K Means

This code is implemented as part of Homework 3 in CS 6375 Machine Learning Course at The University of Texas at Dallas by Harsha Kokel. It is written with java 8.

The task is to analyze image compression achieved by k means clustering for two images `Koala.jpg` and `Penguins.jpg`.
The program initializes cluster centers at with distinct rgb values from the image selected at random. It then clusters the pixel values across these cluster centers, finds the new mean of these clusters and repeats the process till convergence or maximum iterations. Convergence is achieved when the cluster means do not change in consecutive iterations. The distance measure used for finding the nearest cluster center is Euclidean distance with respect to Red, Green, Blue and Alpha values.
For this code maximum Iteration is set to 1000, however in most cases the convergence is reached before 100 iterations.

Usage

```
Kmeans <input-image> <k> <output-image>

<k> is the number of clusters.
```

Results

Theoretically, there is a **trade off** between the image quality and the compression ratio. Lower the compression ratio, higher the image quality. However, with our experiments we see that higher compression ratio can be obtained even with comparable image quality. For Koala.jpg, the compression ratio increases with the increase in K. This is counter intuitive. For Penguins.jps, as expected the compression ratio decreases with increase in k and the image quality decreases.

Koala.jpg

- Original Image Size : 780831 Bytes
- Best value of K : 20
- The table below provides the size (in Bytes) of the compressed image obtained for different K values.

# Clusters	1	2	3	4	5	6	7	8	9	10
2	130163	130151	130163	130151	130853	130853	130151	130151	130853	130151
5	175613	176597	175600	176538	175941	175625	175635	176468	176554	176540
10	163771	164923	163560	165174	163781	165069	163545	163479	163635	163551
15	158598	158316	159583	157002	156964	158206	158997	159376	159047	160200
20	158005	158746	158341	155916	153843	154868	157842	155690	154576	157699

- The table provides the average image size (in Bytes), average compression ratio and variance for different K values.

# Clusters	Average	Compression Ratio	Variance
2	130364	5.99	102501
5	176111.1	4.43	192706
10	164048.8	4.76	164048.8
15	158628.9	4.92	158628.9
20	156552.6	4.99	2833172

Penguins.jpg

- Original Image Size : 777835 Bytes

- Best value of K : 20
- The table below provides the size (in Bytes) of the compressed image obtained for different K values.

# Clusters	1	2	3	4	5	6	7	8	9	10
2	85013	84994	85013	85013	84994	85013	84994	85013	84994	84994
5	108318	105119	105714	105855	106250	108318	106250	104248	105828	107334
10	116596	119824	117621	119588	115873	117801	117602	117643	120468	117966
15	116237	115024	114340	118581	116085	115527	116888	114297	116954	117046
20	116171	116559	115009	116518	115156	116081	115231	115113	116286	116354

- The table provides the average image size (in Bytes), average compression ratio and variance for different K values.

# Clusters	Average	Compression Ratio	Variance
2	85003.5	9.15	90.25
5	106323.4	7.32	1558299
10	118098.2	6.59	1880890
15	116097.9	6.70	1625400
20	115847.8	6.72	366461