Assiggnment 2

1) There are two parts to this assignment and you will do one of them.
2) We will use TURNITIN on your submissions so refrain from copying. Discussion is encouraged.
3) Use of Matlab/Octave will be beneficial. But do not use directly available libraries, write your own code.

Problem set 1

1) Implement the following clustering algorithms on handwritten-digit classification dataset.
    1. K-means
    2. K-medoid
    3. DBScan
2) There are two files - train.dat and test.dat. Location:
   http://www.cse.iitd.ac.in/~prajna/COL868/Assignment2/train.dat
   http://www.cse.iitd.ac.in/~prajna/COL868/Assignment2/test.dat
3) train.dat will be input to the algorithms. Each entry is a 1x256 vector representation of a 16x16 image of a handwritten number. The numbers vary from 0-9. So, there will be 10 clusters. For the demo, you will be asked to display the test data point (from test.dat) using imshow() function (available in both Matlab/Octave) to test whether it is being predicted correctly or not.
4) For each of the algorithm, report the errors.

Problem set 2

1) Implement the following algorithms on nips dataset
    1. Agglomertive
    2. Hierarchical
    3. Birch
2) There is one dataset, dataset2.txt. Location:
   http://www.cse.iitd.ac.in/~prajna/COL868/Assignment2/dataset2.txt
3) The first 3 lines indicate the number of documents in the collection, number of words in the vocabulary and the total number of words in the collection. The remaining lines are in the form:
    1. document number, word number, frequency in the document
4) Here is the vocabulary list:
   http://www.cse.iitd.ac.in/~prajna/COL868/Assignment2/vocab.txt
5) You have to preprocess this dataset to convert in into a matrix form. Use the matrix representation as input to the clustering algorithms.
6) Each row will represent a document. Each column will represent a word from vocabulary. Each entry (i,j) of the matrix will be the frequency of the word in the document.
7) For the top-down clustering approaches, use simple k-means to split the cluster. For the demo, you will be asked to display the dendogram upto level 2 or 3. We will also go through the list of words appearing in each documents in the dendogram.

The assignment of algorithms to different students is given below:

Student	Part1	Part2

NITESH SINGH	√	
AJITA SHREE		√
KUNAL KISHOR	√	
MILAN BHANJIBHAI KATHROTIA		√
VAMSI YALAVARTHI	√	
AKSHAY SURESH BHAT		√
SAHIL YADAV	√	
SHUBHAM SAGAR		√
SURYAKANT PANDEY	√	
VISWA TEJA GAJULAVARTHY		√
EDUBILLI AVINASH	√	
NIKHIL KUMAR		√
ABHISHEK KUMAR	√	
AYUSH VERMA		√
HIRULKAR ANKET PRAKASH	√	