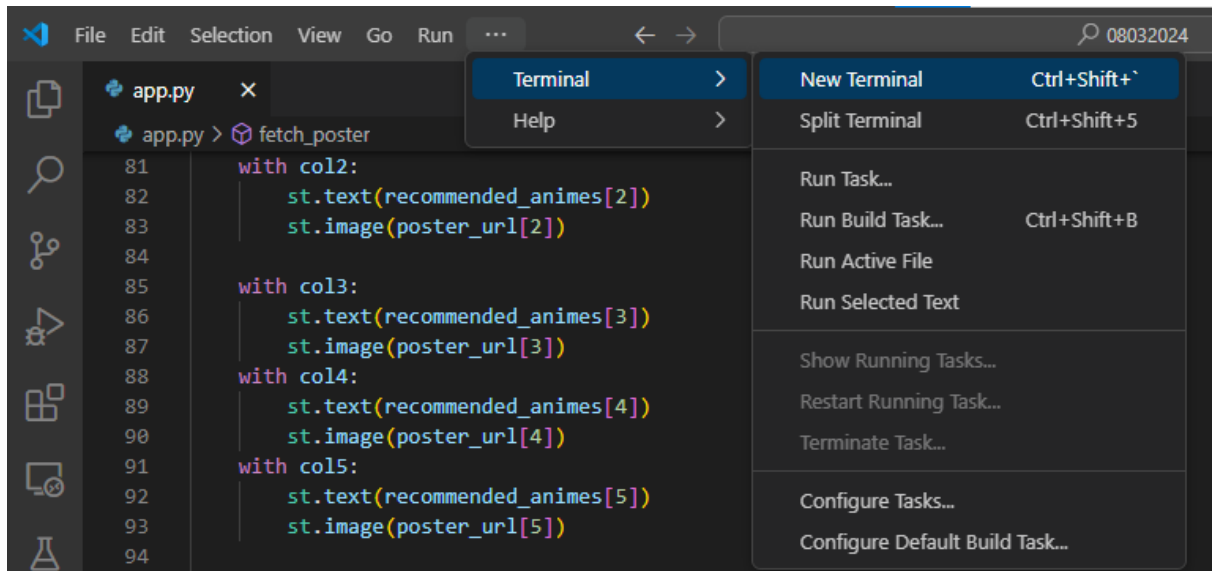
Hình 134: Lệnh thiết lập chức năng trình bày kết quả đề xuất



Hình 135: Giao diện đề xuất cho từ khóa “Doraemon”

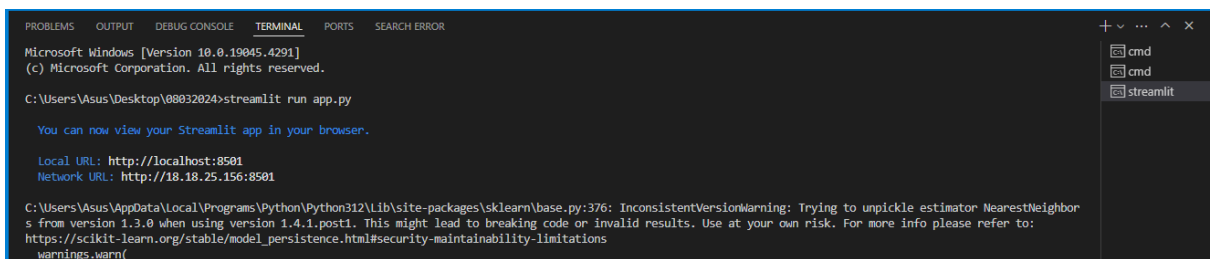
❖ Chạy “app.py”

Trên thanh công cụ tìm Terminal chọn New Terminal hoặc nhấn tổ hợp phím Ctrl+Shift+`.



Hình 136: Tìm cửa sổ lệnh Terminal

Tại cửa sổ lệnh “Terminal” gõ “streamlit run app.py” và nhấn phím “Enter”.



Hình 137: Lệnh chạy trên cửa sổ Terminal

By Nguyễn Thị Huỳnh Như

Date: 25/07/2024

Mail: 20166050@st.hcmuaf.edu.vn / nguyenthihuynhnhu271025@gmail.com