INFSCI 1530/2160: Data Mining

Homework 3

A. You are provided with a dataset (http://pitt.edu/~kpele/movies.csv) that includes information about various movies such as budget, IMBD score, duration etc.

1. Provide a *good* clustering for the data and a description of them (i.e., what type of movies do they include) (25 points)
   Hint: Examine up to 20 clusters.
2. As you will observe, the various features of the data are in different scales. In these cases, normalizing the data helps into obtaining better models – either supervised or unsupervised. Use the sklearn.preprocessing.normalize function to normalize the data and perform the clustering again. What do you observe? (25 points)
3. Visualize the data on a 2D scatter plot, where each point corresponds to a movie and movies that belong to the same cluster have the same color. What do you observe with regards to their *visual separability*? (25 points)

B. Every weekend, you drive into town for your contactless curbside pickup at your favorite restaurant. Across the street from the restaurant are six parking sports, lined up in a row. While you can parallel park, it is certainly not your preference. You will not be required to parallel park when the rearmost of the six spots is available, or when there are two consecutive open spots. If currently there are four cars occupying four of the six spots in a random arrangement, what is the probability that you will have to parallel park? (25 points)