TASK 2

Problem

You are given access to an AWS S3 bucket containing 2,000,000 medical images (MRI scans) in JPEG format. The dataset almost entirely consists of scans of five different body parts: head, neck, spine, abdomen, and pelvis. Contained within this 2,000,000 image dataset are also four distinct black and white images of the CN Tower. There is no difference in image metadata or resolution between the medical images and the images of the CN Tower.

Solution

So here the data has 5 main clusters of images and 4 outliers samples. One way is to get the latent representations from the images and cluster them. there should be 5 main clusters and another very small cluster consisting of 4 samples ideally. So, heads vectors will be close together, pelvis vectors should be in their own cluster, and so should the 4 CN tower images. Most of these solutions will need GPU to perform transformations and matrix multiplications on the large data. Technically they can be done on the cpu, but will take much more time. One can also try outlier detection techniques but they may work differently based on the underlying colour distribution.