HW - 5

- Which distance measure is better and why it is better.

  According to my observations, I think cosine similarity is better than Euclidean distance. Cosine similarity has an advantage over Euclidean distance for text document analysis. Since the length of each text document in the train and test sets varies, the algorithm's conclusions may alter if a term is discovered to be frequent in one document but not so frequent in others. This is where cosine distance comes into play and avoids this.
  The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, the higher the cosine similarity. This is where cosine distance comes into play and avoids this. That is why I think cosine distance is better than Euclidean distance.

- Could you assign a meaningful name to each cluster? Discuss how you interpret each cluster
  I interpreted the clusters by taking out samples from the test set.
  Clusters - Nature, Health, Religion, Business

- How did you pick the parameters such as the number of clusters, variance type etc.?

  I fitted the training data for four distinct models using covariance values of "spherical", "tied" and "diag" for covariance, and I have then selected the model with the lowest BIC (Bayesian Information Criterion) score.
  Later on, I calculated the BIC score for a range of 2 to 8 clusters. We then select the 4 clusters with the lowest scores.
  In this way, I was able to pick the parameters for training the model.

- Compare to Kmeans in Q1, do you achieve better performance by GMM?

  After implementing both, I have observed that K means achieves better performance than that of GMM.

- Based on the top words of each topic, could you assign a meaningful name to each topic?

  After analyzing the top words for each topic, I have assigned the following names -

  Topic 0 - Nature, Topic 1 - Health, Topic 2 - Religion, Topic 3 - Career/ Business

- Although the test subset shows there are 4 clusters, without this information, how do you choose the number of topics?

  If we did not have information from the subset that there were 4 clusters then perplexity can be used to determine the number of clusters. A way of choosing topics is by choosing the topics that give the lowest perplexity score. We chose the topics with the lowest perplexity score as we all know the lower the perplexity score the better-generalized performance we get.

- Does your LDA model achieve better performance than KMeans or GMM?

  From the results, it is clear that Kmeans has achieved the best performance among the three.