TRƯỜNG ĐẠI HỌC NHA TRANG

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO NHÓM**

**THUẬT TOÁN NAIVE BAYES VÀ BÀI TOÁN   
PHÂN LOẠI BÀI HÁT LYRICS CLASSIFICATION**

**Giảng viên hướng dẫn: TS. Đinh Đồng Lưỡng**

**Học viên thực hiện: 1. Nguyễn Trọng Kiên – 61CH092**

**2. Cao Viết Thắng – 61CH097**

**Lớp quản lý: CHCNTT2019-1**

Khánh Hòa, tháng 08/2020

**Tóm tắt**: Phân loại văn bản là một trong những bài toán machine learning có giám sát, và Naive Bayes là một thuật toán phổ biến được sử dụng trong bài toán này. Mục tiêu nghiên cứu này là dự đoán thể loại của bài hát chỉ dựa trên lời bài hát, sử dụng Bag-of-words làm phương pháp trích chọn đặc trưng và thuật toán Naive Bayes để phân loại và dự đoán. Tập dữ liệu training bao gồm 3 nhãn: Nhạc Trẻ, Nhạc Cách mạng, Nhạc Trinh, mỗi nhãn có 100 lời bài hát.

***Keywords: Naive Bayes classifier, lyrics classification, text classification, bag-of-words, machine learning.***

# GIỚI THIỆU VÀ CÁC NGHIÊN CỨU LIÊN QUAN

Phân loại văn bản là một bài toán quan trọng và phổ biến trong học máy có giám sát (supervised learning). Một số ứng dụng của nó như: phát hiện email spam, phân tích cảm xúc, phát hiện ngôn ngữ của văn bản viết tay, phân lớp, v.v. Một vài thuật toán như: Mạng neural (ANN), Support Vector Machines (SVM), thuật toán di truyền, Phân lớp Naive Bayes, K-Láng giềng (KNN) và phân lớp Rocchio [1].

Số lượng các bài hát, đặc biệt là trên internet, đang phát triển một cách nhanh chóng và việc tổ chức quản lý nó là một việc khá khó. Với một bộ sưu tập nhạc khổng lồ, việc phân loại nhạc nên được thực hiện một cách tự động. Phân loại có thể dựa trên các tiêu chí như: thể loại, cảm xúc, người biểu diễn, địa lý, khu vực, v.v.

Để có thể phân loại thành công, người ta có thể dựa vào một số đặc điểm của âm thanh như là: nhịp độ (tempo), nhịp điệu (rhythm), âm sắc (timbre), cao độ (pitch), độ lớn (loudness) hoặc đặc điểm của lời bài hát như là: từ và độ dài của câu, tần số của từ, câu và cấu trúc đoạn, từ đồng nghĩa, từ đồng âm v.v. Theo [2], vấn đề cần làm nhất trong phân loại nhạc tự động theo tâm trang dựa trên đặc điểm của âm thanh (cao độ và nhịp điệu của âm thanh là phổ biến nhất).

* **Các nghiên cứu liên quan:**

Trong [3] có 4 thể loại nhạc phổ biến (cổ điển, jazz, rock và pop) đã đươc chọn để làm nền tảng phân loại âm thanh dựa trên Hệ số bề mặt tần số Mel (Mel Frequency Cepstral Coefficients). Độ chính xác khi dùng Direct Acyclic Graph SVM rơi vào khoảng từ 67% tới 97%. Khi dùng Mạng Nơ-ron (Neural Networks), kết quả đạt được vào khoảng 76% đến 100 % tùy theo thể loại.

Tự động nhận dạng các nghệ sĩ biểu diễn trong các buổi hòa nhạc dương cầm của cùng một bản nhạc là một nghiên cứu thú vị được mô tả trong [4]. Nghệ sĩ dương cầm đã thể hiện 2 tác phẩm của Frederick Chopin. Với kết quả cao với độ chính xác khoảng 70% trong 10-class task.

Fell và Sporleder trong mục [5] đã nghiên cứu về tự động dự đoán thời gian xuất bản gần đúng của một bài hát với lời bài hát. Họ chọn các bài hát pop/rock và chia chúng thành 3 giai đoạn: “2008 và mới hơn”, “từ 1998 tới 2001”, và “phát hành trước 1988”. Kết quả cho thấy những bài hát được phát hành hơn 20 năm trước có thể phân biệt tương đối tốt, nhưng với các bài hát mới hơn kết quả phân loại tương đối thấp.

=> Tùy vào các kiểu phân lớp, cách kết hợp âm thanh và thông tin từ lời bài hát là một cách phổ biến để phân loại nhạc.

Mục tiêu của nghiên cứu này để kiểm tra xem thuật toán phân lớp phân lớp Naive Bayes có thể dự đoán thành công thể loại nhạc dựa trên lời bài hát hay không. Tập dữ liệu training bao gồm 3 nhãn: Nhạc Trẻ, Nhạc Cách mạng, Nhạc Trinh, mỗi nhãn có 100 lời bài hát

# THUẬT TOÁN NAIVE BAYES

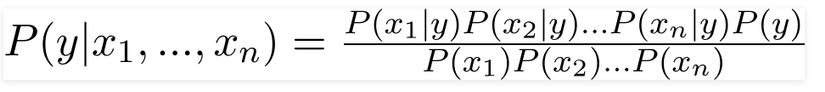
Naive Bayes là một thuật toán phân lớp được mô hình hoá dựa trên định lý Bayes trong xác suất thống kê:

*Trong đó:*

* P(y|X) gọi là posterior probability: xác suất của mục tiêu y với điều kiện có đặc trưng X
* P(X|y) gọi là likelihood: xác suất của đặc trưng X khi đã biết mục tiêu y
* P(y) gọi là prior probability của mục tiêu y
* P(X) gọi là prior probability của đặc trưng X

Ở đây, **X** là vector các đặc trưng, có thể viết dưới dạng:

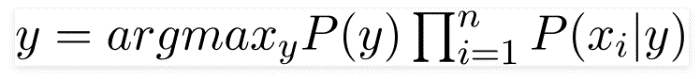
Khi đó, đẳng thức Bayes trở thành:



Trong mô hình Naive Bayes, có hai giả thiết được đặt ra:

* Các đặc trưng đưa vào mô hình là độc lập với nhau. Tức là sự thay đổi giá trị của một đặc trưng không ảnh hưởng đến các đặc trưng còn lại.
* Các đặc trưng đưa vào mô hình có ảnh hưởng ngang nhau đối với đầu ra mục tiêu.

Khi đó, kết quả mục tiêu y để P(y|X) đạt cực đại trở thành:



Chính vì hai giả thiết gần như không tồn tại trong thực tế trên, mô hình này mới được gọi là naive (ngây thơ). Tuy nhiên, chính sự đơn giản của nó với việc dự đoán rất nhanh kết quả đầu ra khiến nó được sử dụng rất nhiều trong thực tế trên những bộ dữ liệu lớn, đem lại kết quả khả quan. Một vài ứng dụng của Naive Bayes có thể kể đến như: lọc thư rác, phân loại văn bản, dự đoán sắc thái văn bản,...

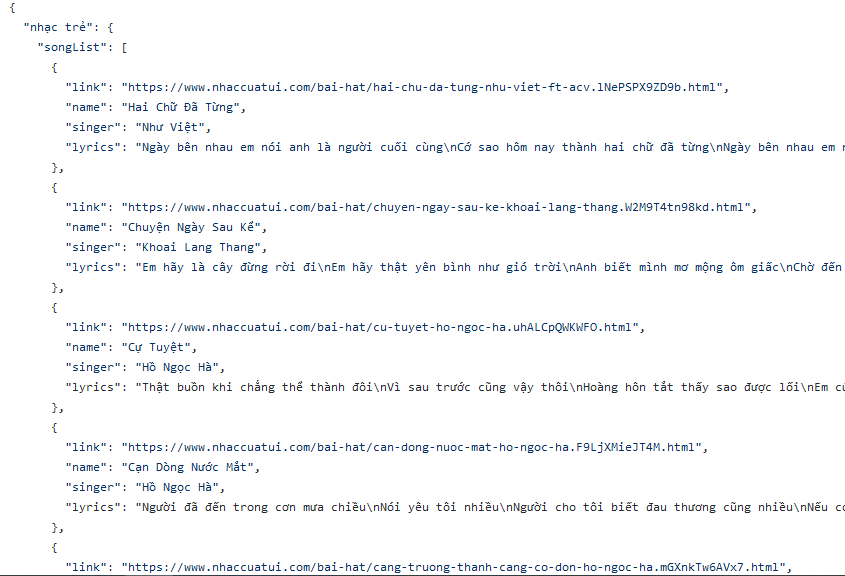
# THỰC NGHIỆM

## Thu thập dữ liệu

Bước đầu tiên là tạo tập dữ liệu. Tất cả các bài hát đã lấy từ trang web nhaccuatui.com.

Tập dữ liệu bao gồm:

* Tập dữ liệu training bao gồm ba thể loại (nhãn), mỗi loại gồm 100 bài hát có thông tin: (thể loại, tên bài hát, ca sĩ và lời bài hát) được biểu diễn dưới dạng JSON của ngôn ngữ Javascript như sau:

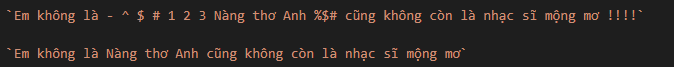


* Tập dữ liệu testing có cấu trúc tương tự như tập training nhưng mỗi loại chỉ có 30 bài hát.

## Preprocessing

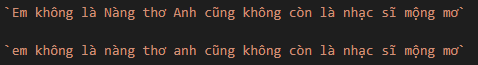
* **Xóa các ký tự đặc biệt:**

**Các ký tự đặc biệt (như !@#$%^&…) hay chữ số không có tác dụng phân loại nhiều nên ta sẽ loại bỏ đi.**



* **Đồng nhất dữ liệu:**

Để dễ dàng cho việc xử lý chuỗi trong thuật toán, ta sẽ chuyển tất cả về dữ liệu về dạng chữ thường để đồng nhất dữ liệu.



* **Tagging - gắn nhãn POS:**

Trong lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP), hầu hết các mô hình cơ bản đều được xây dựng dựa trên phương pháp Bag of Words. Nhưng, các mô hình như vậy không thể xác định được các mối quan hệ về cú pháp giữa các từ.

Ví dụ: với một mô hình phân tích cảm xúc được xây dựng trên phương pháp Bag of Words. Ta không thể xác định được sự khác biệt trong câu :”I like you”, trong đó, động từ “like” là một động từ chỉ tình cảm tích cực. Và tương tự, với một câu khác: “I am like you” thì động từ “like” là là chỉ sự giống nhau giữa hai chủ thể chứ không chỉ tình cảm.

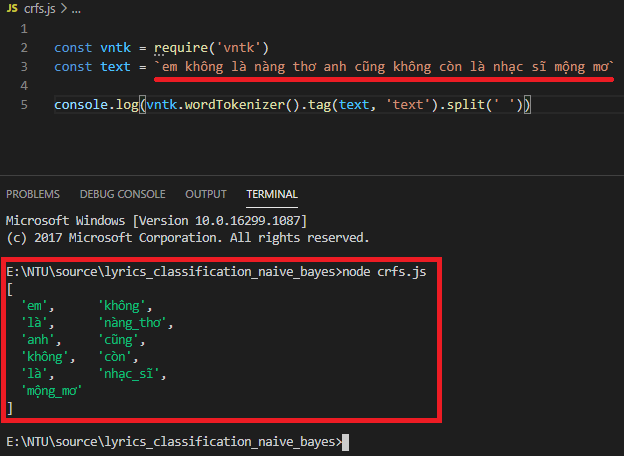
Vậy có cách nào ta có thể cải thiện thuật toán Bag of Words không?

Part of Speech (POS – gắn nhãn từ vựng) là một phương pháp trong xây dựng cây cú pháp mà tại đó ta có thể xác định NERs (các danh từ riêng/chung) và trích xuất quan hệ giữa các từ. Gắn thẻ POS cũng rất cần thiết để xác định được định dạng từ gốc (các dạng từ trong tiếng anh).

Gắn nhãn POS là xác định loại từ của một từ vựng trong ngữ cảnh đoạn văn đó. Để xác định được vấn đề này không hề đơn giản. Vì với một từ cụ thể, trong mỗi ngữ cảnh khác nhau sẽ có những ý nghĩa khác nhau.

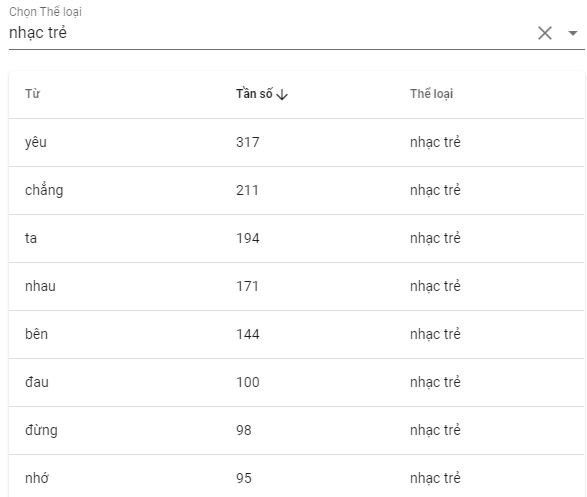
Ta có thể nhìn ví dụ sau: “Give me your answer” thì answer là một danh từ. Nhưng với câu “Answer the question” thì answer là một động từ.

Để máy tính có thể thực sự hiểu nghĩa của một câu và trích xuất được cấu trúc ngữ pháp của nó, thì POS là một bước quan trọng.



* **Feature Extractor**

Ta sẽ sử dụng Bag of Words để tạo ra các vector đặc trưng phục vụ cho classifier.



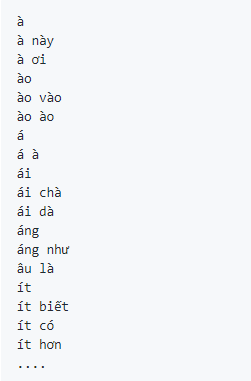
* **Remove Stopwords:**

Sẽ có những từ không cần thiết cho việc phân loại (stopword) nên ta sẽ loại bỏ những từ này đi. Dùng module vietnamese-stopwords.

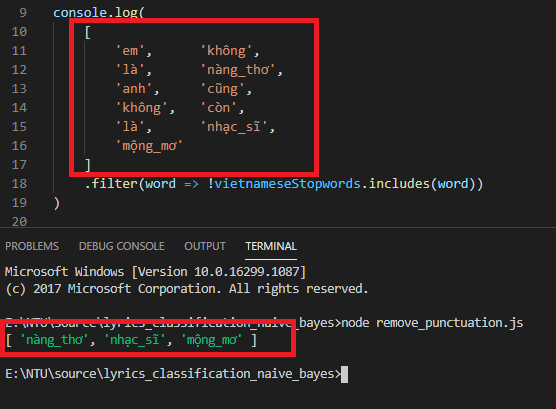
Bộ stopwords của module này có thể xem tại đây:

<https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords-dash.txt>

Một phần của bộ vietnamese-stopwords:



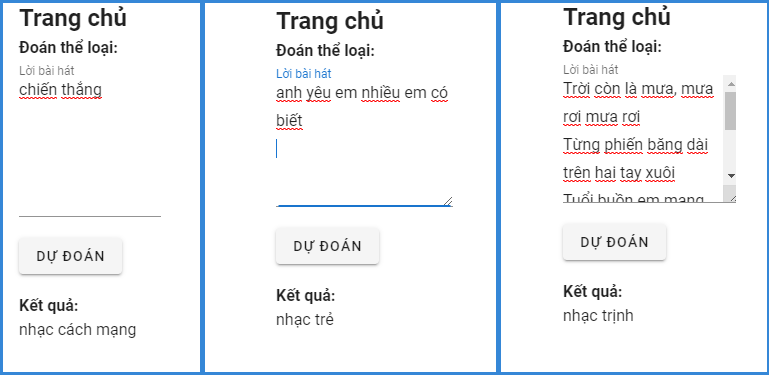
Kết quả sau khi xóa stopwords



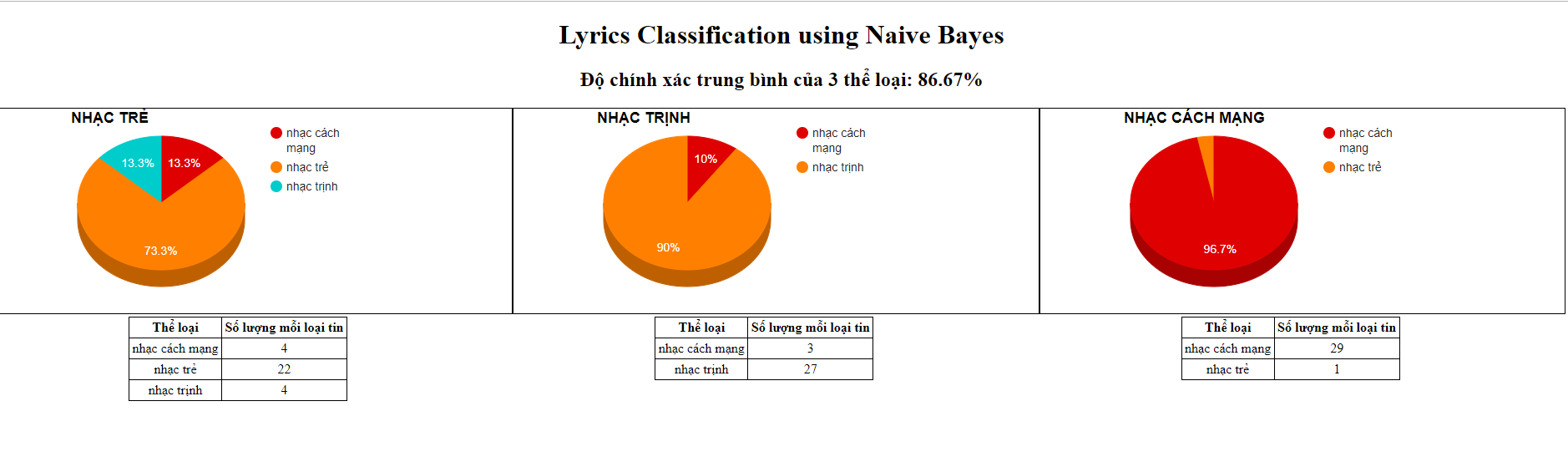
## Kết quả

Sau khi có được data, ta áp dụng thuật toán Naive Bayes dựa trên tập data. Thu được kết quả khá khả quan:

Dự đoán được thể loại bài hát từ các lời bài hát:



Thống kê trên bộ dữ liệu testing.



## Đánh giá

Độ chính xác: 80% - 90%, chưa chính xác hoàn toàn nhưng có thể tạm thời tin tương vào kết quả này.

## Cải tiến

Ngưỡng lượng từ thường xuyên được đặt là 8. Từ xuất hiện ít hơn 8 lần đã bị loại. Mô hình cho kết quả tốt nhất khi ngưỡng là 8 hoặc 9.

Bằng cách tăng hoặc giảm ngưỡng, bộ phân lớp sẽ có độ chính xác lớn hơn.

# KẾT LUẬN

Kết quả của một mô hình được tạo ra là rất tốt. Naive Bayes phân loại là một lựa chọn tốt cho nhiệm vụ này - một lần nữa nó đã chứng minh khả năng của nó. Vì tập dữ liệu khá nhỏ, nó là một ứng cử viên hợp lý cho mô hình.

Toàn bộ source code của bài:

<https://github.com/cvthang56th2/lyrics_classification_naive_bayes>

# TÀI LIỆU THAM KHẢO

1. Viblo.asia, Mô hình phân lớp Naïve Bayes,

<https://viblo.asia/p/mo-hinh-phan-lop-naive-bayes-vyDZO0A7lwj>

1. Blog trí tuệ nhân tạo, Xử lý POS với thuật toán Conditional Random Fields,

<https://trituenhantao.io/kien-thuc/nlp-xu-ly-pos-voi-thuat-toan-conditional-random-fields/>

1. Vũ Hữu Tiệp, Bài 32: Naïve Bayes Classifier,

<https://machinelearningcoban.com/2017/08/08/nbc/>

1. Vũ Hữu Tiệp, Bài 11: Giới thiệu về Feature Engineering,

<https://machinelearningcoban.com/general/2017/02/06/featureengineering/>

1. A. Khan, B. Baharudin, L. H. Lee & K. Khan, "A review of machine learning algorithms for text-documents classification", Journal of advances in information technology, 1(1), 2010, pp. 420.
2. X. Hu & J. S. Downie, "When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis", 11th International Society for Music Information Retrieval Conference (ISMIR), August 2010, pp. 619-624.
3. M. Haggblade, Y. Hong & K. Kao, "Music genre classification",
4. Department of Computer Science, 2011, http://cs229.stanford.edu/proj2011/HaggbladeHongKaoMusicGenreClassification.pdf .
5. E. Stamatatos & G. Widmer, "Music performer recognition using an ensemble of simple classifiers", ECAI, 2002, pp. 335-339.
6. M. Fell & C. Sporleder, "Lyrics-based analysis and classification of music", Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 620-631.