Báo cáo tiến độ

DS310.M11

Đề tài: Phân loại thể loại cho lời bài hát bằng Tiếng Việt

Thành viên nhóm

|  |  |
| --- | --- |
| Họ tên | MSSV |
| Nguyễn Thị Thu Hà | 19521456 |
| Bùi Anh Khôi | 19520649 |

# Giới thiệu bài toán

* thuộc dạng phân loại văn bản (text classification)

# Dữ liệu

## Nguồn dữ liệu

### Hợp âm chuẩn

* được thu thập từ <https://hopamchuan.com/>
* thống kê thể loại

|  |  |
| --- | --- |
| **Thể loại** | **Số lượng bài** |
| nhac tru tinh | 870 |
| nhac tre | 711 |
| nhac ton giao | 479 |
| nhac thieu nhi | 431 |
| rap viet | 426 |

### Nhạc của tui

* được thu thập từ <https://www.nhaccuatui.com/>
* thống kê thể loại

Table

Description automatically generated with medium confidence

|  |  |
| --- | --- |
| **Thể loại** | **Số lượng bài** |
| nhac-tre | 673 |
| tru-tinh | 642 |
| rap-viet | 552 |
| thieu-nhi | 453 |
| rock-viet | 358 |
| tien-chien | 194 |
| cach-mang | 132 |

# Một số vấn đề gặp phải

Dữ liệu ít, chênh lệch giữa các thể loại

Có nên merge

Một số từ đặc trưng của lyric như verse, hook, Đk

Một số bài rap có từ nhạy cảm có dấu “\*”

# Kế hoạch tiếp theo

Hoàn thiện phần xử lý dữ liệu

Phân tích đặc điểm từng thể loại

Tìm hiểu và áp dụng các mô hình ML và DL vào bài toán

Viết báo cáo