Clustering into Groups of 5 Example

The key difference from the initial 100 selection is that for clustering, we:

1. Use average similarity instead of minimum similarity
2. Build clusters one at a time
3. Fill each cluster to exactly 5 images before moving to the next

Let's walk through a simplified example with 15 images to make 3 clusters of 5.

Cluster 1 Formation

Step 1: Random Start

- Randomly select first image for cluster 1, say 'A'
- Cluster 1: [A]
- Remaining: [B, C, D, E, F, G, H, I, J, K, L, M, N, O]

Step 2: Fill Cluster 1

1. Calculate average similarity to current cluster members:

```
For B: avg_similarity(B, [A]) = 0.7
For C: avg_similarity(C, [A]) = 0.3
For D: avg_similarity(D, [A]) = 0.8
...etc
```

2. Select image with lowest average similarity
    - Say 'C' has lowest average
    - Cluster 1: [A, C]
3. Repeat until cluster has 5 images:
    - Calculate average similarity to [A, C]
    - Select lowest average
    - Continue until Cluster 1 has 5 images

Final Cluster 1: [A, C, F, K, N]

Cluster 2 Formation

Start New Cluster

- Randomly select from remaining images for cluster 2
- Repeat filling process
- Each selection minimizes average similarity to current cluster

Final Cluster 2: [B, E, H, L, O]

Cluster 3 Formation

- Use remaining images
- Same process as above

Final Cluster 3: [D, G, I, J, M]

Key Differences from 100 Selection

1. Uses average similarity instead of minimum:

```
# In the code:
similarities = cosine_similarity(remaining_features, selected_features)
avg_similarities = similarities.mean(axis=1)  # Takes average along row
most_diverse_idx = np.argmin(avg_similarities)
```

2. Builds complete clusters sequentially:
    - Finishes one cluster before starting next
    - Each cluster is independent of others
    - Only considers similarity within current cluster
3. Cluster size is fixed:
    - Each cluster must have exactly 5 images
    - No overlap between clusters
    - All images must be used

This approach ensures:

- Images within each cluster are maximally different from each other
- Each cluster is formed independently
- Even distribution of images (5 per cluster)