BÁO CÁO BÀI TẬP GOM NHÓM TRÊN TẬP DATASET CALTECH256

1. Người thực hiện:

Họ tên: Nguyễn Minh Dũng MSSV: 15520138

Link source code: https://github.com/nguyendu392/15520138

2. Mô tả bài toán:

Gom nhóm các ảnh trong tập dataset sử dụng thuật toán gom nhóm Kmeans

3. Các bước thực hiện:

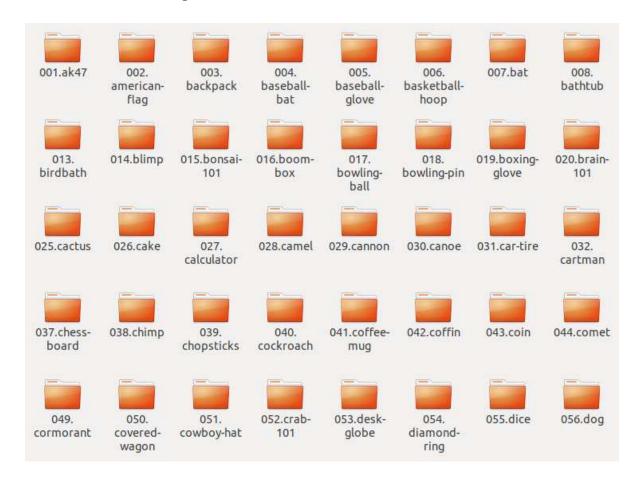
B1: Chuẩn bị dữ liệu B2: Xây dựng Model

B3: Đánh giá

4. Hiện thực:

4.1. Chuẩn bị dữ liệu:

• Dataset: Caltech 256 Image Dataset



 Mỗi ảnh trong dataset được rút trích đặc trưng bằng VGG-16 (output = fc2) dùng để training cho Model.

1	10	1	1	1 10	1	1	1	1
10	10	10	10	10	10	10	10	10
101	101	101	101	101	101	101	101	101
1010	1010	1010	1010	1010	1010	1010	1010	1010
001_0001.	001_0002.	001_0003.	001_0004.	001_0005.	001_0006.	001_0007.	001_0008.	001_0009.
jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl
10 10 101 1010	10 10 101 1010	10 10 101 1010	10 10 101 1010	1 10 101 1016	10 10 101 1010	1 10 101 1010	10 101 1010	10 10 101 1010
001_0013.	001_0014.	001_0015.	001_0016.	001_0017.	001_0018.	001_0019.	001_0020.	001_0021.
jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl
10 10 101 1010	1 10 101 1010	1 10 101 1010	1 10 101 1010	1 10 101 1010	10 101 1010	1 10 101 1010	1 10 101 101 1010	1 10 101 1010
001_0025.	001_0026.	001_0027.	001_0028.	001_0029.	001_0030.	001_0031.	001_0032.	001_0033.
jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl
10 10 101 1010	10 10 101 1010	10 101 101 1010	10 10 101 1010	10 10 101 1010	10 101 1010	1 10 101 1010	10 10 101 1010	1 10 101 1010
001_0037.	001_0038.	001_0039.	001_0040.	001_0041.	001_0042.	001_0043.	001_0044.	001_0045.
jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl
1 10 101 1010	10 101 1010	1 10 101 1010	1 10 101 1010	10 101 101	1 10 101 1016	1 10 101 1010	1 10 101 1010	10 101 1010
001_0049.	001_0050.	001_0051.	001_0052.	001_0053.	001_0054.	001_0055.	001_0056.	001_0057.
jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl
10 10 101 1010	10 10 101 1010	10 101 1010	1 10 101 1010	10 101 1010	10 101 1010	10 101 1010	1 10 101 1010	10 10 101 1010
001_0061.	001_0062.	001_0063.	001_0064.	001_0065.	001_0066.	001_0067.	001_0068.	001_0069.
jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl	jpg.pkl

• Chia random dataset thành 70% training và 30% test (Cross Validation).

4.2. Xây dựng Model:

- Sử dụng features đã rút trích để training cho Model.
- Sử dụng phương pháp KMeans để gom nhóm các features.

4.3. Đánh giá:

• Để đánh nhãn cho các nhóm sau khi phân nhóm bằng Kmeans ta lấy nhãn (theo dataset ban đầu) có tần xuất suất hiện nhiều nhất ở mỗi nhóm.

• Kết quả sau khi chạy trên 3 database được tạo bằng cách random.

	Kmeans
Database 1	56.09%
Database 2	55.65%
Database 3	56.43%

=> Độ chính xác: 56.06±0.41 %

5. Nhận xét:

- Do Kmeans chủ yếu dùng để gom nhóm cho hệ thống **Unsupervised learning**
- Có thể cải thiện độ chính xác bằng cách tìm giải thuật đánh nhãn cho các nhóm sau khi gom khác.