BÁO CÁO ĐỒ ÁN CUỐI KỲ

Manga Recommendation

Ngày 3 tháng 1 năm 2024

Nội dung

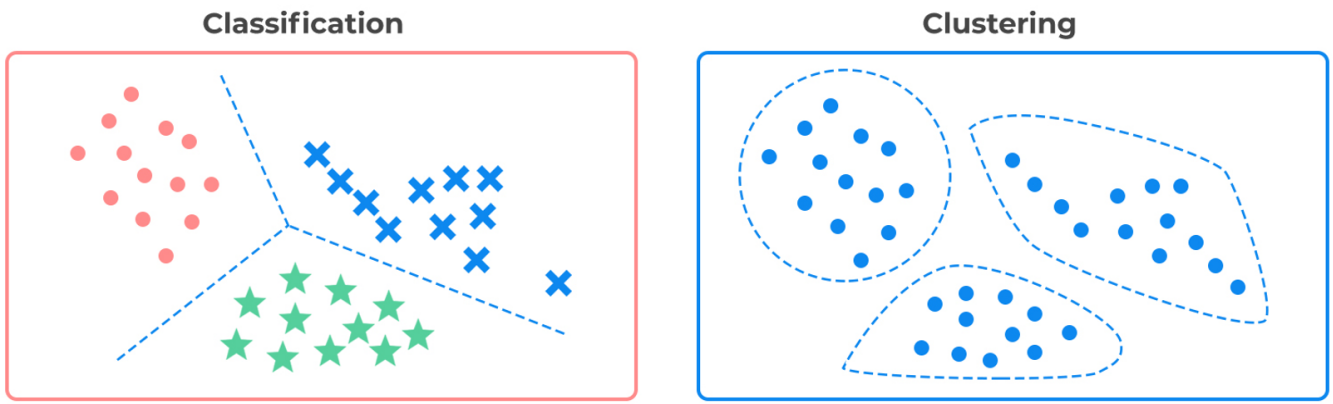
Bối cảnh

Hầu hết ứng dụng của học máy ngày nay dựa vào học có giám sát nơi mà những dữ liệu được gắn nhãn. Tuy nhiên phần lớn dữ liệu có sẵn lại thuộc loại không nhãn. Do đó, tiềm năng lớn nằm ở học không giám sát, nơi mà ta mới chỉ bắt đầu bước chân vào.

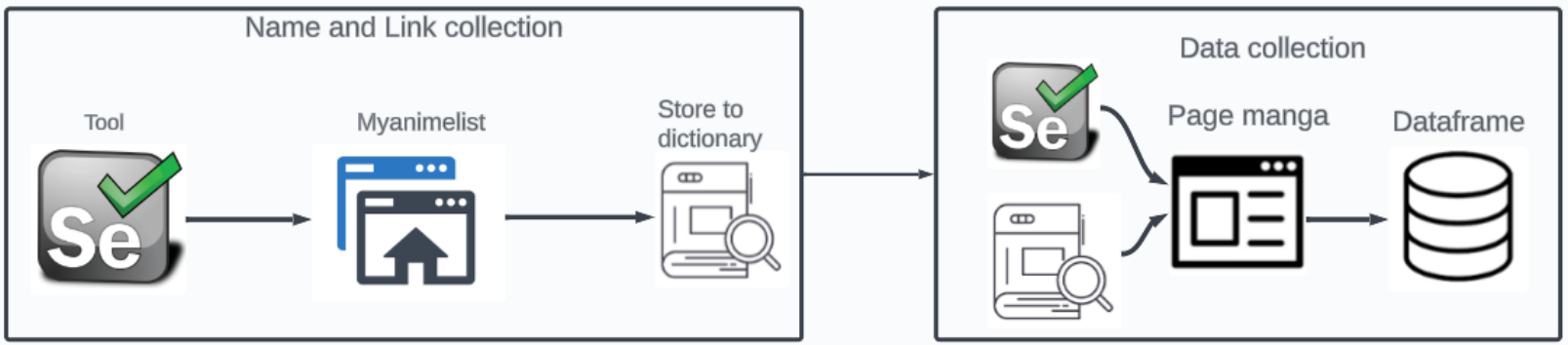
Ở đề tài này, ta sẽ đi tới 1 nhánh trong học không giám sát đó là gom cụm với dữ liệu là top 1000 manga được lấy từ trang web myanimelist nhằm gom nhóm những manga lại với nhau để có thể phát triển lên một hệ khuyến nghị dành cho manga.

Bối cảnh

Phân loại (học có giám sát) với Gom cụm (học không giám sát) :



Quá trình cào dữ liệu



Thuộc tính thu thập

Mỗi dữ liệu manga có các thuộc tính sau được thu thập:

- Ranked
- Score
- Popularity
- Favourite
- Published
- Genre
- Description
- Serialization
- Reading, Completed, Dropped, Plan to Read

Xem qua dữ liệu thu thập được

	name	ranked	score	popularity	favorite	published	genre	description	serialization	reading	completed	dropped	plan_to_read
0	Berserk	1	9.47	1	122,974	Aug 25, 1989	Action, Adventure, Award Winning, Drama, Fanta...	Guts, a former mercenary now known as the "Bla...	Young Animal	364,280	87,307	10,263	154,118
1	JoJo no Kimyou na Bouken Part 7: Steel Ball Run	2	9.30	26	42,935	Jan 19, 2004	Action, Adventure, Mystery, Supernatural	In the American Old West, the world's greatest...	Ultra Jump	34,922	162,358	2,141	50,652
2	Vagabond	3	9.24	15	40,221	Sep 3, 1998	Action, Adventure, Award Winning	In 16th-century Japan, Shinmen Takezou is a wi...	Morning	108,564	81,383	5,054	145,679
3	One Piece	4	9.22	3	114,647	Jul 22, 1997	Action, Adventure, Fantasy	Gol D. Roger, a man referred to as the "King o...	Shounen Jump (Weekly)	465,243	30,242	18,817	46,908
4	Monster	5	9.15	29	20,528	Dec 5, 1994	Award Winning, Drama, Mystery	Kenzou Tenma, a renowned Japanese neurosurgeon...	Big Comic Original	33,413	94,357	2,754	97,297

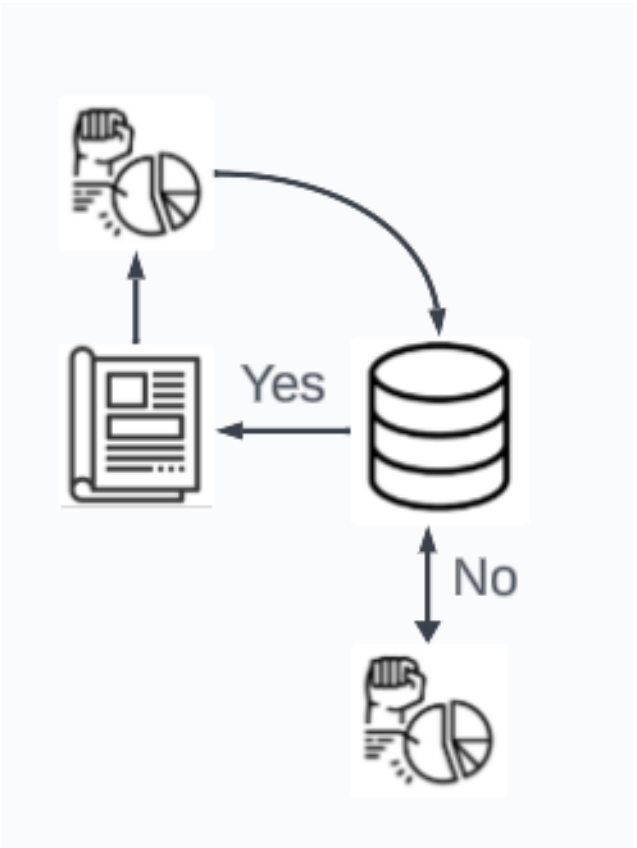
Xác định thuộc tính có giá trị khuyết

Các thuộc tính có giá trị khuyết:

- genre: thể loại
- description: mô tả
- serialization: tạp chí đăng dài kì

Thuộc tính Genre

Thường một tạp chí chuyên đăng một số thể loại cố định nên ta có thể làm theo cách sau:



Thuộc tính description và serialization

Ta điền mẫu câu: "No description information has been added to this title." Cho các description bị thiếu.
Từ: "Anonymous" cho các serialization bị thiếu.

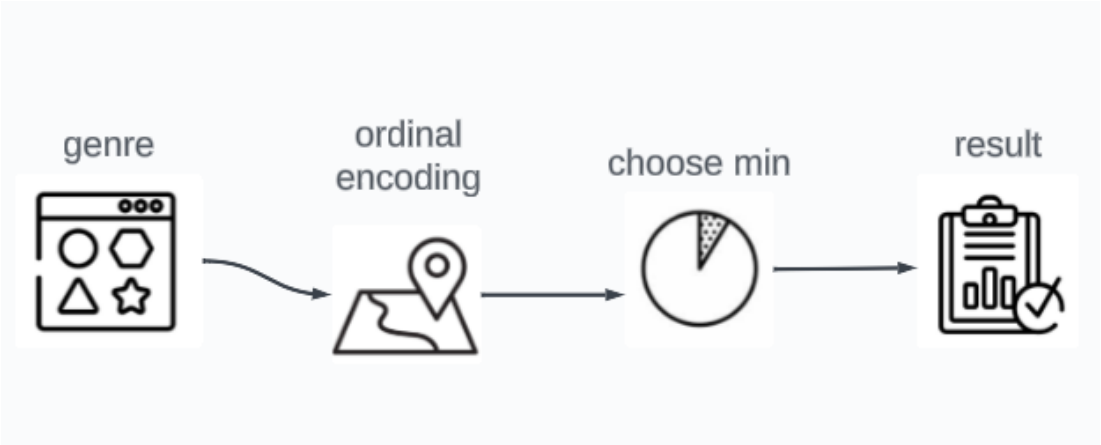
Xử lý dữ liệu numerical

Chuyển sang dữ liệu số với các dữ liệu như favourite, dropped, completed, ...

Chuyển sang dữ liệu datatime với thuộc tính published.

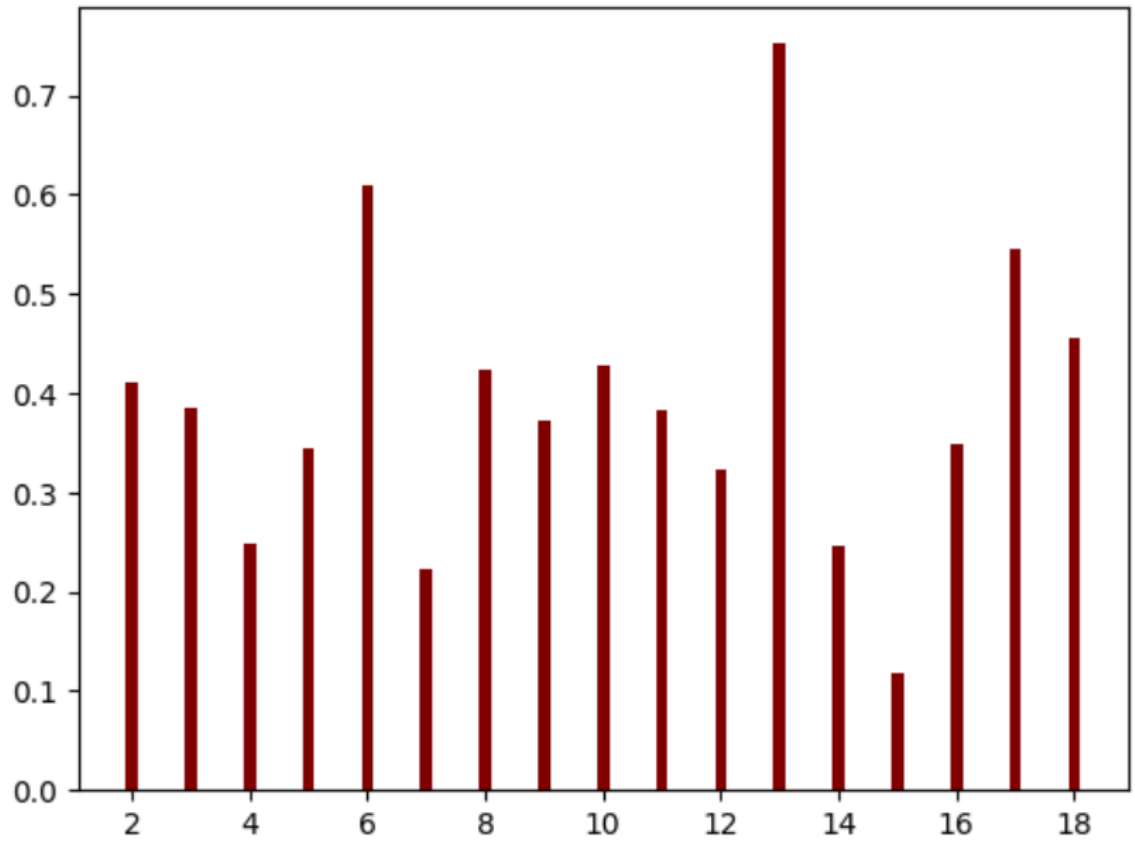
Xử lý thuộc tính categorical

- Xử lý thuộc tính genre



- Xử lý thuộc tính serialization: Dùng OrdinalEncoder của thư viện sklearn.
- Xử lý thuộc tính description: dùng thư viện nltk để xử lý văn bản (chuẩn bị để chuyển sang vector nếu cần).

Mối liên hệ thể loại và tỉ lệ dropped/completed

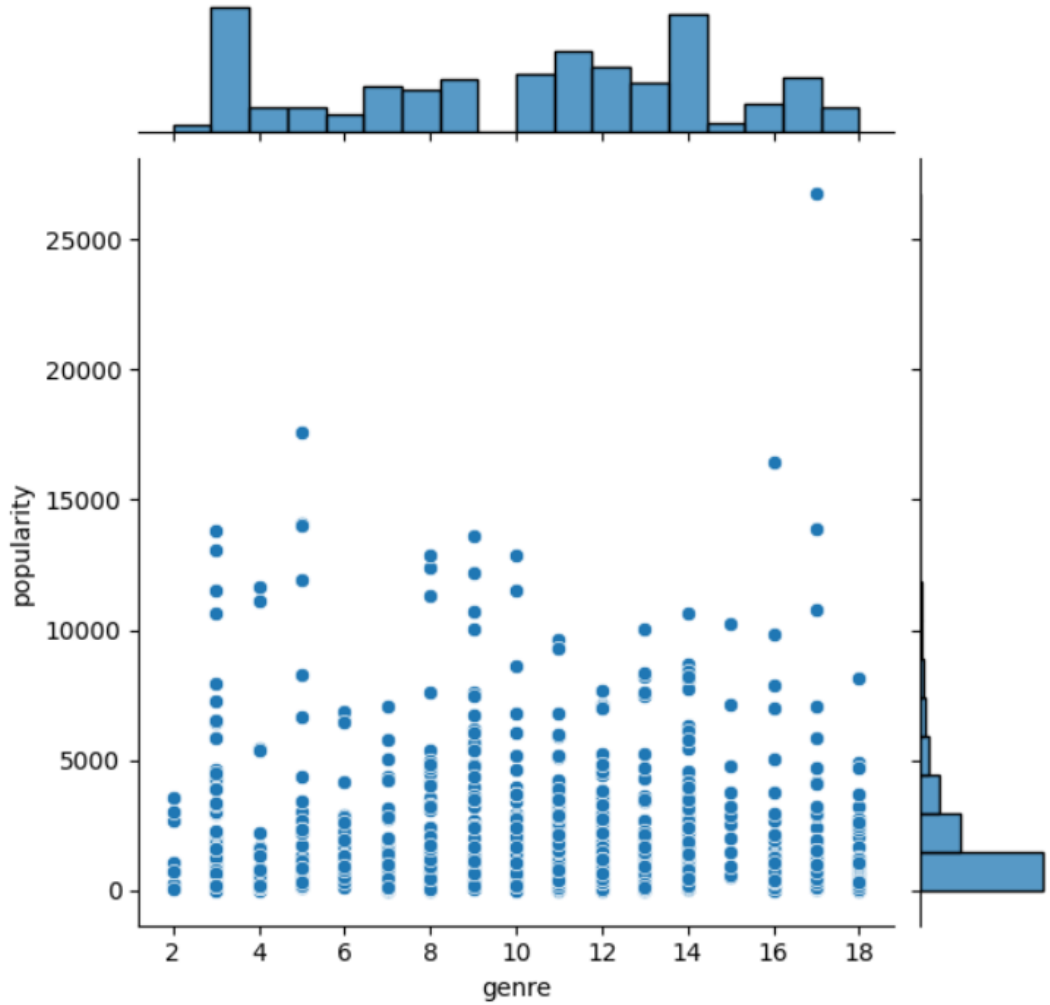


Liệu có thể dự đoán score dựa vào favorite, reading, completed, dropped, plan to read columns

```
X = df2[['favorite', 'reading', 'completed', 'dropped', 'plan_to_read']]
y = df2['score']

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create a linear regression model
model = LinearRegression()

# Fit the model with the training data
model.fit(X_train, y_train)

# Make predictions using the testing set
y_pred = model.predict(X_test)

# The coefficients
print('Coefficients: \n', model.coef_)

# The mean squared error
print('Mean squared error: %.2f' % mean_squared_error(y_test, y_pred))
```

```
Coefficients:
 [ 8.15322622e-06  1.38296727e-07 -6.00046518e-07 -3.41220345e-05
  1.09019258e-05]
Mean squared error: 0.04
```

Mối liên hệ giữa thể loại và phổ biến



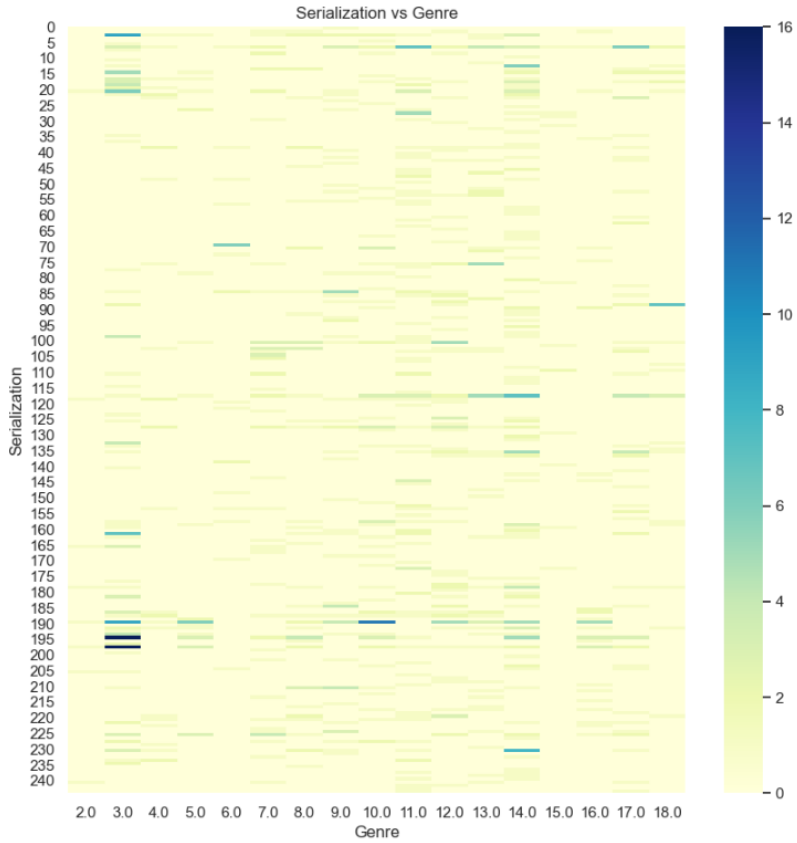
Liệu có thể xác định truyện nào có tiềm năng thăng hạng và bộ nào có tiềm năng rớt hạng?

Cách làm: Trên cùng một score, dựa vào thông số reading để xác định tiềm năng.

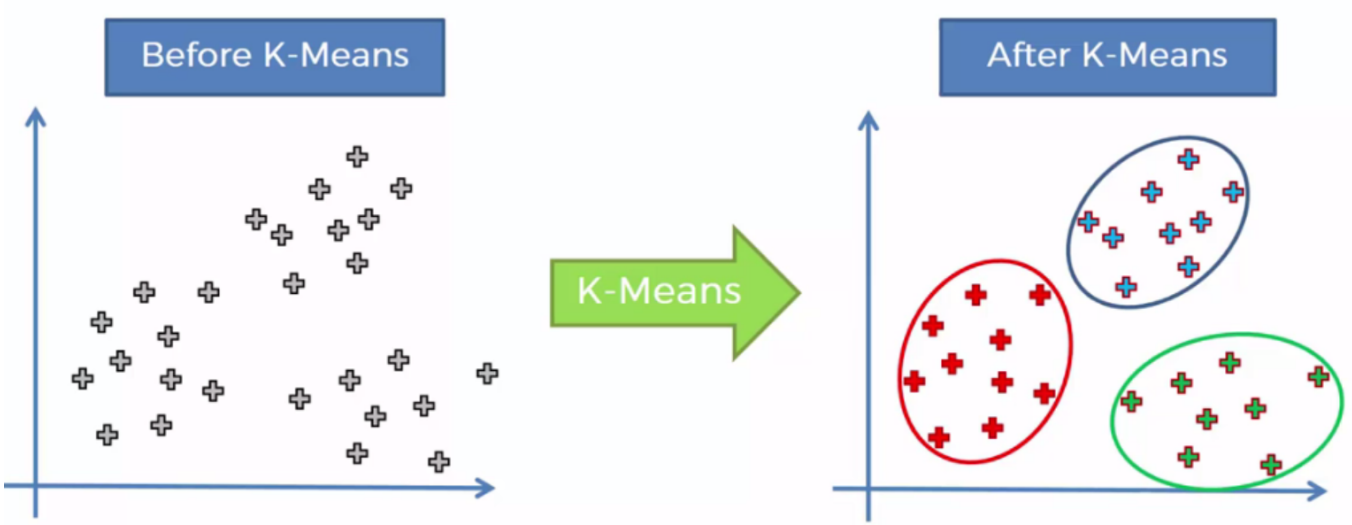
```
df4=data.copy()
df4.columns
df4=df4[['name','score','reading']]
df4_1=df4.groupby('score').max().reset_index() #trên cùng score bộ đang có nhiều người đang đọc nhất

df4_2=df4.groupby('score').min().reset_index() #trên cùng score bộ đang có ít người đang đọc nhất

df4_ins= df4_1[df4_1['name']!=df4_2['name']]
print(df4_ins) #bộ có tiềm năng thăng hạng
df4_dec= df4_2[df4_2['name']!=df4_1['name']]
print(df4_dec) #bộ có tiềm năng rớt hạng
```

Mối liên hệ giữa serialization thể loại

Cách làm: Trên cùng một score, dựa vào thông số reading để xác định tiềm năng.



K Means

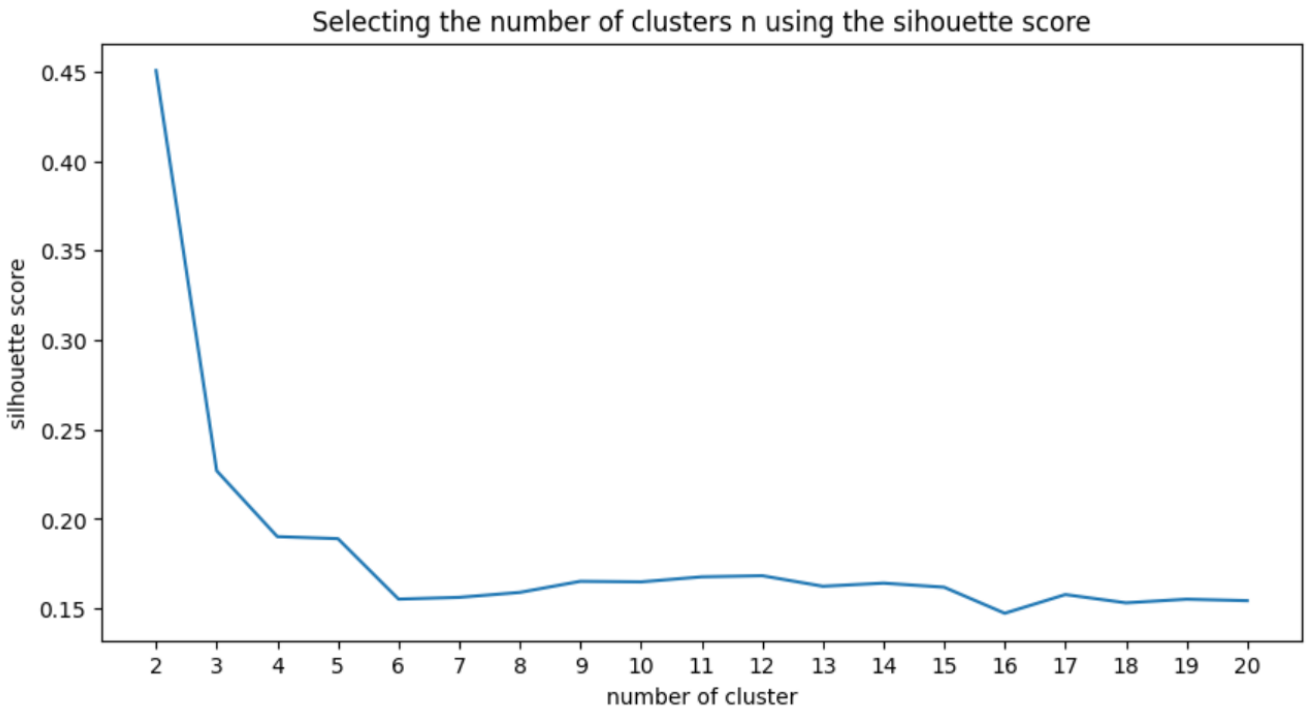


Xác định số cluster bằng silhouette score

silhouette score là mean của hệ số silhouette trên tất cả điểm, mỗi điểm có hệ số silhouette được tính bằng $(b - a)/max(a, b)$.

Trong đó a là khoảng cách trung bình của mỗi điểm tới điểm khác trong cùng 1 cluster và b là khoảng cách trung bình của điểm đó tới cluster gần nhất.
Do đó silhouette score có miền từ -1 đến +1, nếu càng gần 1 nghĩa là điểm được gom nhóm tốt trong cluster đó và xa với các cluster khác, trong khi đó càng gần 0 thì các điểm càng gần cluster boundary.

Xác định số cluster bằng silhouette score

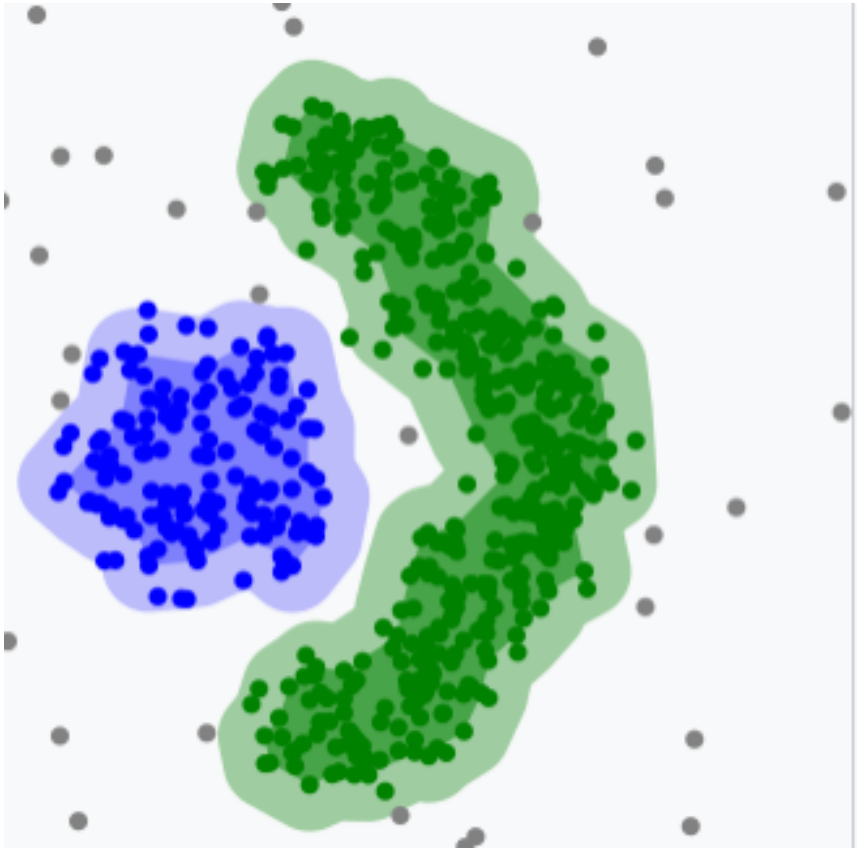


K Means

Ưu nhược điểm thuật toán:

- Ưu: Nhanh chóng và dễ dàng mở rộng.
- Nhược: Dễ gặp cực tiểu địa phương, phải chỉ rõ số cụm cần gom nhóm, không hiệu quả nếu dữ liệu nằm ở các nhóm có kích thước đa dạng,hay đa dạng trong tính tập trung dữ liệu và các nhóm không có tập trung theo dạng hình cầu

DBSCAN



Cách hoạt động

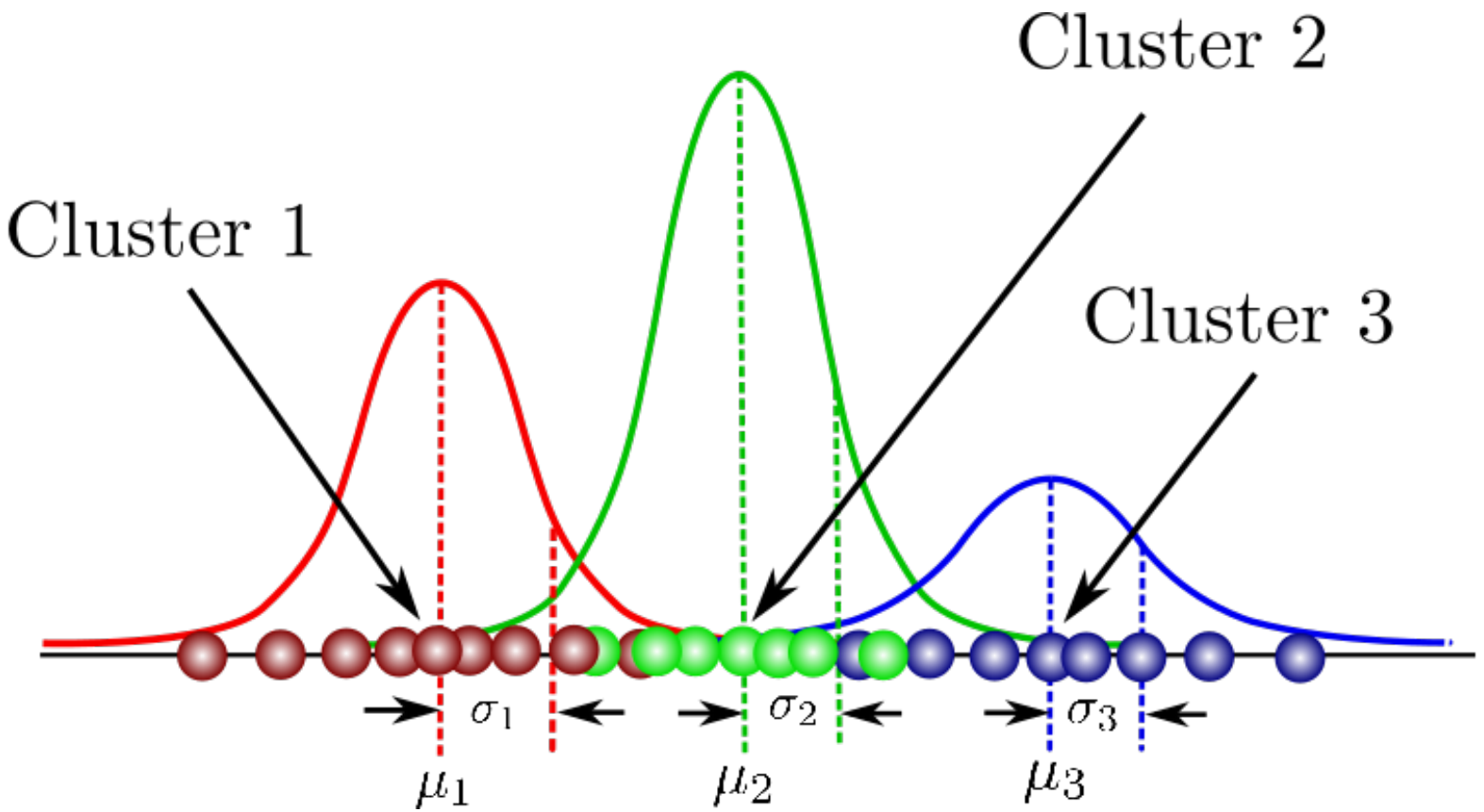
Cách tiếp cận của thuật toán cho phép xác định cluster của hình dáng bất kỳ (thay vì chỉ tốt ở hình cầu như K-means)

- Với khoảng cách cho trước, thuật toán đếm có bao nhiêu điểm trong cùng 1 khoảng cách ϵ từ điểm đó. Vùng này gọi là ϵ-neighborhood của điểm đó.
- Nếu điểm đó có ít nhất min samples điểm trong vùng ϵ-neighborhood (bao gồm bản thân), thì sẽ được gọi là điểm core. Tức điểm core là điểm nằm ở các vùng có mật độ cao.
- Tất cả điểm trong neighborhood của điểm core thuộc cùng 1 cluster.
- Tất cả điểm không là điểm core và không có 1 điểm nào thuộc neighborhood của nó thì xem là điểm bất thường.

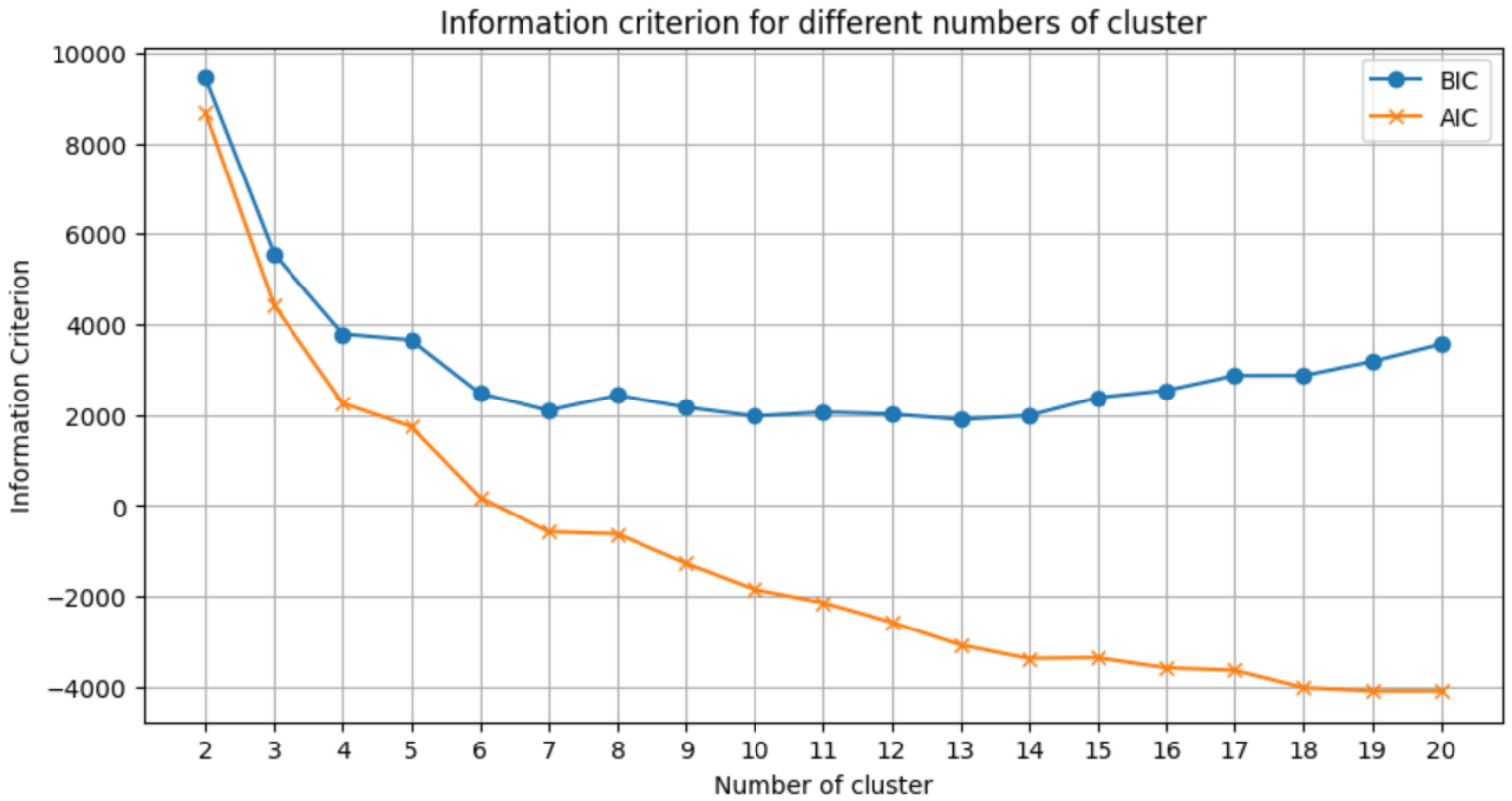
DBSCAN

Thuật toán Không thích hợp cho dữ liệu này vì dữ liệu chúng ta không tách biệt tốt bởi các vùng mật độ thấp. Nếu ϵ thấp thì đa số điểm là điểm dị thường còn nếu ϵ cao thì đa số dữ liệu chỉ tách biệt từ 1-2 cluster

Gaussian Mixtures



Tìm kiếm số cluster tối ưu dựa trên BIC và AIC



Gaussian mixtures

Ưu nhược điểm thuật toán:

- Ưu: Có thể giải quyết những cụm có đa dạng kích thước, mật độ, hướng hay hình dáng elip thay vì chỉ tốt ở hình cầu như k-means. Cho phép dữ liệu có xác suất thuộc cụm khác (soft clustering). Mạnh mẽ trước nhiễu.
- Nhược: Dễ gặp cực tiểu địa phương, phải chỉ rõ số cụm cần gom nhóm, không hiệu quả nếu dữ liệu thuộc loại categorical, chi phí tính toán cao và tốc độ thấp so với các thuật toán khác như k-means

Tổng kết

- Biết cách dùng trello, git để theo dõi và thực hiện đồ án.
- Học được các bước cần có từ việc thu thập đến xây dựng mô hình từ dữ liệu (cách cào dữ liệu, cách làm sạch dữ liệu, ...)
- Hiểu được cách hoạt động của các thuật toán gom cụm như k mean, gaussian mixture, dbscan.

Hạn chế

- Chưa ứng dụng mô hình bằng các công cụ triển khai như bentoML,...
- Dữ liệu chưa đủ lớn để có thể so sánh thời gian chênh lệch giữa các thuật toán.

Phương hướng phát triển

Triển khai ứng dụng hệ khuyến nghị manga dựa vào những manga người dùng đọc trước đó và thuật toán gom cụm.