CALIFORNIA STATE UNIVERSITY, LOS ANGELES
PROGFEST 2012

Problem 1
Image Segmentation using Clustering

It is the year 2112. The earth has run out of natural resources, and the fate of humanity depends on finding a suitable earth-like replacement.

Years ago, NASA had predicted this would occur, and had sent several probes to investigate potential candidates. For the past 50 years, the probes have been returning imagery of each planet they encountered. To aid scientists in their search, you have been chosen to narrow down the list of potential candidates. One simple criteria would be to identify planets that have an appropriate amount of water, land, and vegetation. For example, the earth is about 70% water, 20% land, and 10% vegetation.

For this problem, you will implement the k-means algorithm to partition the image pixels into 3 groups (a.k.a. clusters). Each pixel is represented by a vector of its red, green, and blue components (collectively referred to as RGB values). The algorithm works by assigning pixels to a group, and iteratively refining those groups until they best represent their constituent pixels. More specifically, it can be described as follows:

1. Initialize the cluster centers to a random vector of pixels
2. Using euclidean distance, assign each pixel to a group based on its closest cluster. The euclidean distance between the cluster center vector $\mathbf{c}$ and pixel vector $\mathbf{p}$ is defined as:

$$d(\mathbf{c}, \mathbf{p}) = \sqrt{(c_{red} - p_{red})^2 + (c_{green} - p_{green})^2 + (c_{blue} - p_{blue})^2}$$

3. Compute a new cluster center by computing the average of all pixels in a group
4. Repeat from (2) until convergence (the pixels do not change groups)

The input to your program will be given in a text file, to be specified as a command line argument. To facilitate testing, we will provide you the initialization vector in the input text file. Thus, the format for the input will be as follows:

- Line 1: Initial vector of cluster centers given as comma-separated RGB values.
  Eg. $cluster1_{red}, cluster1_{green}, cluster1_{blue}, \ldots, cluster3_{red}, cluster3_{green}, cluster3_{blue}$
- Lines 2 to EOF: A pixel of satellite imagery, represented by a comma-separated RGB values

Your program must output the fraction of pixels that belong to each of the 3 classes, each on a new line, rounded to 3 decimal digits. Ideally, these would correspond to the fraction of the planet consisting of water, land, and forest.

SAMPLE INPUT:

```
255.0, 0.0, 0.0, 0.0, 255.0, 0.0, 0.0, 0.0, 255.0
183.0, 180.0, 81.0
46.0, 108.0, 147.0
67.0, 71.0, 207.0
```

```
152.0, 96.0, 99.0
41.0, 200.0, 183.0
173.0, 33.0, 215.0
16.0, 57.0, 87.0
69.0, 187.0, 40.0
159.0, 84.0, 154.0
224.0, 250.0, 125.0
```

SAMPLE OUTPUT:

```
0.300
0.300
0.400
```