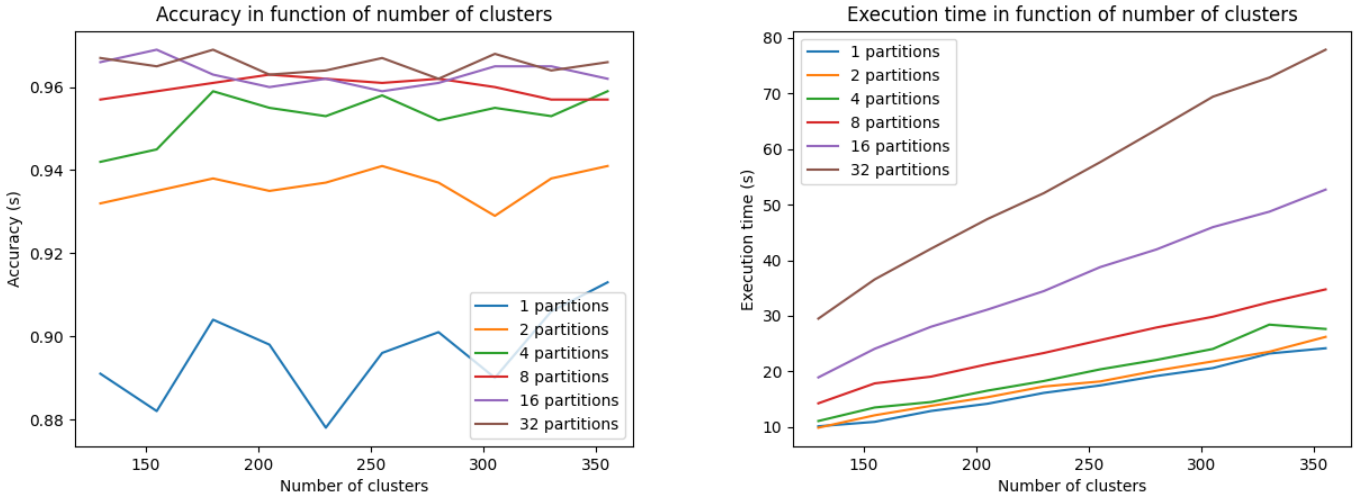
Assignment 3: BDAP [B-KUL-H00Y4A]

Andreas Hinderyckx

Question 1 The approximate product quantization (PQ) method is faster than the naive nearest neighbors (NN) because it limits the amount of distance computations, which dominate execution time. Concretely: for naive NN N distances are computed in the case of N training examples. When using PQNN, the N training instances are mapped onto `nclusters` $\ll N$ centroids. Only the distances between the query example and the `nclusters` must be calculated, hence the computation time drops significantly.

Question 2 The effect of `npartitions` and `nclusters` is shown in figure 1.



(a) Accuracy

(b) Execution time

Figure 1: Effect of number of clusters and number of partitions on accuracy (1a) and execution time (1b) of `ProdQuanNN` executed on the `covtype` dataset

Question 3 From figure 1b it can be seen that the execution time grows linearly with the number of partitions and the number of clusters.