Q2:Performance comparison between Q1 and Q2:

actual class	Disaster and Accident	News and Economy	Travel & Transportation
cluster			
0	39	5	127
1	160	5	46
2	11	196	11

Cluster 0: Topic Travel & Transportation
Cluster 1: Topic Disaster and Accident
Cluster 2: Topic News and Economy

	precision	recall	f1-score	support
Disaster and Accident	0.76	0.76	0.76	210
News and Economy	0.90	0.95	0.92	206
Travel & Transportation	0.74	0.69	0.72	184
micro avg	0.81	0.81	0.81	600
macro avg	0.80	0.80	0.80	600
weighted avg	0.80	0.81	0.80	600

actual class	Disaster and Accident	News and Economy	Travel & Transportation
cluster			
0	68	0	151
1	132	7	8
2	10	199	25

Cluster 0: Topic Travel & Transportation
Cluster 1: Topic Disaster and Accident
Cluster 2: Topic News and Economy

	precision	recall	f1-score	support
Disaster and Accident	0.90	0.63	0.74	210
News and Economy	0.85	0.97	0.90	206
Travel & Transportation	0.69	0.82	0.75	184
micro avg	0.80	0.80	0.80	600
macro avg	0.81	0.81	0.80	600
weighted avg	0.82	0.80	0.80	600

actual class	Disaster and Accident	News and Economy	Travel & Transportation
cluster			
0	70	0	145
1	131	13	25
2	9	193	14

Cluster 0: Topic Travel & Transportation
Cluster 1: Topic Disaster and Accident
Cluster 2: Topic News and Economy

	precision	recall	f1-score	support
Disaster and Accident	0.78	0.62	0.69	210
News and Economy	0.89	0.94	0.91	206
Travel & Transportation	0.67	0.79	0.73	184
micro avg	0.78	0.78	0.78	600
macro avg	0.78	0.78	0.78	600
weighted avg	0.78	0.78	0.78	600

actual class	Disaster and Accident	News and Economy	Travel & Transportation
cluster			
0	65	0	133
1	140	8	23
2	5	198	28

Cluster 0: Topic Travel & Transportation
Cluster 1: Topic Disaster and Accident
Cluster 2: Topic News and Economy

	precision	recall	f1-score	support
Disaster and Accident	0.82	0.67	0.73	210
News and Economy	0.86	0.96	0.91	206
Travel & Transportation	0.67	0.72	0.70	184
micro avg	0.79	0.79	0.79	600
macro avg	0.78	0.78	0.78	600
weighted avg	0.79	0.79	0.78	600

- Above figure shows results for Q1(K-Means) when running 4 times with same parameters.
- As we can notice in every run, classification report/metrics changes. This tells us that Kmeans clustering is not that much robust algorithm as compared to LDA. Because, final clusters pretty much depend on initialization of K-centroids.
- When I ran Q2(LDA) multiple times with same parameters, I am getting same results only.
- Hence, for topic distribution task LDA proves to be better than K-Means clustering

Hyperparameter tuning:

- **min_df:** We are trying to find clusters of topic/ topic distribution for given question. Hence, to ignore the rare words we need to set this parameter. I chose to remove words which appear in less than 5 documents.
- **Stop words:** removing stop words is necessary in topic classification problem, otherwise some of these words might get considered as topic while classifying.
- **iterations:** Normally iterations/epochs should stop when there is no change in assigned topics/topic distribution. For given problem, I chose 25 iterations in both questions, which gives me good overall classification of topics.
- **max_df:** This parameter allows us to control the words which appear too frequently in overall corpus of documents. By setting it to 90% (0.9), I am removing words which appear in more than 90% documents. As these words appear in almost every document, they won't add much value to final classification.