Homework 3
Statistical Data Mining 2

1) Consider the Utility Matrix Below that represents the ratings, on a 1-5 scale, of eight items, a through h, by three users: A, B, and C. Compute the following from the data of this matrix.

		items						
		A	b	c	d	e	f	g	h
Users	A	5	4		5	2		3	2
	B		3	4	4	2	2	1	
	C	3		1	4		4	5	3

a) Treating the utility matrix as Boolean, compute the Jaccard distance between each pair of users.
b) Repeat Part A, but use the cosine distance.
c) Treat ratings of 3, 4, and 5 as 1, and ratings 1, 2, and blank as zero. Compute the Jaccard distance between each pair of users.
d) Repeat Part C, but use the cosine distance.
e) Normalize the matrix by subtracting from each nonblank entry the average value for its user.
f) Using the normalized matrix from Part E, compute the cosine distance between each pair of users.

2) ISLR Chapter 10 Question 2

3) Adapted from ISLR Chapter 10 Question 10.
a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total) and 50 variables.
Hint: there are a number of functions in R that you can use to generate data, rnorm() and runif() are two options. Be sure to add a mean shift to the observations in each class so that there are three distinct classes – remember to set.seed().

b) Perform k-means clustering of the observations with K=3. Using the rand index and adjusted rand index, assess how well do the clusters that you obtained in K-means clustering compare to the true labels?

c) Using silhouette plots, select the optimal number of clusters.

d) Using the gap statistics, select the optimal number of clusters.