Answer 1

Part A:

1. I tried the grid search first and the C is in a wide range as [0.001,0.003,0.01,0.03,0.1,0.3,1,3,10,30,100,300,1000] because I am not sure the exact range of C should be in while too large C may cause overfitting and too little C may cause the oppsite effect.Since the the accuracy function of C should be continuous, it may be reasonable run iteration to pick smaller domain around the appropriate grid picked in the last iteration to get better performance in some circumstance.

2. Best C is 6

3. Accuracy on test set is 0.9058

PartB

1. I chose to tune the vector size and the window size.Similar to A, I used grid search but without shrinking the domain.The vector size is in range of [5,20,100,200] and the window size is in range of [1,5,10,50].

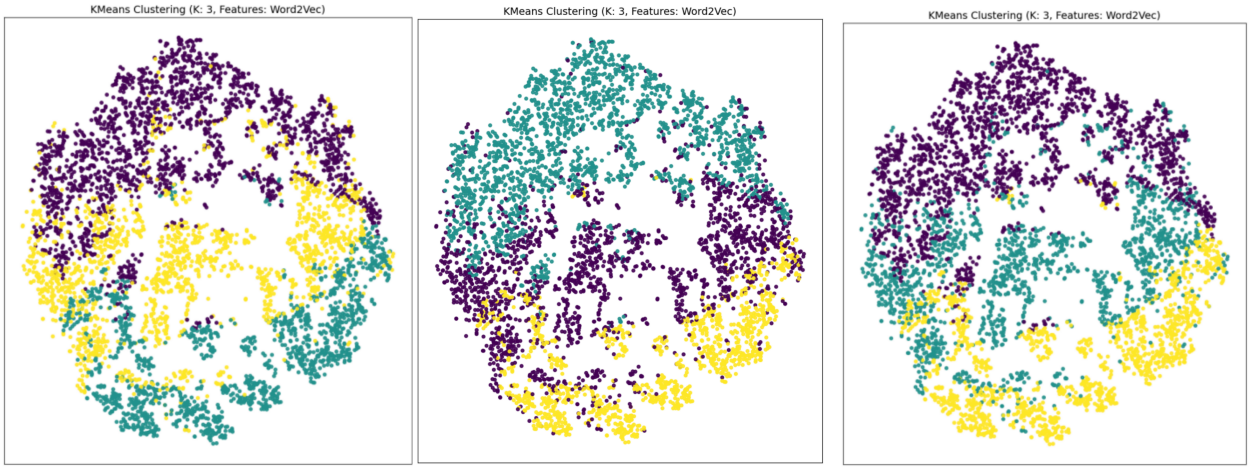
2. Best vector size and window size are 200 and 50.Since the larger size include the information of smaller size, it should perfom better.

3. Best C for (2) is 30.Accuracy on test set (word2vec): 0.8689

Answer 2

1. *TF-IDF & Word2Vec*:
I have tried K=2,3,5,10,100.And found 100 may be too big and hard to distingusih.From my view,the hidden sturcture of word cluster may means the pattern of people express with the same words or phrases, so the K may be small which means 2,3,5,10 may be appropriate.

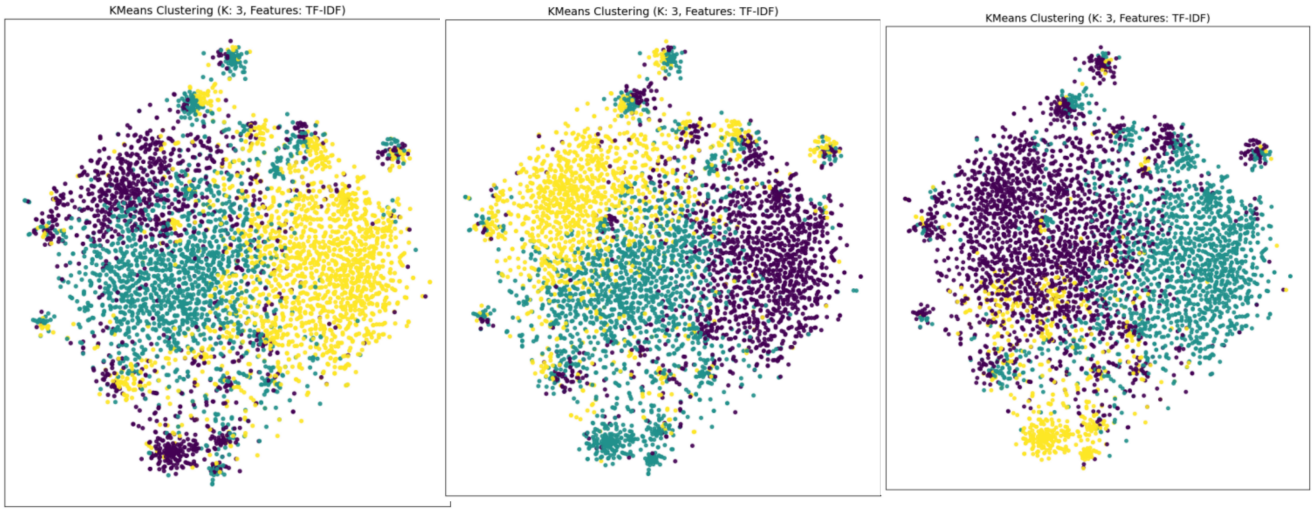
2. Word2Vec:
I randomly choose 5000 of the 50000 data.From the picture shown below ,for a particular K=3,the pattern or the distrubuation of the same cluster do not change except the color which is random.
So I think the seed did not affect the clusters especially after many iteration the distance should converge to a fix value.If the distance do not converage after a fixed iteration numbers the cluster also show some similar results



which means seeds matters the classification of cluster little.

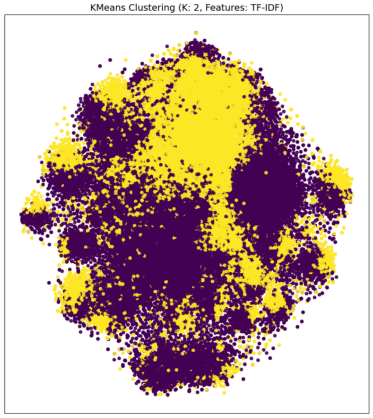
However, when I test the clusters, I found different seeds may affect the numbers of iteration for the distance to converage which means how fast it coverages or iteration stops before the max of iteration we input.It is reasonable because for some points chosen initally are good while others are bad.

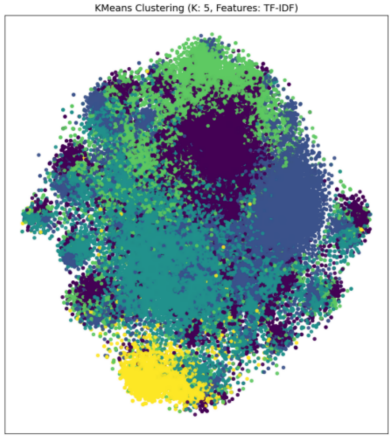
TF-IDF:

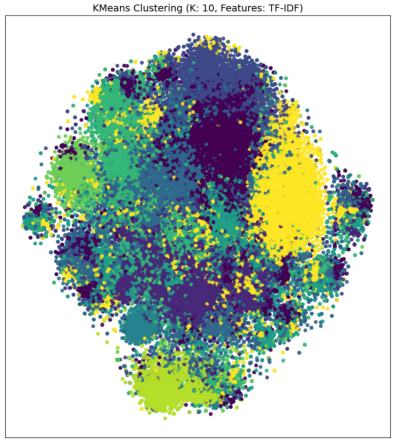


The same results and conclusion as shown in Word2Vec from the picture above.

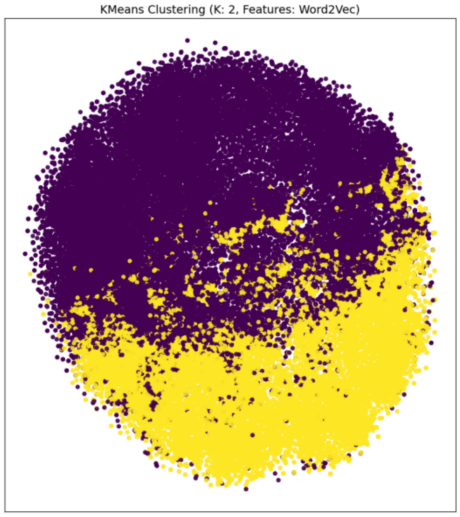
3. I have tried the tf-idf on the whole data sets with 2,5,10,100 and found 5 may be the best choice.According to the distance function applied, the visulation of clusters should show picture like a pie chart.I think for the tfidf ,the K=5 perform the best to distinguish and show most likely to a pie chart.







When it comes to Word2Vec the K=2 perform well.It seems K =2 is appriopriate



Answer3:
Argmax: John Barder

Random:
Asiad Lone Hontk
Jok Loviolamy
Gram Horzson
Jan Prock
Danilie Wery
Borid Bewman
Miot Wae Jofy
Nytom Tarellin
Olfer Sanjire
Kmohmen Tur