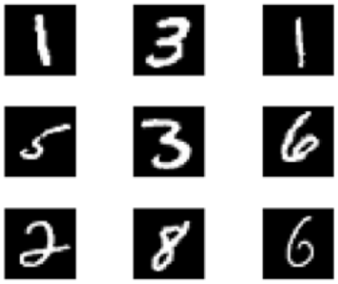
Assignment 4 Due Date: 30/11/2023

Goal: Performing K Means clustering on **image** data and text data

Image data: MNIST data set



1. Use google colab. Google colab has GPU so your program will be fast in google golab.
2. We will in this assignment cluster images from MNIST (Mixed National Institute of Standards) database of handwritten digits. MNIST train data set containing *N* = 50000 grayscale images of size 28 × 28 as shown above. We will follow the steps specified below:
3. After reading the train dataset, you will get a 784 × 50000 matrix digits and a 1 × 50000 matrix labels.
4. We will not need labels for clustering. However, for external validation, we will use them. Each column of digits is a 28 × 28 grayscale image, stored as a vector of length $28^2 = 784$ with elements between 0 and 1
5. You are asked to apply the k-means algorithm to this set of N = 50000 vectors, with k = 10 groups, and starting from a random initial group assignment
6. Calculate the quality of the final clusters using following criteria
    a. Internal Validation
        i. $$J = \frac{1}{N}\sum_{i=1}^{N} \min_{j=1,\ldots,k} \|x_i - z_j\|^2$$
        ii. Davies–Bouldin index
        iii. Dunn index
    b. External validation:
        i. Purity
        ii. Rand index

7. Plot the 3D plots for the randomly chosen 1000 data from the MNIST using the tsne. Use different colors for the points that are assigned to different clusters. Also plot the centroids in bigger markers.

Assignment 4 Due Date: 30/11/2023

Text clustering: Clustering Prothom-alo news

1. Use the dataset described here: https://github.com/banglanlp/bnlp-resources/tree/main/news_categorization
2. Download 1200 different news. 1200 will be from 6 different categories (200 from each category)
3. Now use following procedures to calculate feature vectors for the text associated with each editors' column.
    a. Use the tokenizer to extract all words for all sentences
    b. Use stemming.
    c. Normalize the words
    d. Discard stop words.
    e. Use Zipf's law to discard most frequent and least frequent words
    f. Build the vocabulary/term list
    g. Calculate TF matrix for all comments/documents. **(You can't use library functions)**
    h. Calculate normalized TF matrix **(You can't use library functions)**
    i. Calculate IDF for all terms **(You can't use library functions)**
    j. Calculate final weighted matrix for all documents. **(You can't use library functions)**
    k. Now use the rows as feature vectors for corresponding documents/columns. We don't use the category for clustering.
    l. We will use this information for external validation
    m. Use K-mean clustering with the feature vectors to cluster the 1200 documents into 6 categories.
4. Calculate the quality of the final clusters using following criteria **(You can't use library functions)**
    a. Internal Validation
        i. $$J = \frac{1}{N}\sum_{i=1}^{N} \min_{j=1,\ldots,k} \|x_i - z_j\|^2$$
        ii. Davies–Bouldin index
        iii. Dunn index
    b. External validation: Here you will use the ground truth categories (that you have collected during the scrapping)
        i. Purity
        ii. Rand index

5. Plot the 3D plots for the 1200 documents using three-dimension data that you will get from the tsne library. The tsne will take high dimension data and return 3-dimension data for plotting. Use different colors for the points that are assigned to different clusters. Also plot the centroids in bigger markers.