Assignment 5

1. **Document clustering using Fuzzy C-Means Algorithm**

Use following steps:

i) Consider NewsGroup20 dataset. Preprocess them by performing stopword removal and stemming operation.

ii) Count number of distinct words considering all documents, say it is n.

iii) Represent documents by an m X n matrix where m = number of documents.

iv) For each document, compute TF-iDF (i.e., Term Frequency-inverse Document Frequency) value of each word.

v) Discretize all the values and get a decision system DS = (D, W, C) where, D = set of documents, W = set of conditional features (i.e., words) and D = Decision feature (NewsGroup20 dataset contains 20 different classes).

vi) Apply QuickReduct generation Algorithm and get the reduced set of words. Thus the decision system is also reduced.

vii) Remove the decision class from the reduced dataset and apply Fuzzy C-Means algorithm.

viii) Display the result.