

Nhập Môn Khoa Học Dữ Liệu

Đề tài: Phân loại đa nhãn review phim

GVHD: Thầy Nguyễn Ngọc Duy



Thành Viên - Nhóm 04



Lê Trung Nguyên

..... N21DCCN057➔



Hà Gia Huy

..... N21DCCN038➔



Trần Bình Phương Nhã

..... N21DCCN059➔



Nội Dung Bài Báo Cáo

- 01** Giới thiệu đề tài
- 02** Xây dựng bộ dữ liệu
- 03** Xử lý dữ liệu và trích xuất đặc trưng
- 04** Huấn luyện và đánh giá mô hình



Phần 01.

Giới thiệu đề tài

Tổng quan về đề tài



Nhiệm vụ phân loại đa nhãn review, đánh giá của người xem về phim nhằm mục đích xác định những trải nghiệm, suy nghĩ khác nhau của người xem về các bộ phim mà họ đã thưởng thức.

Các tiến bộ trong lĩnh vực NLP và công nghệ thông tin đã tạo điều kiện cho sự ra đời của các ứng dụng mới, đồng thời mở rộng không gian nghiên cứu dựa trên dữ liệu văn bản ngày càng phong phú. Phân loại văn bản là một trong những bài toán cơ bản của NLP.

Có nhiều cách phân loại văn bản trong NLP, nếu dựa trên số lượng nhãn được gán cho mỗi văn bản, có thể chia phân loại văn bản thành 2 loại là phân loại nhị phân (Binary) hay phân loại đa lớp (Multi-class) và phân loại đa nhãn (Multi-label).

Đối với ngữ cảnh cần thu thập ý kiến người dùng về một sản phẩm hay vấn đề, bài toán phân loại nhị phân hay đa lớp sẽ tỏ ra kém hiệu quả vì ngôn ngữ được dùng trong những trường hợp này phức tạp hơn, gán nhãn đơn lẻ sẽ làm hạn chế khả năng trích xuất thông tin.

Thuật toán học máy sử dụng trong đề tài

01

0.47

Naïve Bayes

NB là một thuật toán máy học được dựa trên định lý Bayes về giả định tính độc lập giữa các đặc trưng.

02

0.75

K-Nearest Neighbors

KNN là thuật toán đi tìm đầu ra của điểm dữ liệu bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất.

03

0.28

Support Vector Machine

SVM là một mô hình phân loại hoạt động bằng việc xây dựng một siêu phẳng trong không gian n chiều sao cho nó phân loại các lớp một cách tối ưu nhất.

04

0.73

Random Forests

RF tạo ra cây quyết định trên các mẫu được chọn ngẫu nhiên, tiến hành dự đoán riêng và chọn giải pháp tốt nhất bằng cách bỏ phiếu.

05

0.60

Logistic Regression

LR là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X.



Phần 02. Xây Dựng Bộ Dữ Liệu

Thu thập dữ liệu và chọn nhãn

Nhãn cảm động (touching)

Liên quan đến yếu tố cảm xúc “cảm động” mà bộ phim có thể mang đến cho người xem.
VD: “xem xong khóc muốn trôi rập”

Nhãn hài hước (comedy)

Liên quan đến yếu tố cảm xúc “hài hước” mà bộ phim có thể mang đến cho người xem. VD: “phim hài thiệt sự, coi mà cười đau cả bụng”

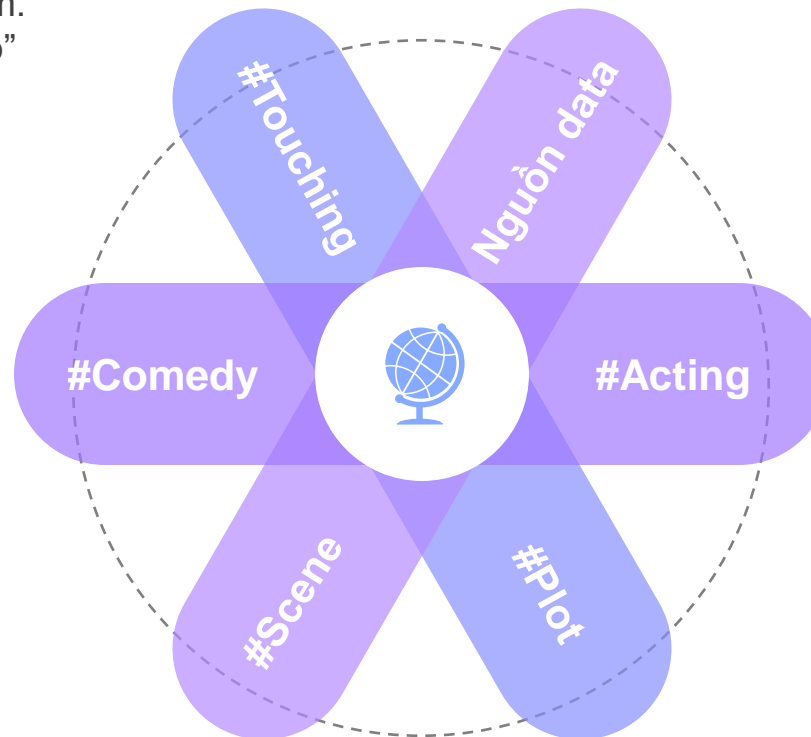
Nhãn góc quay, kỹ xảo (scene)

Liên quan đến yếu tố bối cảnh, góc quay, kỹ xảo được sử dụng trong phim. VD: “công nhận CGI đỉnh thật, quá choáng ngợp”

Trang review movie của MOMO

<https://www.momo.vn/cinema/review>

Sau khi thu thập và sà lọc, chúng tôi thu được bộ dataset có **5530** dòng dữ liệu và được phân loại dựa vào **5** nhãn



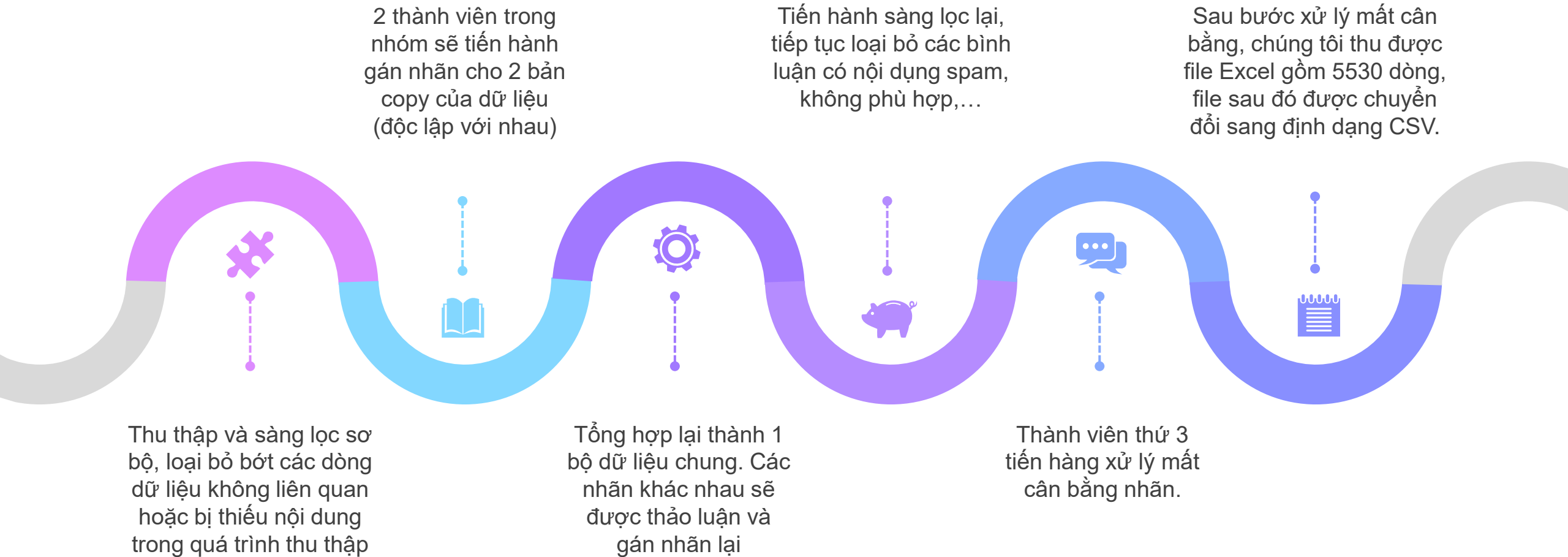
Nhãn diễn xuất (acting)

Liên quan đến yếu tố diễn xuất của các diễn viên. VD: “nhân vật Chú Ba diễn rất hay và nhập tâm”

Nhãn bối cảnh, cốt truyện (plot)

Liên quan đến yếu tố cốt truyện, kịch bản, nội dung phim. VD: “kịch bản có nhiều plot twist, làm người xem bất ngờ”

Quy trình đánh nhãn và tạo dataset



Trực quan hóa dữ liệu

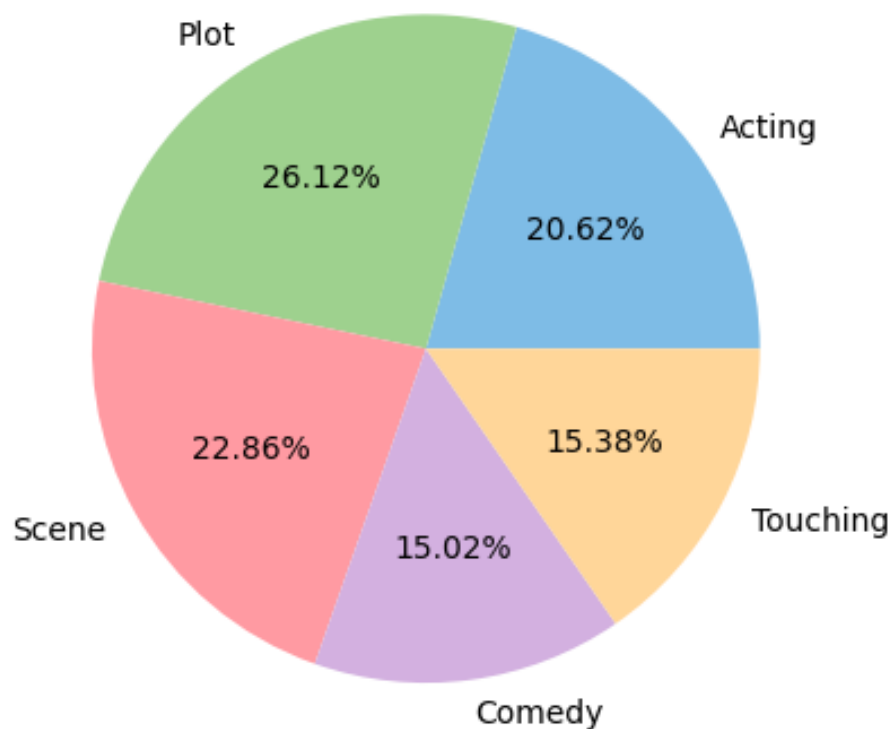
Nhận xét #plot

Nhãn cốt truyện (plot) có tần suất xuất hiện cao hơn, bởi cốt truyện, kịch bản là yếu tố then chốt quyết định chất lượng của hầu hết bộ phim.

Các nhãn mang yếu tố cảm xúc

Ngoài phụ thuộc vào thông điệp và khả năng truyền tải cảm xúc của bộ phim, các nhãn mang yếu tố cảm xúc còn phụ thuộc vào cảm nhận cá nhân của khán giả.

Tags Statistic



Nhận xét #comedy #touching

Các nhãn hài hước (comedy) và cảm động (touching) có tần suất xuất hiện thấp nhất, vì nó có liên quan mật thiết đến chủ đề, thông điệp và khả năng truyền tải cảm xúc của bộ phim nên không phải phim nào cũng được đề cập.

Sự liên quan #acting và #plot

Dựa vào dataset, các review có đề cập yếu tố diễn xuất (acting) hoặc yếu tố bối cảnh (scene) thường cũng nhắc đến yếu tố cốt truyện (plot).

Phần 03.

Xử lý dữ liệu và trích xuất đặc trưng



Biến đổi dữ liệu thô

Xóa các ký tự đặc biệt,
biểu tượng cảm xúc và xóa
các ký tự không cần thiết

Xóa khoảng
trắng thừa

Đưa toàn bộ dữ liệu
về chữ viết thường

Xóa các ký tự
cố ý viết dài

Chuẩn hóa Unicode
(chuyển chuỗi từ dạng
mã hóa Window-1252
sang Unicode utf-8)

Chuẩn hóa dấu câu
cho đúng vị trí

Thay các từ teen
code thành dạng
tiêu chuẩn

Tách từ - word
tokenization. Chia câu
review thành các từ

Bỏ các từ stop word
không mang nhiều ý nghĩa
trong quá trình training



Trích xuất đặc trưng

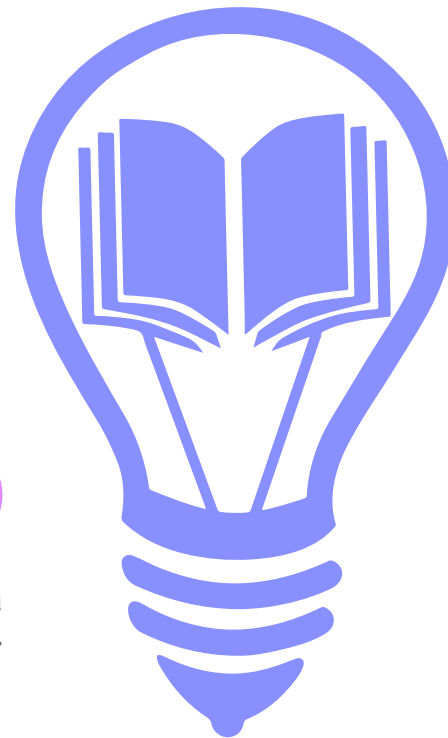
Công thức tổng quát

- TF (term frequency): Tính toán số lần xuất hiện của một từ trong văn bản.
- IDF (inverse document frequency): Ước lượng độ quan trọng của từ đó trong văn bản.
- TF-IDF được tính như sau:

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

Nhận xét

TF-IDF hoạt động tốt với các tập tài liệu nhỏ và vừa mà không yêu cầu nhiều tài nguyên tính toán. TF-IDF chỉ dựa trên tần suất từ xuất hiện mà không hiểu được ngữ nghĩa của từ trong ngữ cảnh.



Phương pháp TF-IDF

TF-IDF (Term frequency - inverse document frequency) là một trong những kỹ thuật cơ bản trong xử lý ngôn ngữ giúp đánh giá mức độ quan trọng của một từ trong văn bản.

Đầu ra

Đối với các từ có điểm cao, đồng nghĩa với mức độ đặc biệt và hiếm gặp trên bộ văn bản. Trái lại, nếu có giá trị thấp thì các từ đó có thể là các từ xuất hiện đại trà trong toàn bộ tập dữ liệu.

Phần 04.

Huấn luyện và đánh giá mô hình

Phương pháp training model



Problem Transformation

Chia bài toán phân loại đa nhãn thành các bài toán phân loại đơn nhãn.



Adapted Algorithm

Xử lý bài toán đa nhãn một cách trực tiếp mà không thông qua biến đổi thành các bài toán phụ.



Ensemble Approaches

Kết hợp nhiều classifier hoặc model để cải thiện hiệu suất phân loại đa nhãn.

**Phương pháp giải quyết
bài toán phân loại đa nhãn**

Phương pháp được áp dụng trong đề tài



Binary Relevance

Mỗi nhãn sẽ được xem là một bài toán phân loại đơn nhãn riêng biệt.



Label Powerset

Xem xét các trường hợp có các nhãn là giống nhau mà gom chúng lại thành một lớp.



Classifier Chains

Xây dựng một chuỗi các classifier trong đó mỗi classifier sử dụng các dự đoán của các classifier trước đó làm các feature bổ sung.



Các chỉ số đánh giá (Evaluation Metrics)

Macro-average precision score

Macro-average precision là trung bình cộng của các precision theo từng class.

$$MacroP = \frac{\sum_{c_i \in C} P(D, c_i)}{|C|}.$$

Macro-average recall

Macro-average recall là trung bình cộng của các recall theo từng class.

$$MacroR = \frac{\sum_{c_i \in C} R(D, c_i)}{|C|}.$$

Macro-average F1 score

Macro-average F1 score là giá trị trung bình điều hòa của macro P và macro R, được tính bằng cách tính F1 score cho từng class riêng lẻ sau đó tính giá trị trung bình.

$$MacroF1 = \frac{1}{N} \sum_{i=0}^N F1.$$

Micro-average precision score

Micro-average precision tổng hợp sự đóng góp của tất cả các class để tính toán precision tổng thể.

$$MicroP = \frac{\sum_{c_i \in C} TPs(c_i)}{\sum_{c_i \in C} TPs(c_i) + FPs(c_i)}.$$

Micro-average recall score

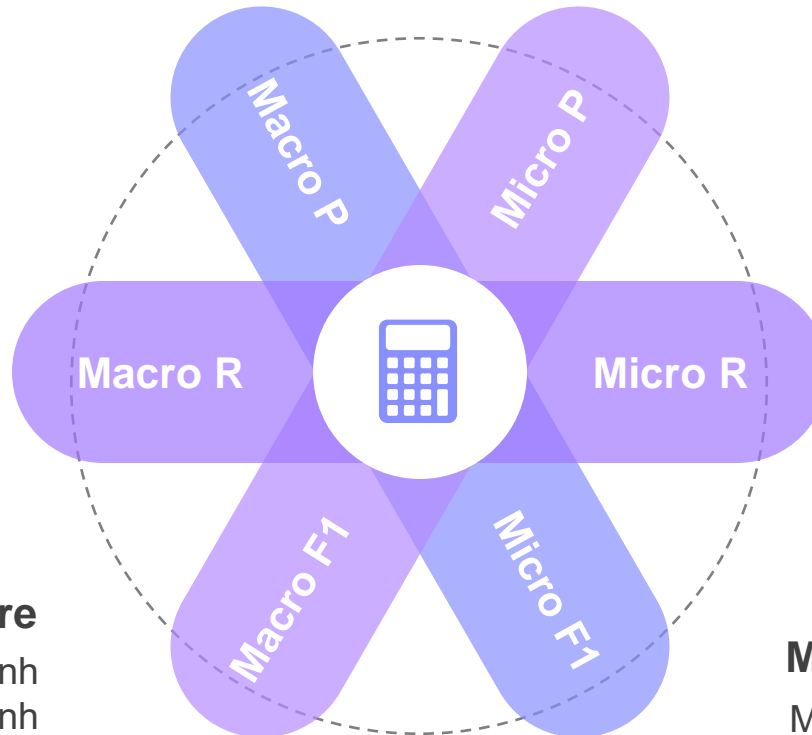
Micro-average recall tổng hợp sự đóng góp của tất cả các class để tính toán recall tổng thể.

$$MicroR = \frac{\sum_{c_i \in C} TPs(c_i)}{\sum_{c_i \in C} TPs(c_i) + FNs(c_i)}.$$

Micro-average F1 score

Micro-average F1 score là giá trị trung bình điều hòa của micro P và micro R. Được tổng hợp từ toàn bộ các nhãn trên phạm vi toàn cục.

$$MicroF1 = 2 \cdot \frac{MicroP \cdot MicroR}{MicroP + MicroR}.$$



Kết quả training và kết luận

Method	Classifier	Micro F1	Macro F1	Micro P	Macro P	Micro R	Macro R
Binary Relevance	NB	0.47084	0.45532	0.86291	0.88371	0.32420	0.31383
	RF	0.74712	0.75481	0.86619	0.86918	0.65707	0.66927
	KNN	0.27737	0.29294	0.60750	0.61529	0.17972	0.19586
	SVC	0.73163	0.73565	0.87898	0.88799	0.62671	0.62903
	LR	0.59853	0.58945	0.90414	0.91333	0.44759	0.43862
Label Powerset	NB	0.36228	0.35249	0.85716	0.87853	0.22809	0.22355
	RF	0.61889	0.62140	0.87332	0.87742	0.47933	0.48277
	KNN	0.27820	0.29374	0.60629	0.61558	0.18054	0.19645
	SVC	0.59445	0.59478	0.86718	0.88195	0.45250	0.45111
	LR	0.49325	0.49061	0.89388	0.90780	0.34094	0.33835
Classifier Chains	NB	0.47688	0.46384	0.86783	0.88586	0.32928	0.32050
	RF	0.74509	0.75257	0.86956	0.87253	0.65195	0.66384
	KNN	0.27253	0.29290	0.60642	0.61465	0.17972	0.19586
	SVC	0.73081	0.73481	0.88006	0.88873	0.62501	0.62725
	LR	0.57750	0.56709	0.90007	0.90834	0.42537	0.44645



Phương pháp cho kết quả tốt nhất (dựa theo Micro F1 score) là: *Binary Relevance*



Thuật toán học máy tương ứng cho kết quả tốt nhất là: *Random Forest*



Giá trị Accuracy Score thu được từ mô hình này là: *0.7242314647377939*



Giá trị Hamming Loss thu được từ mô hình này là: *0.06672694394213381*



Cảm ơn thầy và các bạn đã lắng nghe

