Problem 2: Visualizing k-means

1. No, the initial centroids differ across runs in the simulation since they are a randomly chosen subset of the points in our dataset
(cvisualization.py >> initialCentroids = random.sample(points, k)

2. No, the clusters are not identical even if the simulation is run on the same 2 features. This is because the clusters obtained depend upon which points are initially selected as centroids; as they are chosen randomly, how points are grouped into clusters at each iteration will differ across runs.

>> include 2 screen shots of final clusters

3. ~# of iterations for clusters to be stable for features (ArtistHotness and SongHotness, k = 6):

When reducing cutoff from 0.1 to 0.01:
> the number of iterations required for clusters to be stable changes by ~ _____
> the maximum cluster diameter changes by _____

>> include 2 screenshots of final clusters

4. Small clusters are useful because they provide information about more fine-grained relationships between variables; however, as k gets very large, the relationships between the data points in the clusters becomes more arbitrary and less meaningful (at the limit, each data point can be considered its own cluster—at which point, it would not be providing any relational information)

Problem 3: k-means and individual songs
** 4 questions

Problem 4: predicting song hotness with k-means