Homework 3: due May 16, 2016

Problem 1: Download the dataset `crescents.mat` from the Class website and load it into Matlab. It contains two-hundred points in two dimensions. If you plot it, you see that the data form two half-moon-like clusters. Clearly, k-means directly applied to this dataset will fail to cluster the data according to these two shapes. Use the graph Laplacian or diffusion maps (followed by k-means) to try to cluster the data as good as possible according to the half-moon shapes. You can use Matlab's k-means function to do the actual clustering once you transformed the data.

Problem 2: Download the dataset `genomedata.mat` from the Class website; it contains Single Nucleid Polymorphisms data from the Human Genome Diversity Project. The data consists of an array consisting of 5000 rows, each row has 1043 different strings. The 5000 rows are Single Nucleid Polymorphisms, the columns correspond to 1043 different individuals. The entries are not numerical values (quite annoyingly), but contain the characters 'AA','CC','GG','TT','AG','AC','TC','TG', and (even more annoyingly) also '`--`', the latter represents missing measurements. Your goal is to cluster the data into a small set of clusters. After loading the file into Matlab, you need to convert the characters into numerical values. It is up to you which conversion you use (you can use the file `gen2vec.m` to do the actual conversion, once you have chosen a conversion rule).

Since the data are high-dimensional you first need to reduce the dimension before clustering. You should attempt the dimension reduction via PCA as well as via diffusion maps. In both cases you need to decide how many dimensions you want to use. Also, in both cases you may want to use Matlab's k-means function to do the actual clustering after dimension reduction.

Note: Your results may differ from mine, because you will likely choose a different conversion rule. I did not get a meaningful clustering via PCA, but did achieve reasonable clustering via diffusion maps.