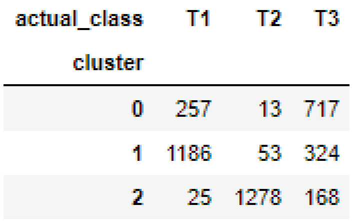
Assignment 7 Analysis:

Task 1) K-Mean Clustering

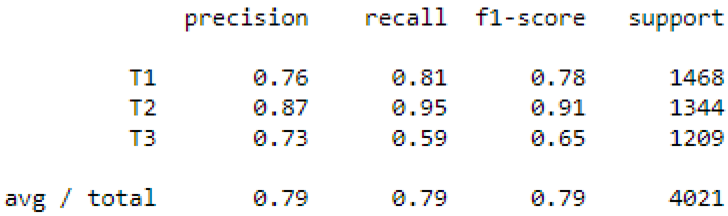
Results:

Confusion Matrix:



actual_class	T1	T2	T3
cluster			
0	257	13	717
1	1186	53	324
2	25	1278	168

Classification Report:



```
             precision    recall  f1-score   support

         T1       0.76      0.81      0.78      1468
         T2       0.87      0.95      0.91      1344
         T3       0.73      0.59      0.65      1209

avg / total       0.79      0.79      0.79      4021
```

Based on cluster centroids/samples, give a meaningful name (instead of T1, T2, T3) to each cluster.

Top words for each cluster are as follows:

Cluster 0: said; crash; bus; rail; plane; train; police; passengers; car; cruise; airlines; speed; flight; road; says; driver; traffic; accident; airport; people

Cluster 1: said; oil; people; bp; japan; water; spill; gulf; disaster; nuclear; earthquake; pakistan; tsunami; quake; floods; coast; million; plant; government; killed

Cluster 2: percent; tax; said; year; economy; rate; obama; comment; government; economic; billion; budget; debt; bank; growth; new; market; spending; jobs; report

Cluster Names:

Cluster 0 (T3): Travel

Cluster 1 (T1): Bad News Natural and Manmade

Cluster 2 (T2): Finance

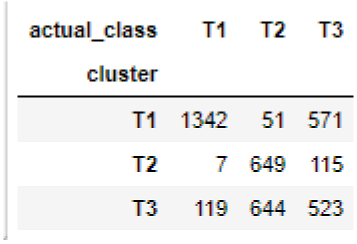
Task 2:

1) Print out: Feature Names and Shapes:

['abandoned', 'abc', 'ability', 'able', 'aboard', 'abroad', 'absolutely', 'abuse', 'accept', 'access']

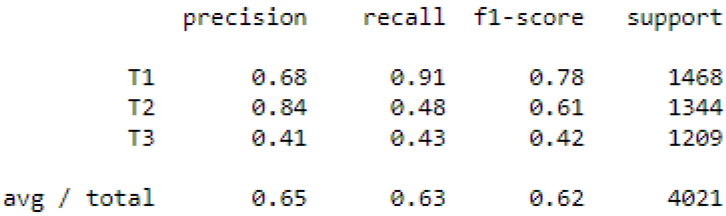
(4021, 2749)

2) Confusion Matrix:



actual_class	T1	T2	T3
cluster			
T1	1342	51	571
T2	7	649	115
T3	119	644	523

3) Classification Report:



```
             precision    recall  f1-score   support

         T1       0.68      0.91      0.78      1468
         T2       0.84      0.48      0.61      1344
         T3       0.41      0.43      0.42      1209

avg / total       0.65      0.63      0.62      4021
```

4) Display Topics:

Topic 0:

said people comment oil news sign users water japan police city bp officials report killed says new told year crash hit time government rate coast plane disaster world nuclear area state just gulf passengers spill million video miles ap earthquake

Topic 1:

percent said year economy rate market economic government growth bank billion new debt report china financial prices month week global news unemployment jobs recession markets fed banks crisis data rates time quarter months expected recovery world european finance airlines central

Topic 2:

said tax new obama year com million state president rail high house money budget federal government billion business years time www people plan service security speed health information make public travel program work pay taxes cruise services help company spending

5) Cluster Names:

Topic 0: Accidents and Disaster

Topic 1: Global Finance

Topic 2 (T2): Government Topics

Analysis:

For me, I am getting a higher accuracy for K-Mean than LDA.

But if we increase clusters than the accuracy will increase for both the models, I assume that.