Q1. Briefly discuss:

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

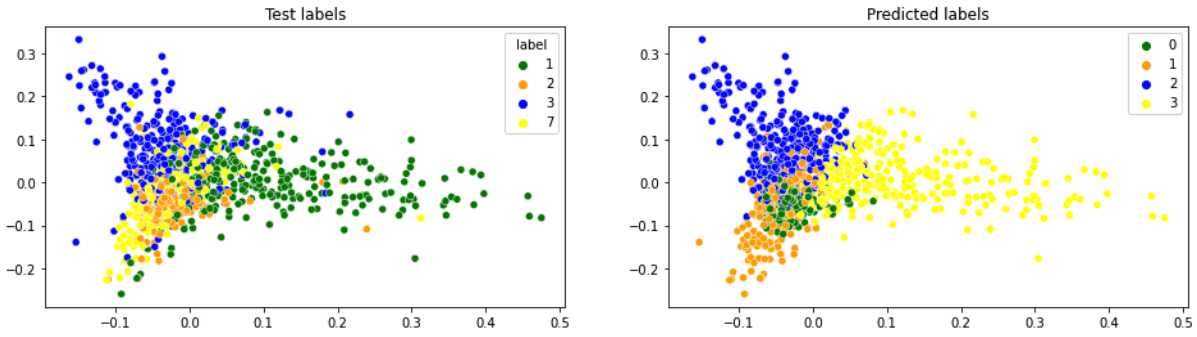
ANS. Euclidean distance is very commonly used, easy to compute and it works well with datasets with compact or isolated clusters but it is sensitive to outliers.

On the other hand, cosine measurement is independent of vector length and invariant to rotation which is what makes cosine distance a better measurement for text similarity applications.

2. Could you assign a meaningful name to each cluster? Discuss how you interpret each cluster.

ANS. - In this example the data set does not have labels with meaning names attached to them. Which makes it unsupervised learning.

- Some ways to give meaning names to these clusters can be as follows:
  - We can randomly or manually assign them names looking at a few samples in a given cluster
  - We can make use of unsupervised multi-document keyword extraction.
  - I visualized the clusters by plotting them in a 2 dimensional space, which was achieved by applying PCA




Q2. Write your analysis on the following:

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

ANS. By iterating through the parameters and checking the Bayesian Information Criterion (BIC) we can pick the optimal parameter.

- Here we are looking for the parameters that give a model with the lowest BIC. Smaller the BIC, better the model. In other words, a complex model is penalized due to its large number of parameters.
- I used different initial means (i.e. `n_init` parameter) when fitting the model to achieve model stability

2. Compared to Kmeans in Q1, do you achieve better performance by GMM?

ANS. No, the accuracy of Kmeans was higher compared to GMM

Q3. Finally, please analyze the following:

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

ANS. The model extracts the top words of each topic, which helps us intuitively assign a meaning name to each topic.
For example the 1st cluster in my case could possibly be called 'financial news' because some of the most common words are news, market, finance etc.

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

ANS. - There are no fixed or golden rules to choose the number of topics. Perplexity may be one way to find the number of topics.

- The best number of topics should be around the lowest perplexity.
- We can manually scan and figure out the possible topics in the data.

To choose the best number of topics:

- We can interpret the number of topics K by plotting perplexity again various numbers of K

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

ANS. Yes, the LDA model has a better accuracy compared to the Kmeans and GMM

Q4. Parameters of the model:

num_topics = 4

alpha = 0.001
iteration = 30

- In order to measure the coherence of the topic I have used both original and preprocessed text.
- I got a coherence score of 0.43499 with original text and a coherence score of 0.746 with preprocessed text
- I have calculated the perplexity score and plotted an interactive graph showing the cluster separation.
- Topic separation can be visualized as follows

Selected Topic: 1 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

λ = 1

0.0 0.2 0.4 0.6 0.8 1.0