Q2:Performance comparison between Q1 and Q2: actual class Disaster and Accident News and Economy Travel & Transportation

actual class Disaster an	d Accident	News and	Economy	Travel & Transporta	tion
cluster					
0	39		5		127
1	160		5		46
2	11		196		11
Cluster 0: Topic Travel &					
Cluster 1: Topic Disaster		nt			
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 an	d Accident	News and	Economy	Travel & Transporta	tion
cluster		News and		Travel & Transporta	
	68	News and	Economy 0 7	Travel & Transporta	151
cluster 0	68 132	News and	0 7	Travel & Transporta	151 8
cluster 0 1 2	68 132 10		0	Travel & Transporta	151
<pre>cluster 0 1 2 Cluster 0: Topic Travel &</pre>	68 132 10 Transporta	tion	0 7	Travel & Transporta	151 8
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster	68 132 10 Transporta and Accide	tion	0 7	Travel & Transporta	151 8
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and	68 132 10 Transporta and Accide	tion nt	0 7		151 8
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and	68 132 10 Transporta and Accide Economy	tion nt	0 7 199	support	151 8
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster and Accident	68 132 10 Transporta and Accide Economy precision	tion nt recall	0 7 199 f1-score	support	151 8
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and	68 132 10 Transporta and Accide Economy precision 0.90	tion nt recall 0.63	0 7 199 f1-score 0.74	support 210	151 8
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster and Accident News and Economy Travel & Transportation	68 132 10 Transporta and Accide Economy precision 0.90 0.85	tion nt recall 0.63 0.97	0 7 199 f1-score 0.74 0.90	support 210 206	151 8
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster and Accident News and Economy	68 132 10 Transporta and Accide Economy precision 0.90 0.85 0.69	tion nt recall 0.63 0.97 0.82	0 7 199 f1-score 0.74 0.90 0.75	support 210 206 184	151 8
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster and Accident News and Economy Travel & Transportation micro avg	68 132 10 Transporta and Accide Economy precision 0.90 0.85 0.69	tion nt recall 0.63 0.97 0.82	0 7 199 f1-score 0.74 0.90 0.75	support 210 206 184 600	151 8

actual class Disaster and	Accident	News and	Economy	Travel & Tr	ansportation
cluster	70				4.45
0	70		0		145
1 2	131		13		25
The same of the sa	9		193		14
Cluster 0: Topic Travel &					
Cluster 1: Topic Disaster a Cluster 2: Topic News and I		nc			
	recision	nocoll	f1-score	support	
Pi	recision	recall	T1-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		
weighted avg	0.78	0.78	0.78		
actual class Disaster and	Accident	News and	Economy	Travel & Tr	ansportation
actual class Disaster and cluster		News and	Economy	Travel & Tr	ansportation
cluster 0	65	News and	0	Travel & Tr	133
cluster 0 1	65 140	News and	0 8	Travel & Tr	133 23
cluster 0 1 2	65 140 5		0	Travel & Tr	133 23
cluster 0 1 2 Cluster 0: Topic Travel &	65 140 5 Transporta	tion	0 8	Travel & Tr	133 23
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster	65 140 5 Transporta and Accide	tion	0 8	Travel & Tr	133 23
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster 2: Topic News And Disas	65 140 5 Transporta and Accide Economy	tion nt	0 8 198		133 23
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster 2: Topic News And Disas	65 140 5 Transporta and Accide	tion nt	0 8		133 23
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster 2: Topic News And Disas	65 140 5 Transporta and Accide Economy	tion nt	0 8 198	support	eansportation 133 23 28
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and P	65 140 5 Transporta and Accide Economy recision	tion nt recall	0 8 198 f1-score	support 210	133 23
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster and Accident	65 140 5 Transporta and Accide Economy recision 0.82	tion nt recall 0.67	0 8 198 f1-score 0.73	support 210 206	133 23
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster at Cluster 2: Topic News and Disaster at Cluster and Accident News and Economy Travel & Transportation	65 140 5 Transporta and Accide Economy recision 0.82 0.86	tion nt recall 0.67 0.96	0 8 198 f1-score 0.73 0.91	support 210 206 184	133 23
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster Cluster 2: Topic News and Disaster and Accident News and Economy	65 140 5 Transporta and Accide Economy recision 0.82 0.86 0.67	tion nt recall 0.67 0.96 0.72	0 8 198 f1-score 0.73 0.91 0.70	support 210 206 184 600	133 23
cluster 0 1 2 Cluster 0: Topic Travel & Cluster 1: Topic Disaster at Cluster 2: Topic News and I property of the Cluster and Accident News and Economy Travel & Transportation micro avg	65 140 5 Transporta and Accide Economy recision 0.82 0.86 0.67	tion nt recall 0.67 0.96 0.72 0.79	0 8 198 f1-score 0.73 0.91 0.70	support 210 206 184 600	133 23

- 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.