CSE351
HW 4
Due July 7th at 11:59 pm
Acceptable format: ipynb

Download the speech files of 10 celeberities from VoxCeleb1 at VoxCeleb . The task of this HW is to build a speaker identification model. In this model, you input an audio from one of these 10 selected celebs and the model identifies the celeb from his/her voice.
Let $x_j^i(t)$, for $t = 1, \ldots, T$ be the jth speech file of the ith speaker of the length T. In here, we only build the model for 10 speakers, i.e. $i = 1, \ldots, 10$. For each speaker you select 100 speech files, $j = 1, \ldots, 100$. In total, you should have at least 2 minutes speech files for each speaker. Do the following steps to build the speaker identification model:

- **Prepration:** Concatenate all speech files of one speaker to one file. we denote it as $x^i(t)$ for the ith speaker where $i = 1, \ldots, 10$. Note that you should have at least 2 minutes speech for each speaker.

- **Feature extraction:** transform $x^i(t)$ to a mel spectral feature matrix. You can do this using librosa.feature.melspectrogram function available at librosa . Reduce the sampling frequency to 8 khz and use the window of the length 30 msec with a frameshift of 10 msec. Read the function document and example in the given link how to load a speech file and transfom it to the melspectrogram. After applying this transormation you have your feature vector of the form of a matrix $X_{N^i \times K}$ where N^i and K denote the number of samples for the ith speaker and K is the dimension of the feature space and is the same for all speakers. You preform this process for all speakers and create the $A_{N^1+N^2+\ldots+N^{10},K} = [X_{N^1 \times K}; X_{N^2 \times K}, \ldots, X_{N^{10} \times K}]$. This matrix is the training A matrix.

- **Labels:** for each sample, i.e. a row in matrix A, you need to assign a label to identify to which speaker this sample belongs to. So we create the vector $\overrightarrow{y}$ of dimension $N^1 + N^2 + \ldots + N^{10}$ that contains the abels for each sample (aka frame); this is the trian y vector.
- **Training:** Having A, the input feature matrix and $\overrightarrow{y}$, the output labels. build a classifier using a feedforward neural networks with one hidden layer and 256 hidden units and train it with 1000 epoch.
- **Confusion matrix**: report the performace of your speaker identification model using a confusion matrix.
- **Clustering**: using the k-means algorithm cluster A into 10 clusters and find the membership assignment for each sample. Calculate how many samples are clustered together for each speaker.