TEXT CLASSIFICATION ASSIGNMENT 2

Team 10

Ahmed Abdo Amin Abdo

Ahmed Abd Ellatife Mohamed Salah Eldine

Alaa Tohamy Mohamed AbdElwahab

Sherif Mohamed Abdelaziz Ahmed

1) Preparing the data

we have chosen 5 books from Gutenberg NLTK library

1- "austen-emma.txt",
2- "milton-paradise.txt",
3- "chesterton-brown.txt",
4- "shakespeare-caesar.txt",
5- "carroll-alice.txt"

then we have chosen 200 random partitions of each book every partition is a list containing 150 words

2) Preprocessing data

we have removed stop words and all characters that are not important in the data and lowered all the words to be efficient in training and testing

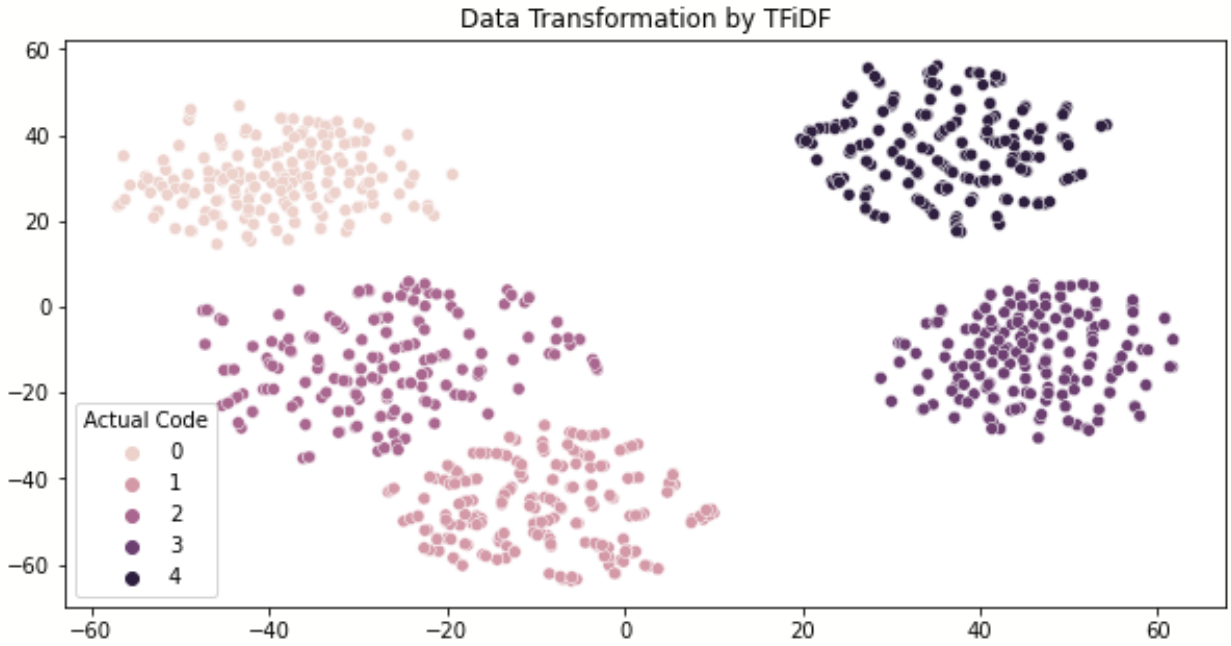
arranged books partitions, book name, and author name in a data frame

3) Feature engineering

In this step, we have used BOW, TF-IDF, LDA, and Word-Embedding to transform the words into numeric values so the machine can understand easily and train it

BOW Data Transformation

TF-IDF Data Transformation

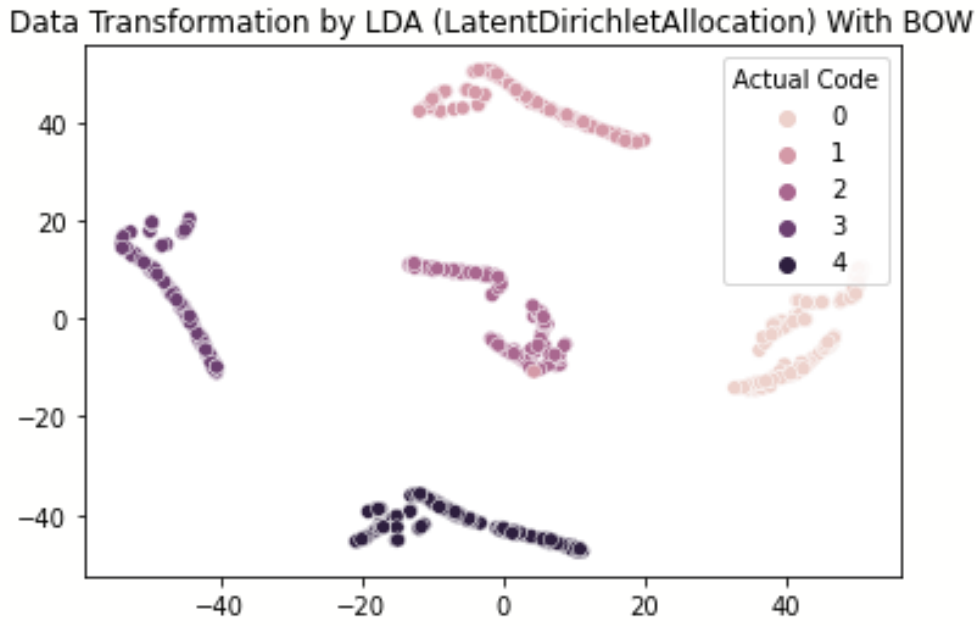


Feature Selection

we have used the feature selection method using "**SKlearn Chi-Square"** to select the best 10% Features

```
Before features reduction
BOW : (1000, 8535)
TFiDF : (1000, 12032)
After features reduction
BOW : (1000, 854)
TFiDF : (1000, 1204)
```

LDA With BOW Data Transformation



THE TOP 10 WORDS FOR TOPIC #0

['betray', 'afternoon', 'accept', 'bidding', 'addressed', 'approaching', 'befriend', 'admirest', 'adjoining', 'absurd']

THE TOP 10 WORDS FOR TOPIC #1

['bigge', 'ashamed', 'blush', 'ascending', 'bates', 'assert', 'betrayed', 'backgammon', 'abrupt', 'beauteous']

THE TOP 10 WORDS FOR TOPIC #2

['avenged', 'bewilderment', 'apply', 'beg', 'authority', 'aught', 'bosh', 'although', 'agent', 'authorised']

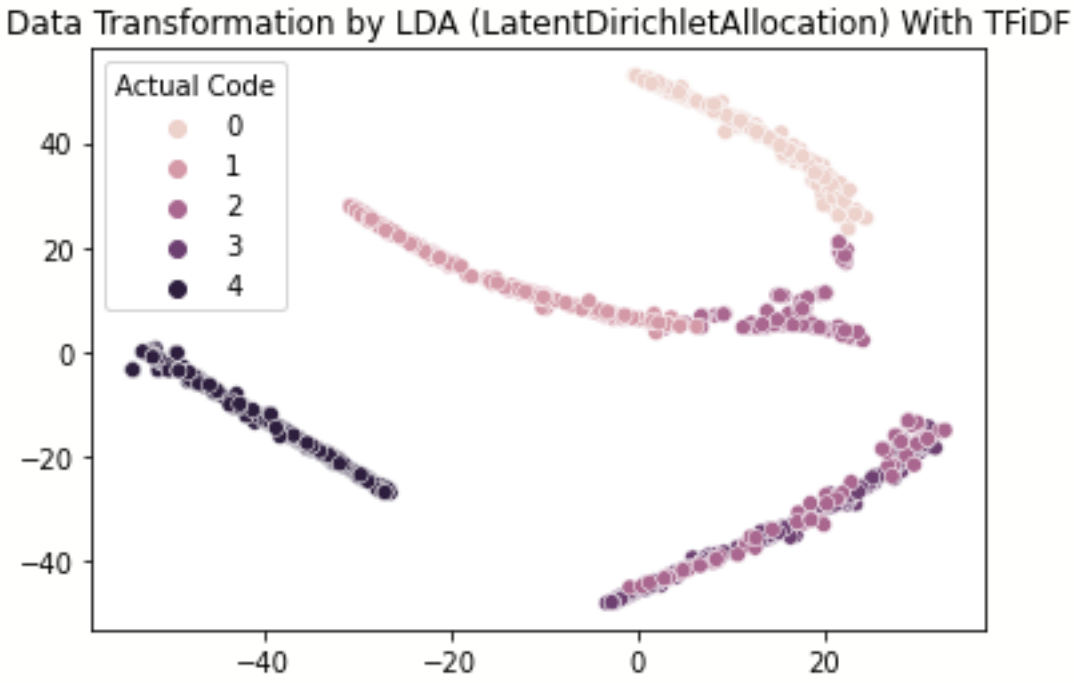
THE TOP 10 WORDS FOR TOPIC #3

['articulate', 'back', 'anarchist', 'address', 'backgammon', 'attachment', 'assented', 'approbation', 'betrayed', 'beauteous']

THE TOP 10 WORDS FOR TOPIC #4

['array', 'bin', 'better', 'betrayed', 'bind', 'arbitrator', 'bit', 'bidding', 'ayre', 'absurd']

LDA With TFIDF Data Transformation



THE TOP 10 WORDS FOR TOPIC #0

['audibly', 'admiral', 'absorbing', 'aule', 'achilles', 'ambitious', 'assisted', 'acquit', 'acknowledgement', 'absent']

THE TOP 10 WORDS FOR TOPIC #1

['auoyded', 'anointed', 'background', 'annoy', 'artimedorus', 'antonio', 'audience', 'argues', 'abomination', 'aspire']

THE TOP 10 WORDS FOR TOPIC #2

['approving', 'augur', 'amazed', 'assisting', 'appropriate', 'appointed', 'banker', 'affords', 'admissible', 'approbation']

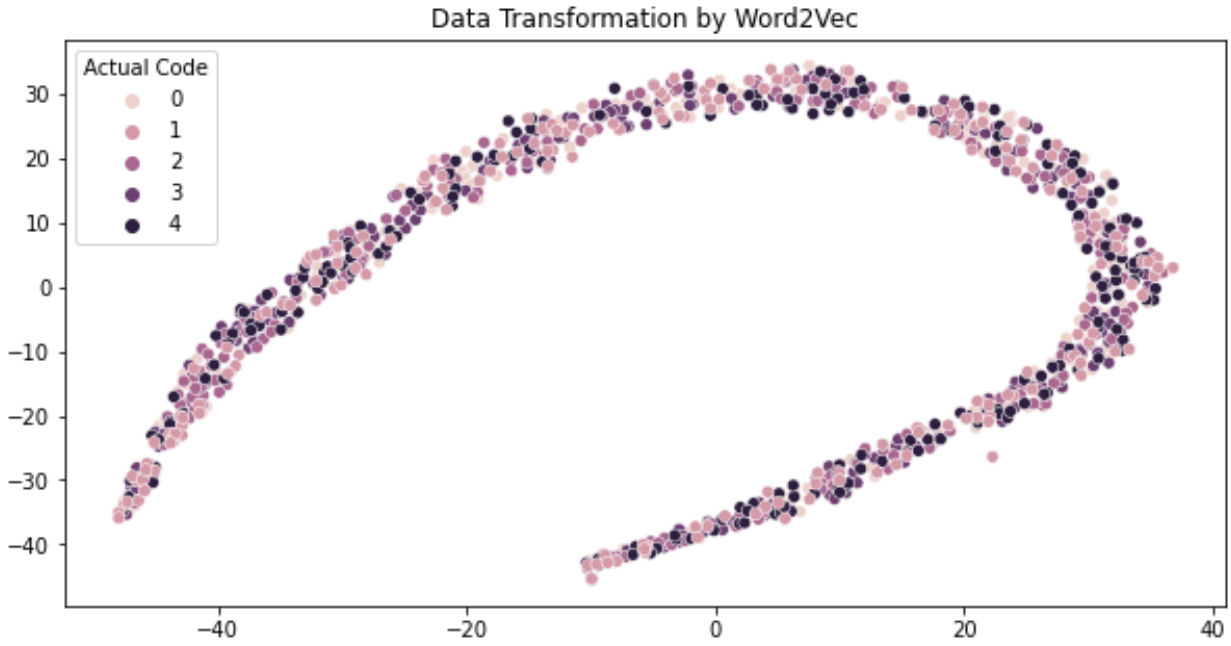
THE TOP 10 WORDS FOR TOPIC #3

['animal', 'argue', 'ahaz', 'achieving', 'argues', 'apparent', 'antonies', 'ambrosia', 'audience', 'aspire']

THE TOP 10 WORDS FOR TOPIC #4

['angelical', 'austere', 'auditress', 'audience', 'austerely', 'amid', 'aux', 'aule', 'ardent', 'absent']

Data transformation With Word2vec



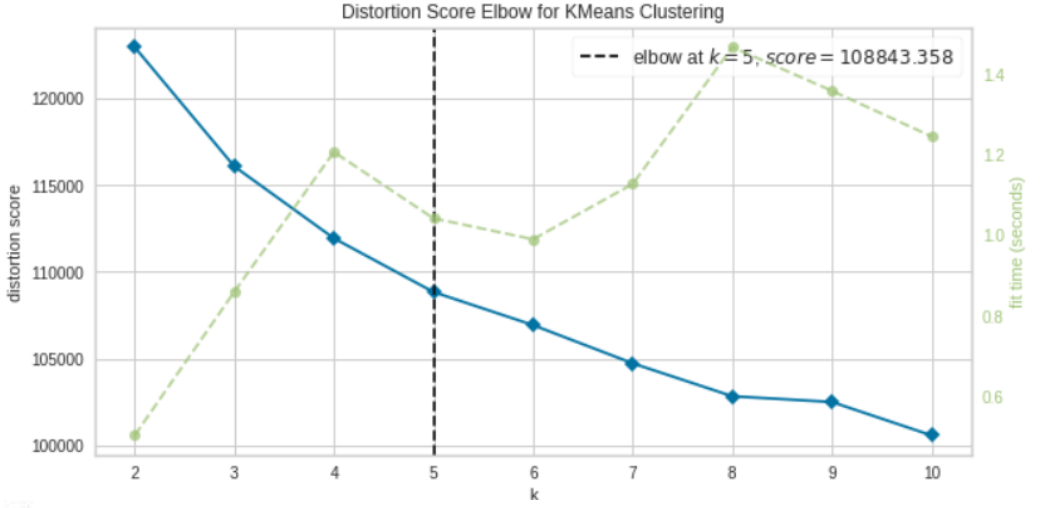
4) Evaluation

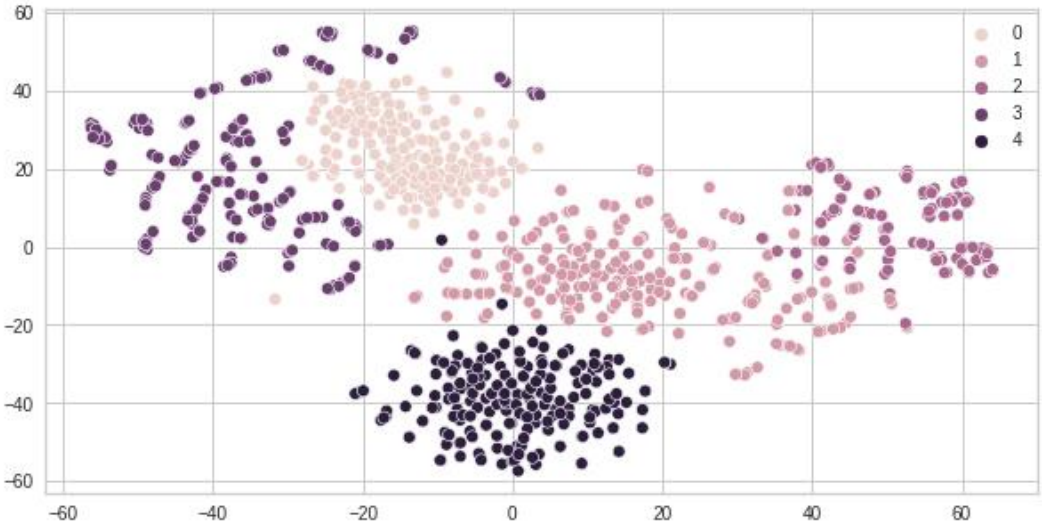
We used 3 algorithms with each transformer (K-means, EM, and Hierarchical clustering)

The result of these algorithms is:

1) K-Means On BOW

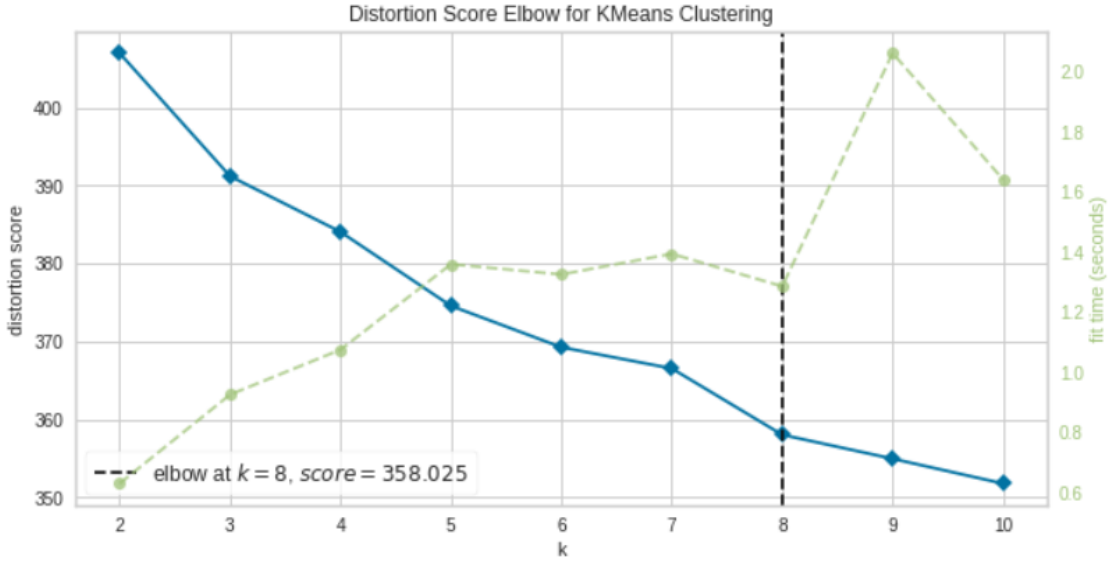
```
For n_clusters = 2 The average silhouette_score is : 0.1323738030297936
For n_clusters = 3 The average silhouette_score is : 0.06637812616380055
For n_clusters = 4 The average silhouette_score is : 0.041286579104121285
For n_clusters = 5 The average silhouette_score is : 0.05027128239892194
For n_clusters = 6 The average silhouette_score is : 0.048192967833504054
For n_clusters = 7 The average silhouette_score is : 0.04973225719541655
For n_clusters = 8 The average silhouette_score is : 0.04674520732938387
For n_clusters = 9 The average silhouette_score is : 0.055427536878161536
For n_clusters = 10 The average silhouette_score is : 0.049210312280807716
```

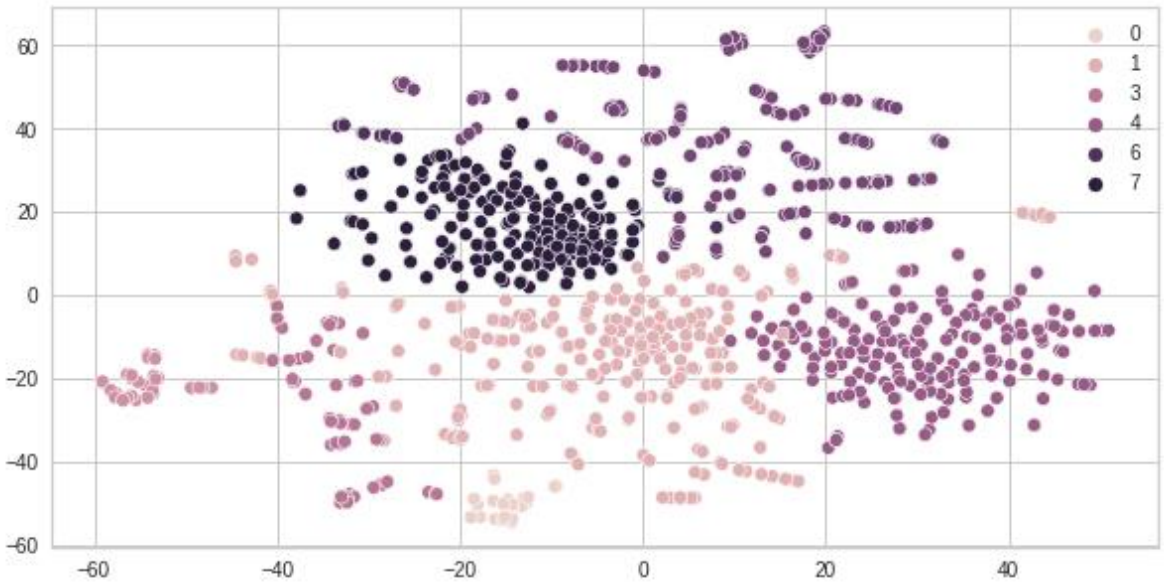




2) K-Means On TF-IDF

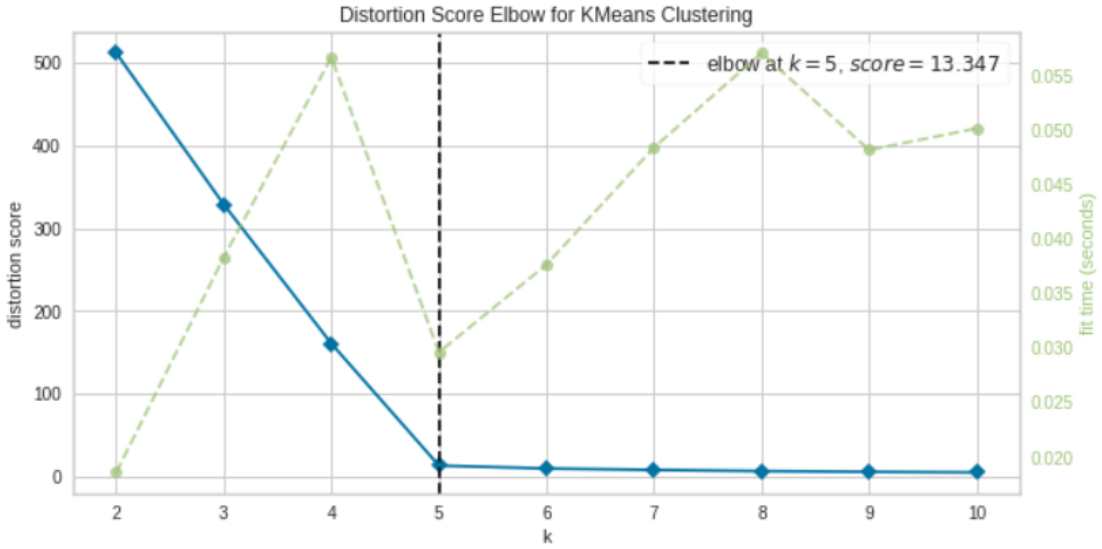
```
For n_clusters = 2 The average silhouette_score is : 0.04309632414758285
For n_clusters = 3 The average silhouette_score is : 0.08442764242986743
For n_clusters = 4 The average silhouette_score is : 0.09053717698492951
For n_clusters = 5 The average silhouette_score is : 0.02902303691776729
For n_clusters = 6 The average silhouette_score is : 0.034754286660432156
For n_clusters = 7 The average silhouette_score is : 0.025278297786776455
For n_clusters = 8 The average silhouette_score is : 0.03457015297009799
For n_clusters = 9 The average silhouette_score is : 0.035541874436295764
For n_clusters = 10 The average silhouette_score is : 0.045418487844285 48
```

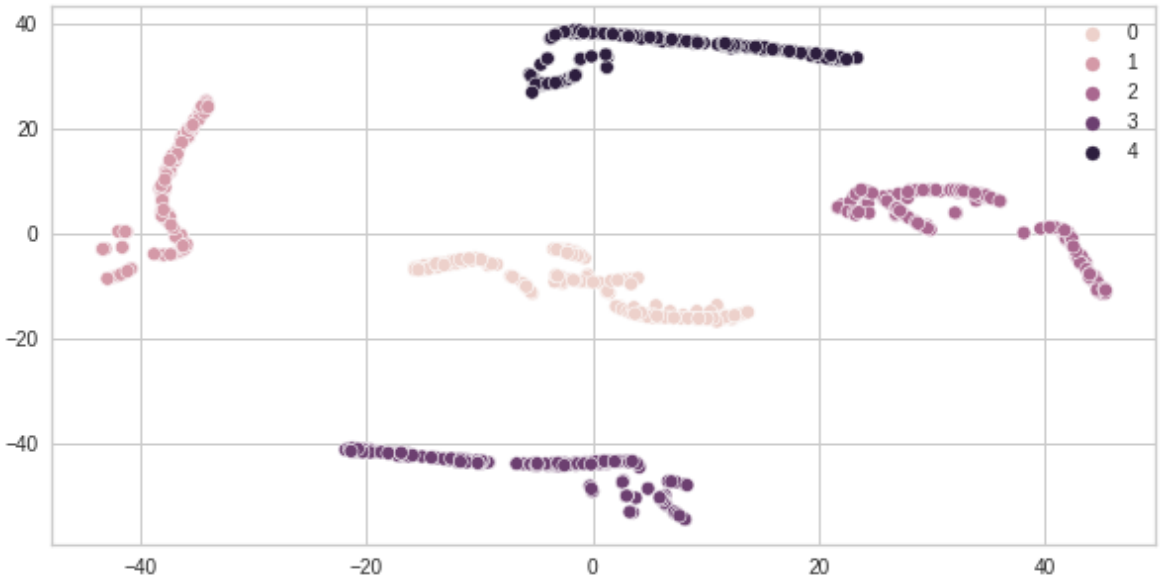




3) K-Means On BOW-LDA

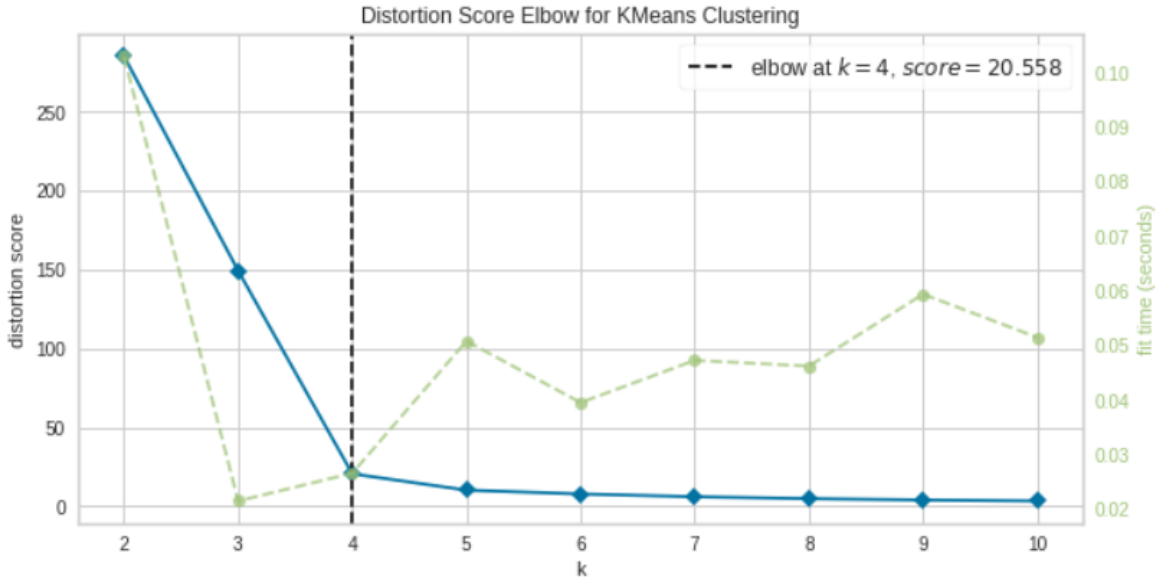
```
For n_clusters = 2 The average silhouette_score is : 0.3944806755479702
For n_clusters = 3 The average silhouette_score is : 0.5817413878021643
For n_clusters = 4 The average silhouette_score is : 0.750110704336471
For n_clusters = 5 The average silhouette_score is : 0.9106970812339511
For n_clusters = 6 The average silhouette_score is : 0.8665047030079577
For n_clusters = 7 The average silhouette_score is : 0.8134924249990083
For n_clusters = 8 The average silhouette_score is : 0.7971160490303818
For n_clusters = 9 The average silhouette_score is : 0.7993143122638098
For n_clusters = 10 The average silhouette_score is : 0.7918492701848875
```

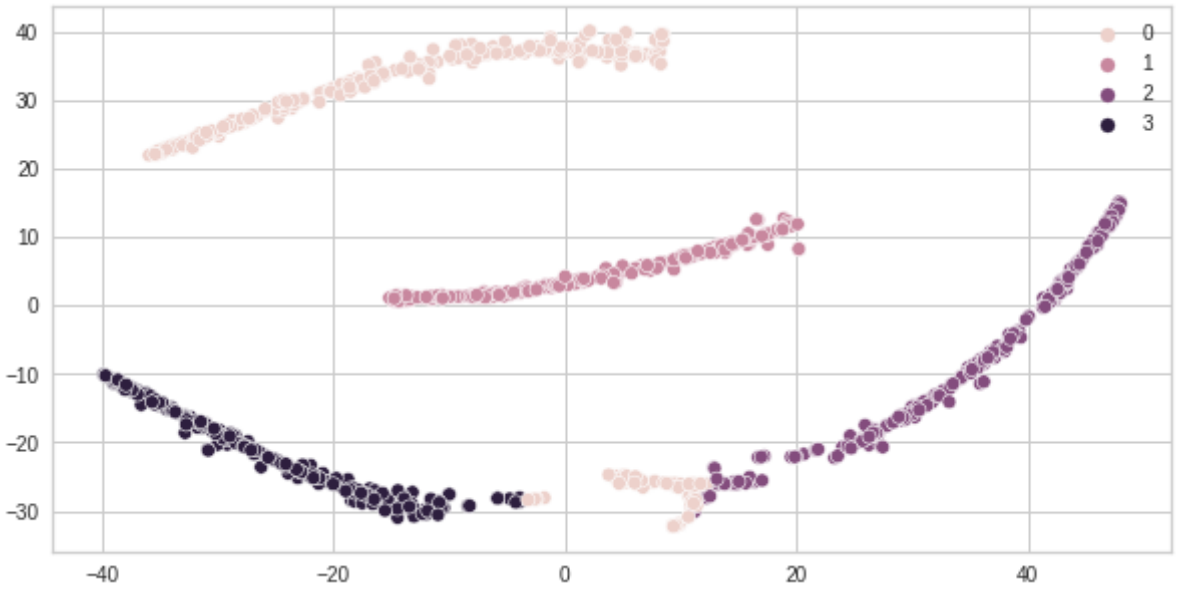




4) K-Means On TFIDF-LDA

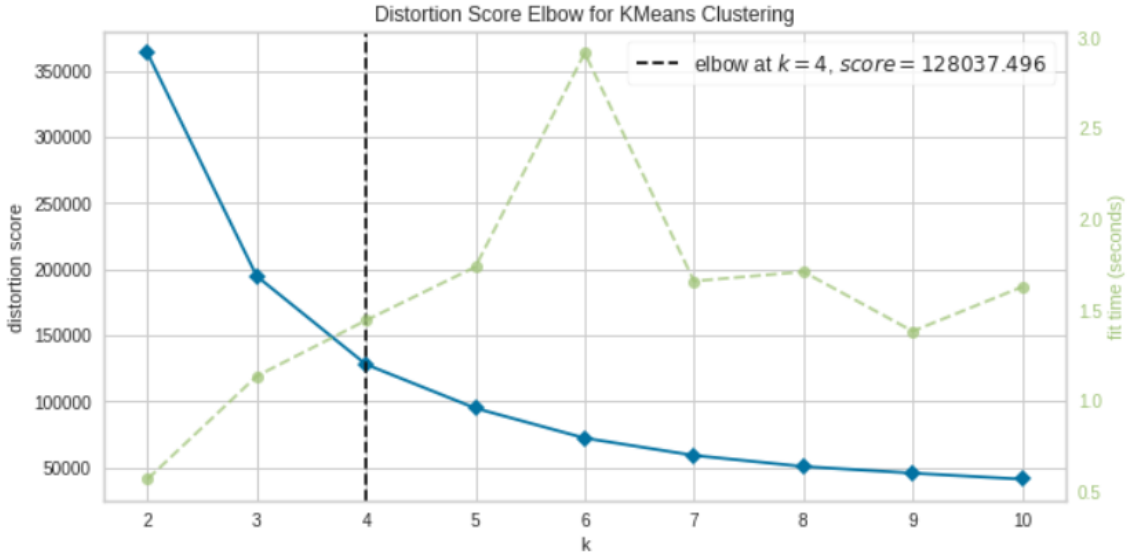
```
For n_clusters = 2 The average silhouette_score is : 0.44498025820376175
For n_clusters = 3 The average silhouette_score is : 0.6624085131522006
For n_clusters = 4 The average silhouette_score is : 0.8673328508347292
For n_clusters = 5 The average silhouette_score is : 0.8571208443344636
For n_clusters = 6 The average silhouette_score is : 0.8470742874773066
For n_clusters = 7 The average silhouette_score is : 0.8530455160076854
For n_clusters = 8 The average silhouette_score is : 0.85448520206955
For n_clusters = 9 The average silhouette_score is : 0.8569948138111622
For n_clusters = 10 The average silhouette_score is : 0.8508996532312137
```

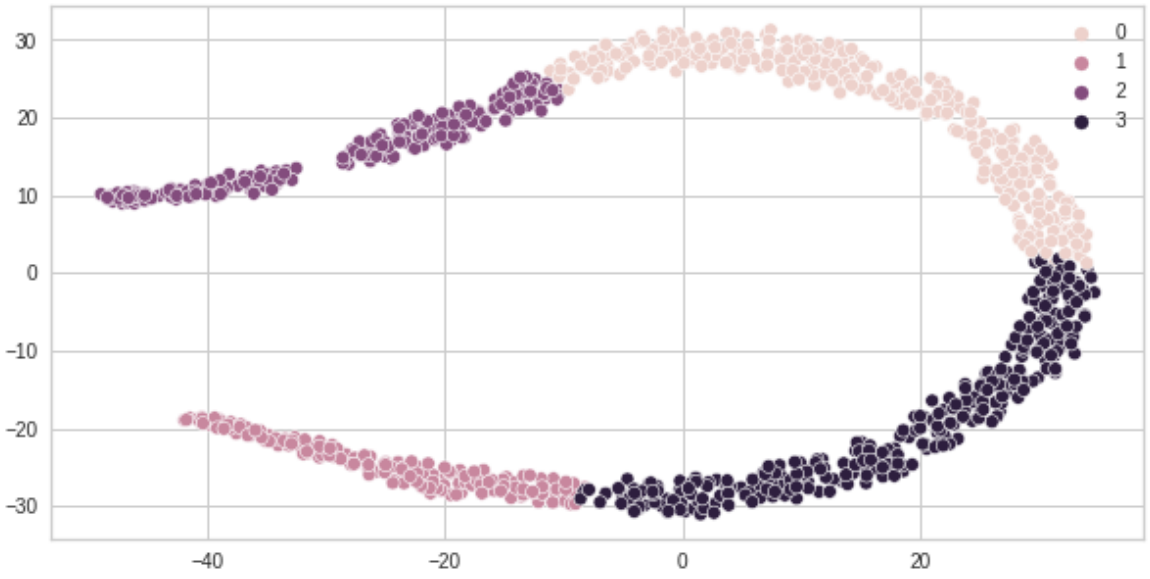




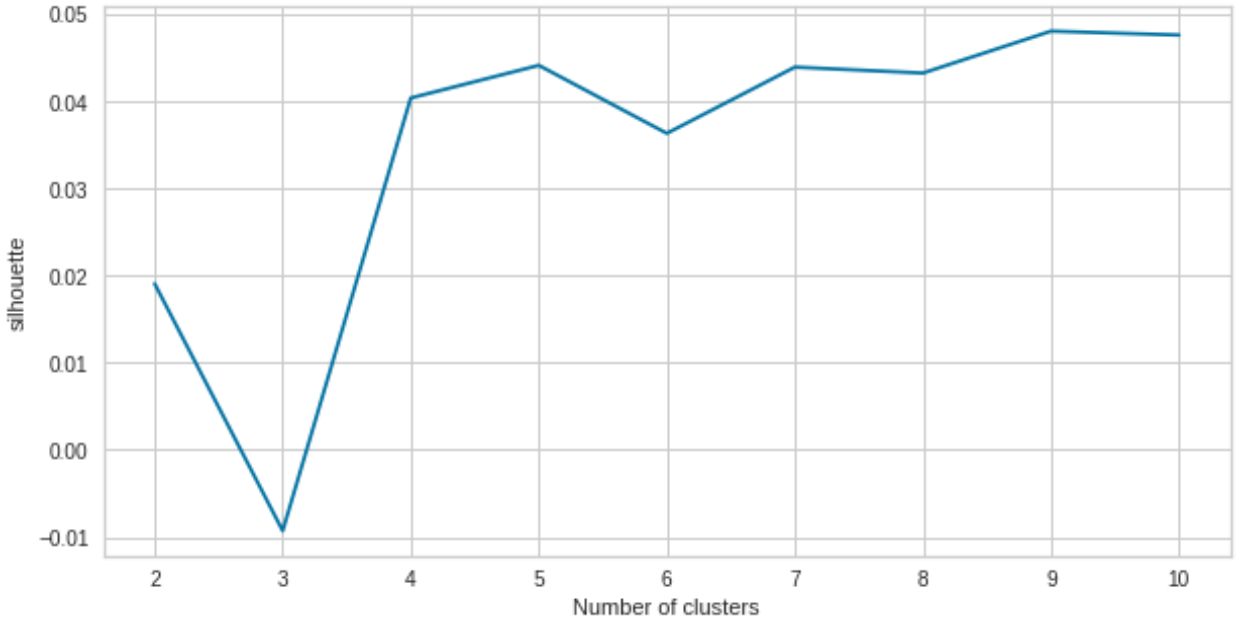
5) K-Means on Word2Vec

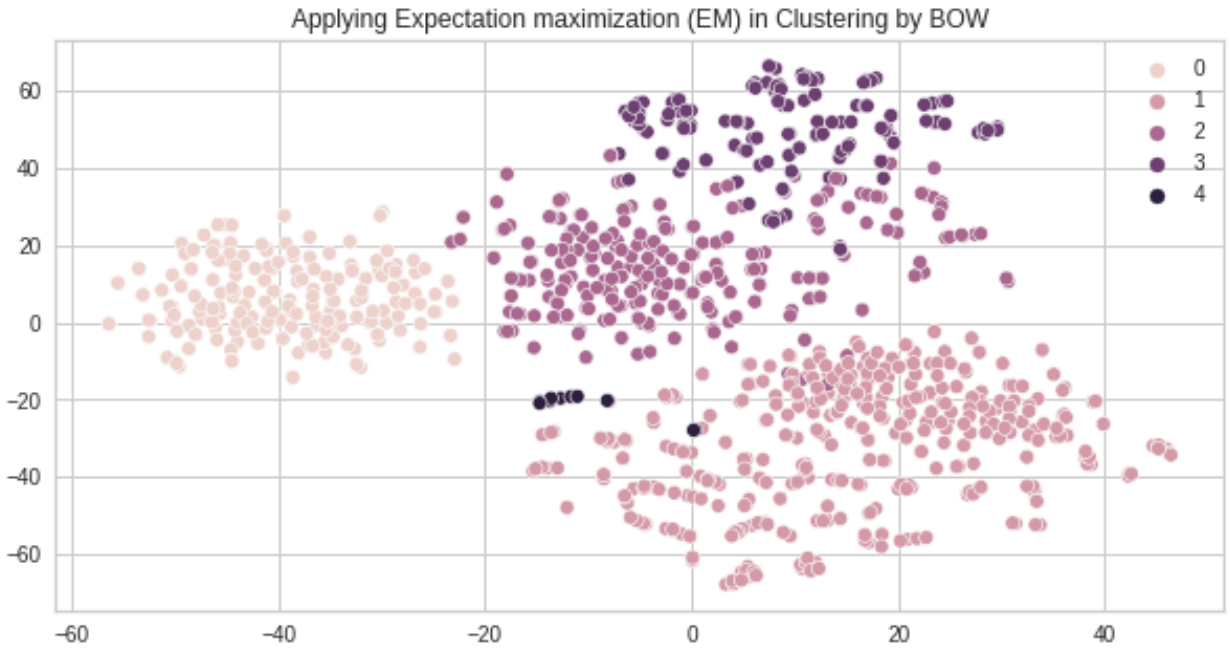
```
For n_clusters = 2 The average silhouette_score is : 0.5253677
For n_clusters = 3 The average silhouette_score is : 0.48340592
For n_clusters = 4 The average silhouette_score is : 0.446919
For n_clusters = 5 The average silhouette_score is : 0.42204693
For n_clusters = 6 The average silhouette_score is : 0.40839303
For n_clusters = 7 The average silhouette_score is : 0.38578594
For n_clusters = 8 The average silhouette_score is : 0.3587154
For n_clusters = 9 The average silhouette_score is : 0.34341392
For n_clusters = 10 The average silhouette_score is : 0.31744805
```



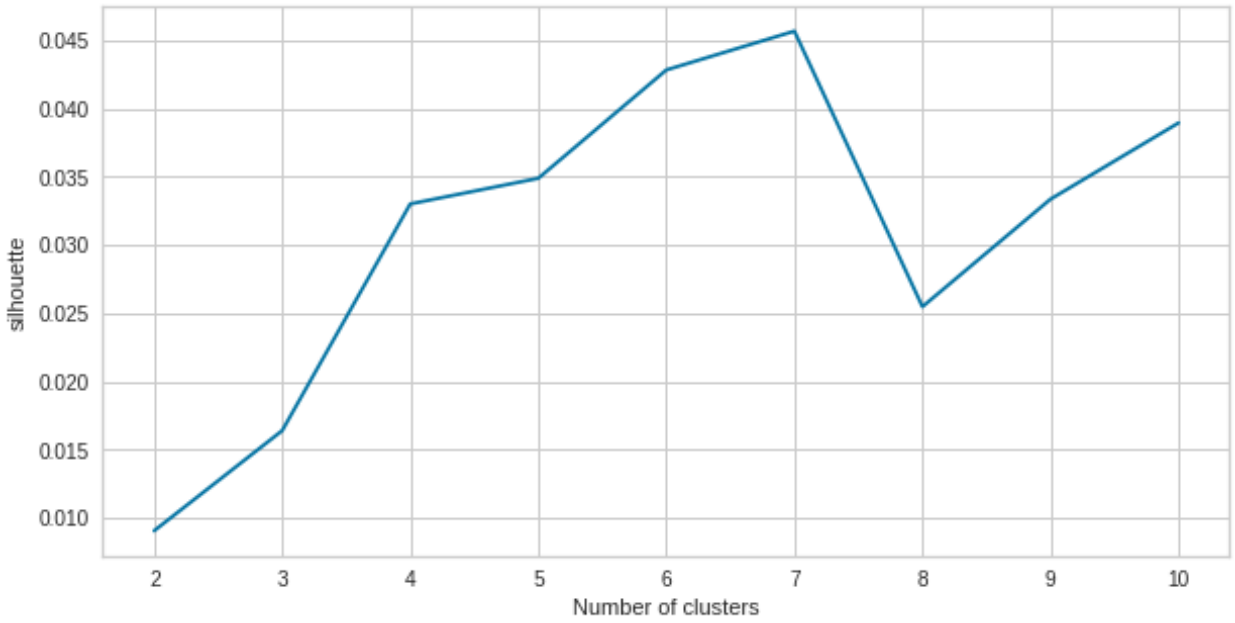


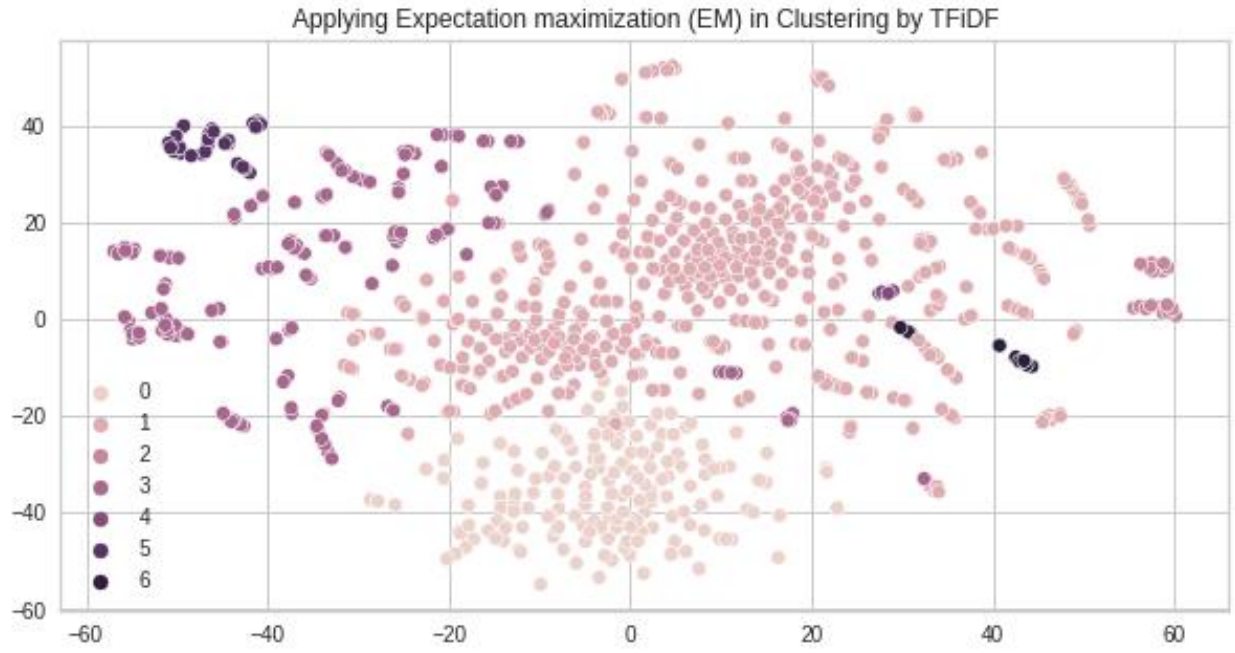
6) EM on BOW


0.05
0.04
0.03
0.02
0.01
0.00
−0.01
silhouette
2
3
4
5
6
7
8
9
10
Number of clusters

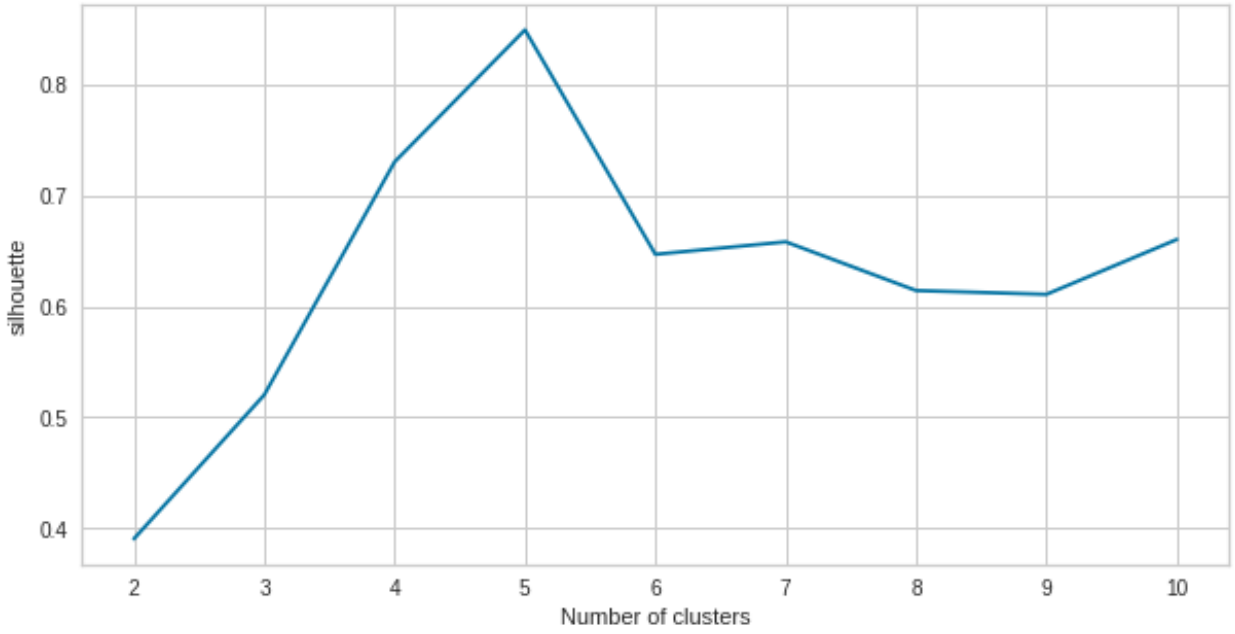

Applying Expectation maximization (EM) in Clustering by BOW
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
0
1
2
3
4

7) EM on TF-IDF

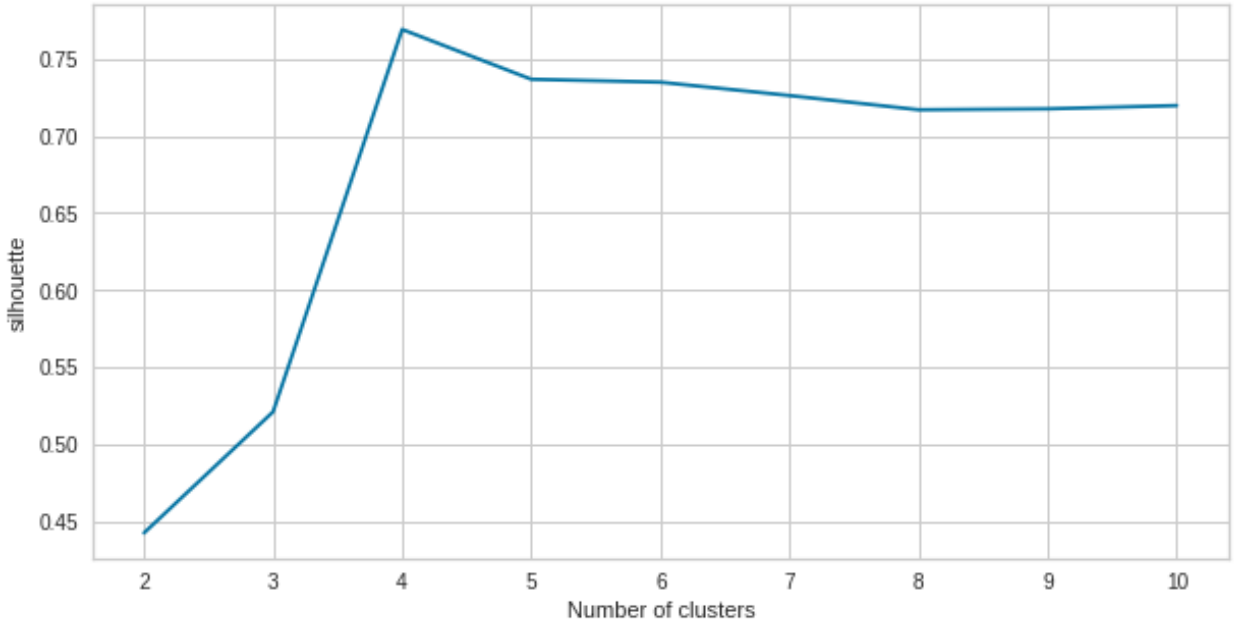

0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
silhouette
2
3
4
5
6
7
8
9
10
Number of clusters

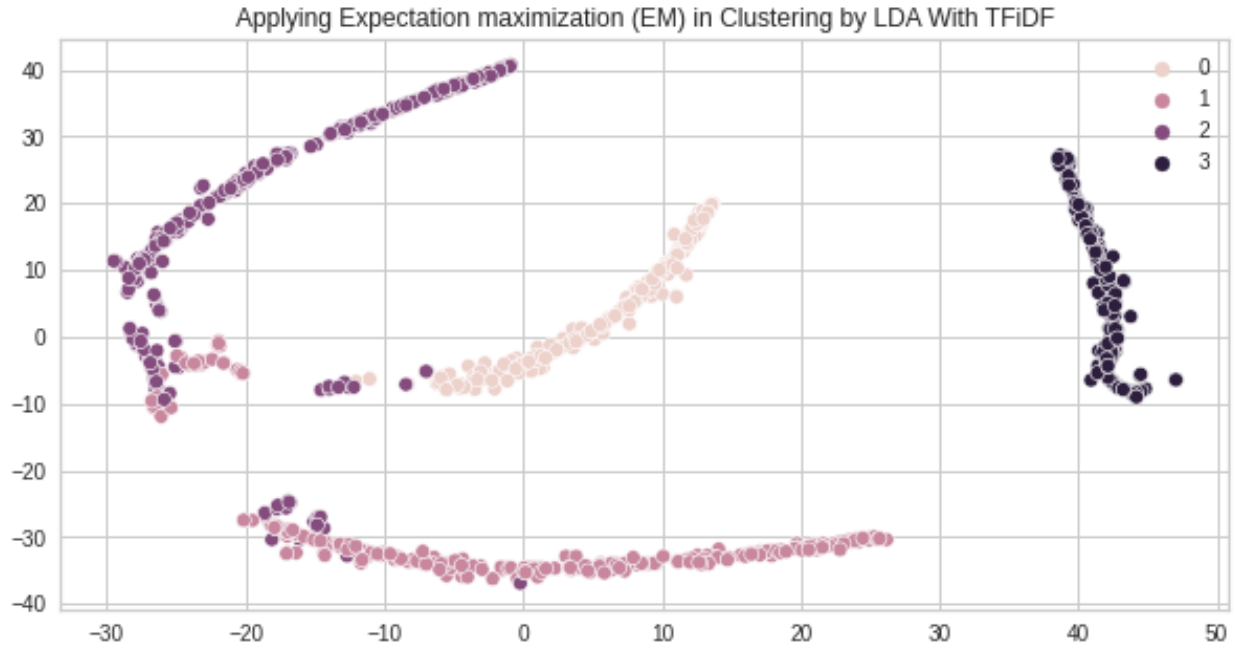

Applying Expectation maximization (EM) in Clustering by TFiDF
40
20
0
−20
−40
−60
0
1
2
3
4
5
6
−60
−40
−20
0
20
40
60

8) EM on BOW-LDA

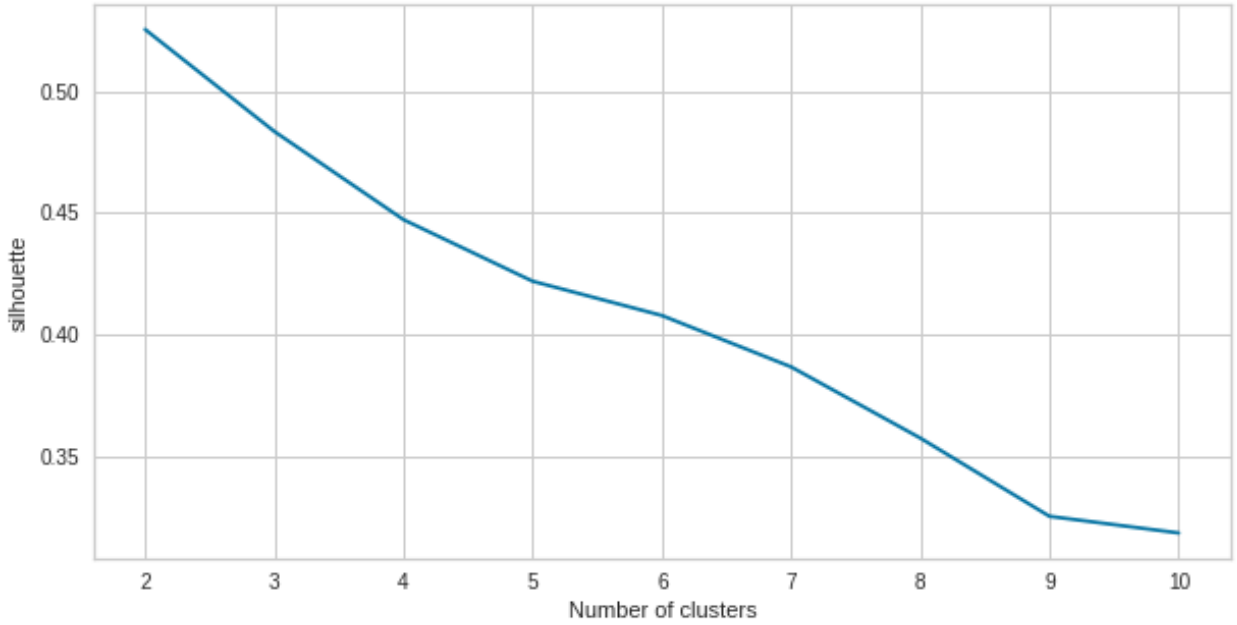


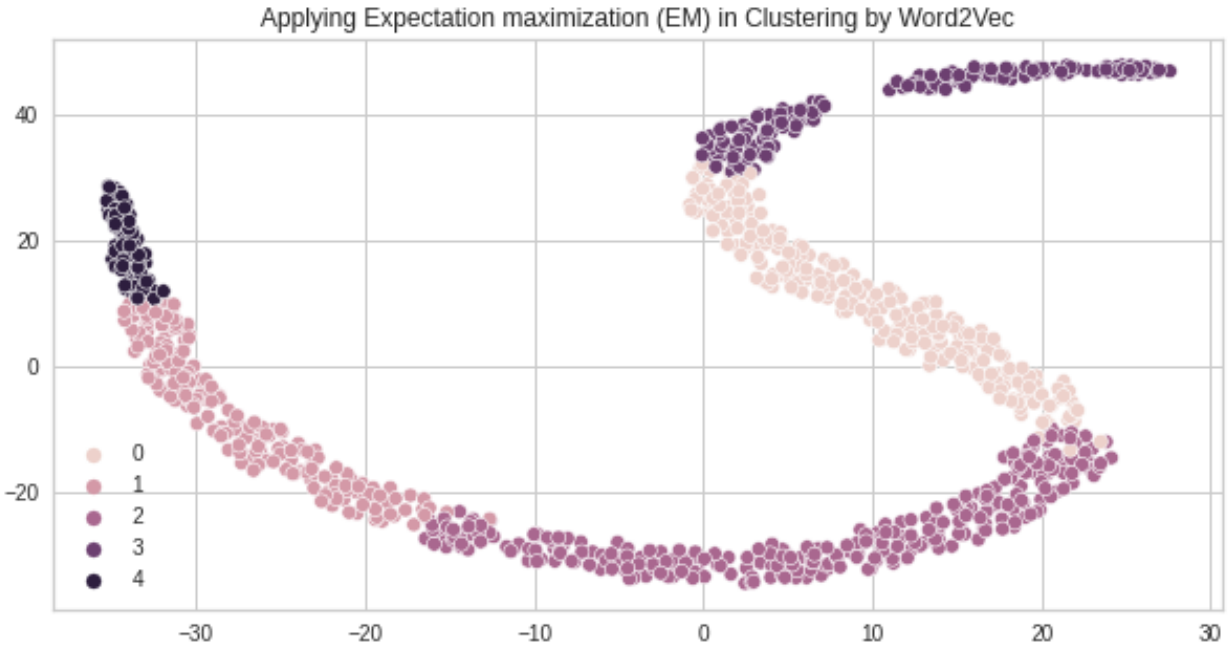
9) EM on TFIDF-LDA





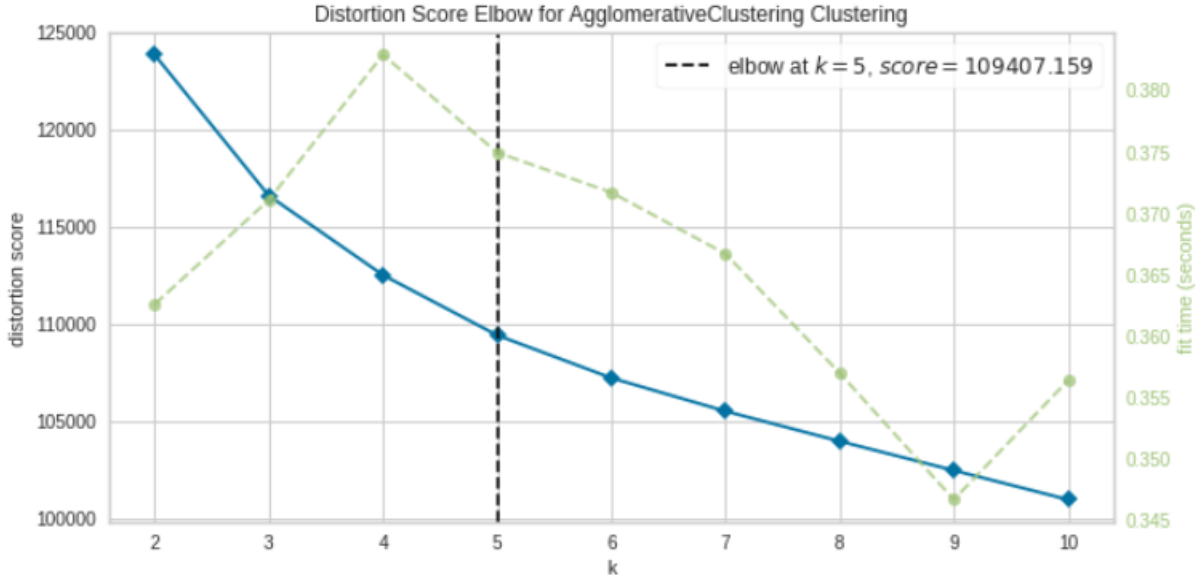
10) EM on Word2Vec

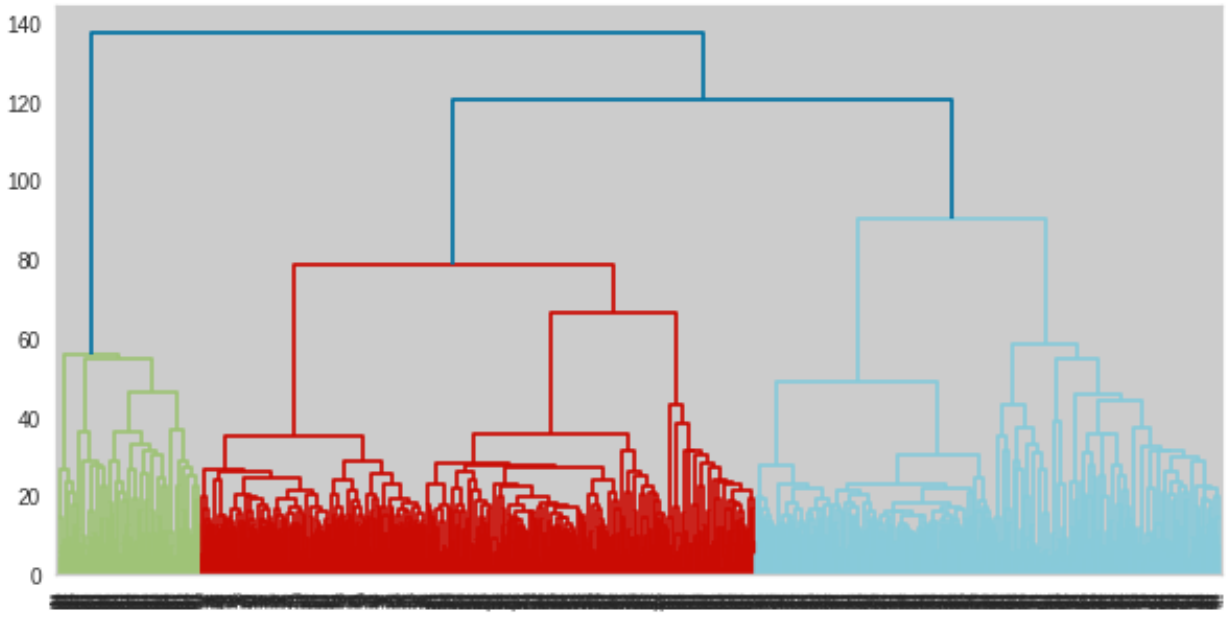




11) Hierarchical Clustering on BOW

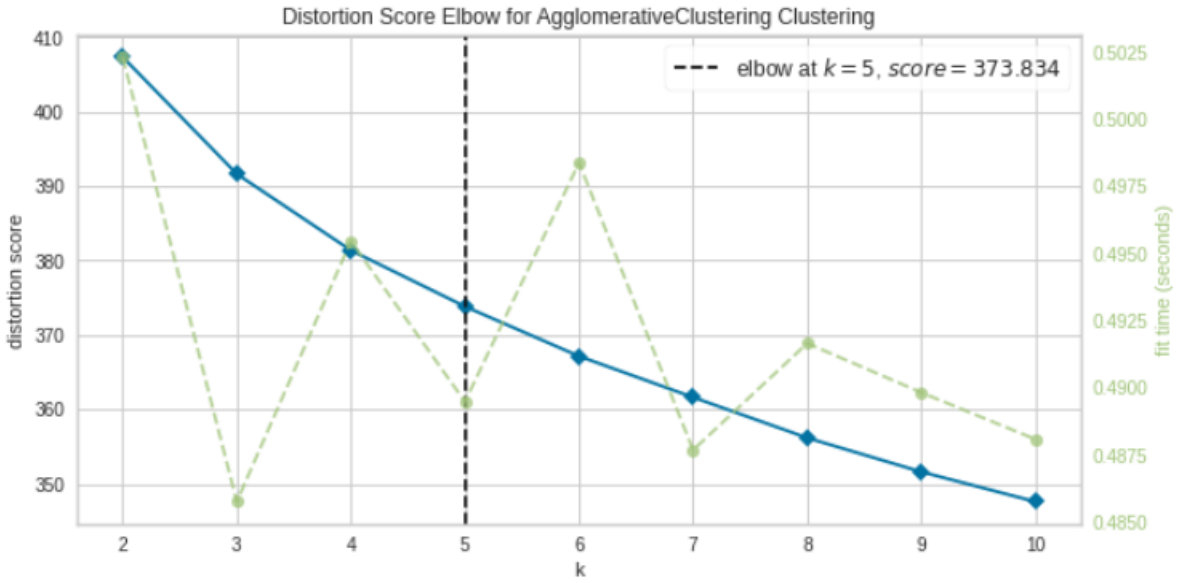
```
For n_clusters = 2 The average silhouette_score is : 0.15866895633128406
For n_clusters = 3 The average silhouette_score is : 0.06643590362064727
For n_clusters = 4 The average silhouette_score is : 0.039273708940442145
For n_clusters = 5 The average silhouette_score is : 0.046599535438943886
For n_clusters = 6 The average silhouette_score is : 0.04151148421013291
For n_clusters = 7 The average silhouette_score is : 0.047176564182434724
For n_clusters = 8 The average silhouette_score is : 0.050437777333335446
For n_clusters = 9 The average silhouette_score is : 0.055343936857065894
For n_clusters = 10 The average silhouette_score is : 0.05971271228952116
```

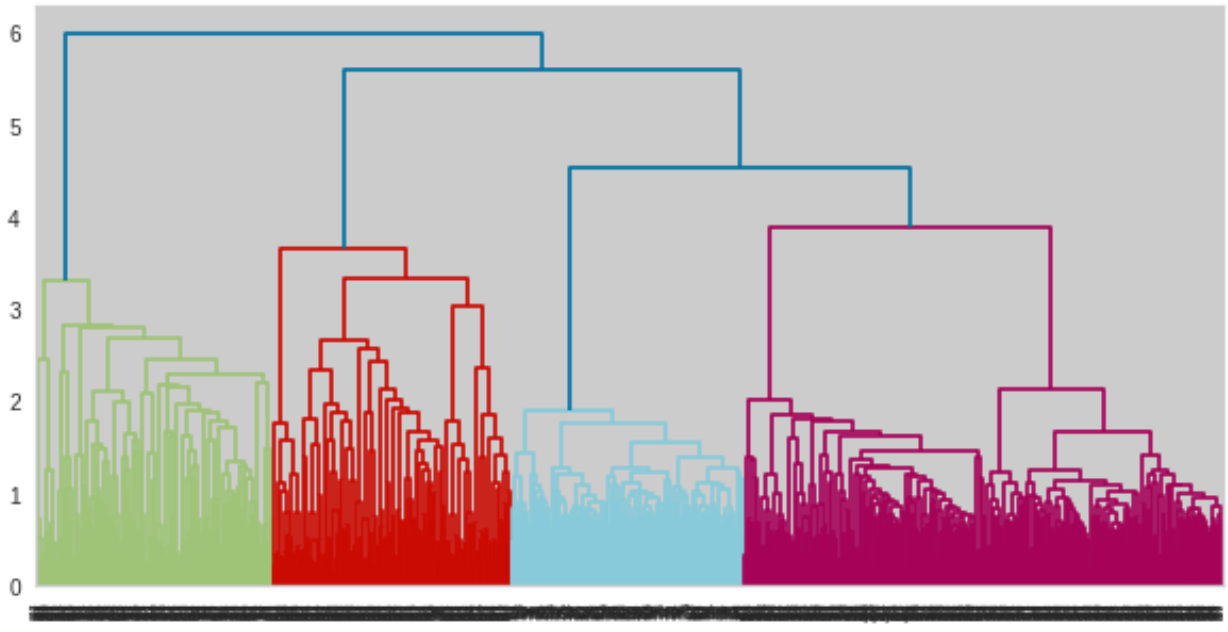




12) Hierarchical Clustering on TF-IDF

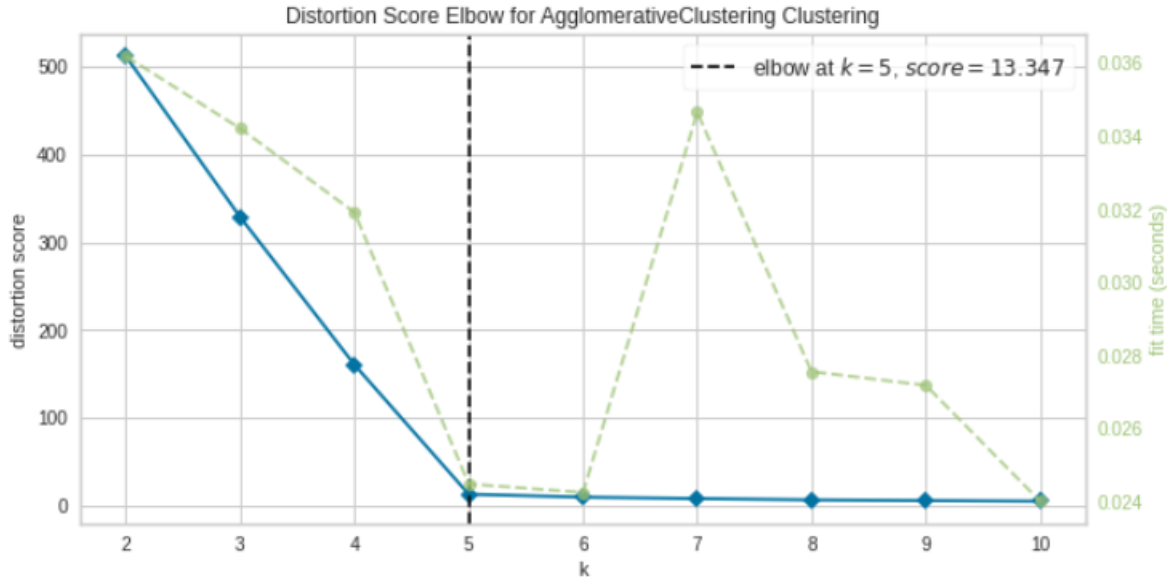
```
For n_clusters = 2 The average silhouette_score is : 0.0994963832728562
For n_clusters = 3 The average silhouette_score is : 0.08139686666731757
For n_clusters = 4 The average silhouette_score is : 0.01927141761861875
For n_clusters = 5 The average silhouette_score is : 0.009101909264157364
For n_clusters = 6 The average silhouette_score is : 0.016390364074212356
For n_clusters = 7 The average silhouette_score is : 0.02236285891195139
For n_clusters = 8 The average silhouette_score is : 0.028724613899336674
For n_clusters = 9 The average silhouette_score is : 0.033602159772748115
For n_clusters = 10 The average silhouette_score is : 0.03789403833611998
```

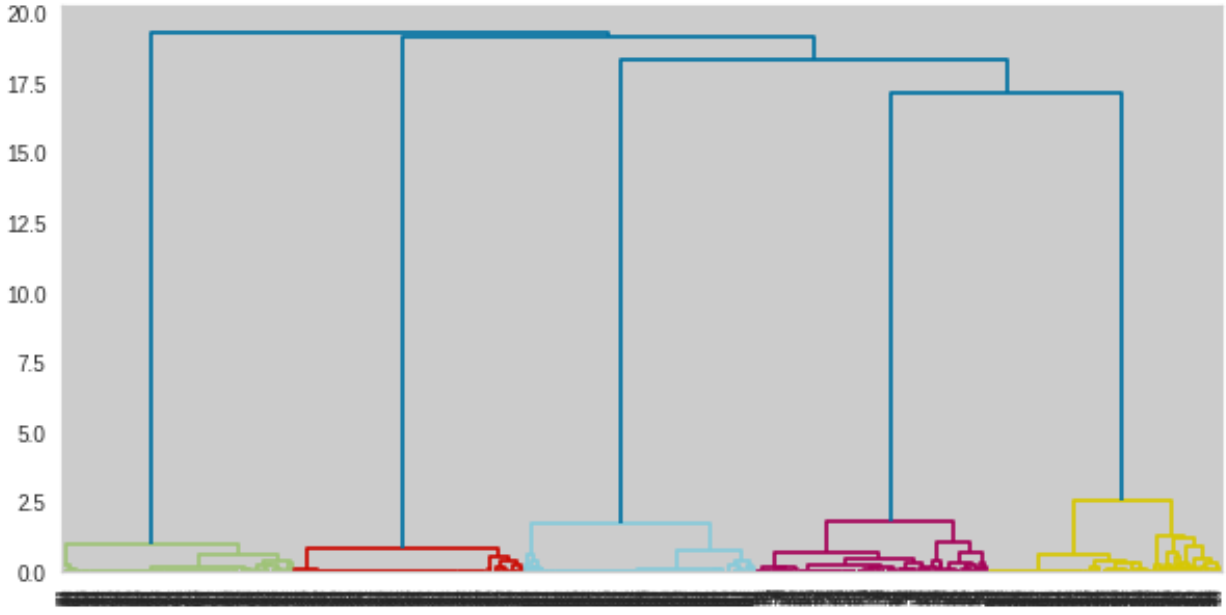




13) Hierarchical Clustering on BOW-LDA

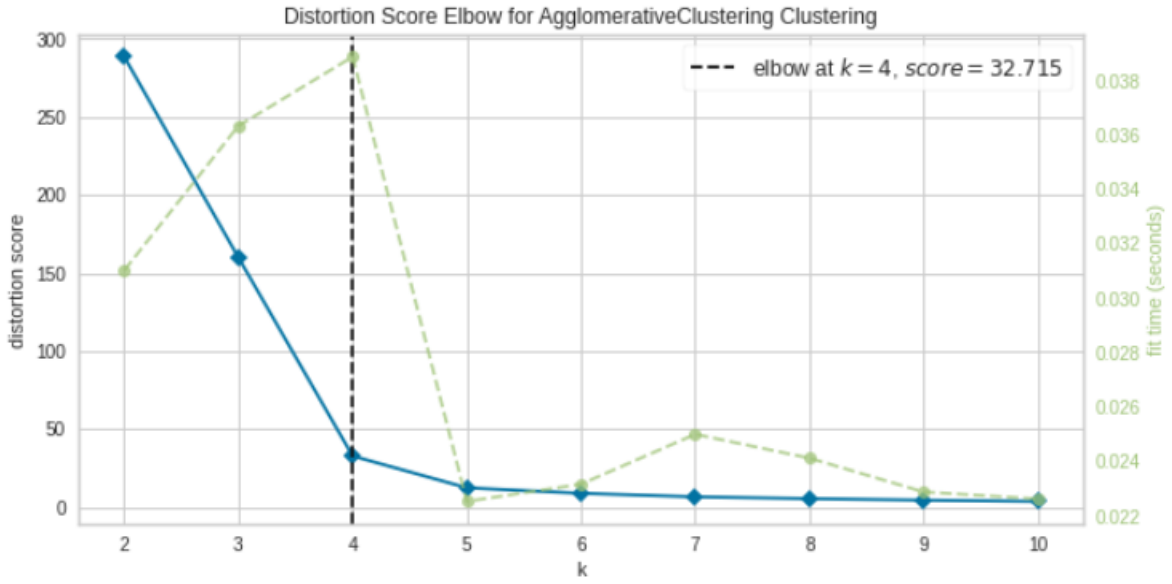
```
For n_clusters = 2 The average silhouette_score is : 0.3944806755479702
For n_clusters = 3 The average silhouette_score is : 0.5817413878021643
For n_clusters = 4 The average silhouette_score is : 0.750110704336471
For n_clusters = 5 The average silhouette_score is : 0.9106970812339511
For n_clusters = 6 The average silhouette_score is : 0.8604474836830458
For n_clusters = 7 The average silhouette_score is : 0.8012346885197351
For n_clusters = 8 The average silhouette_score is : 0.7864020125580111
For n_clusters = 9 The average silhouette_score is : 0.78576426101291
For n_clusters = 10 The average silhouette_score is : 0.7875083579751528
```

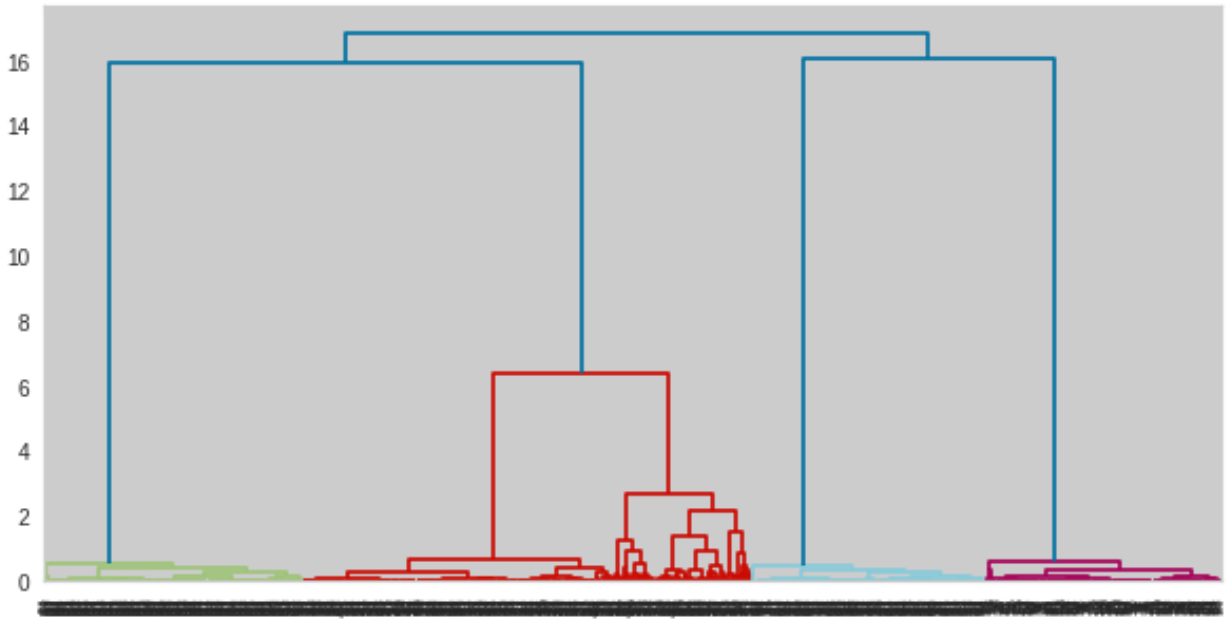




14) Hierarchical Clustering on TFIDF-LDA

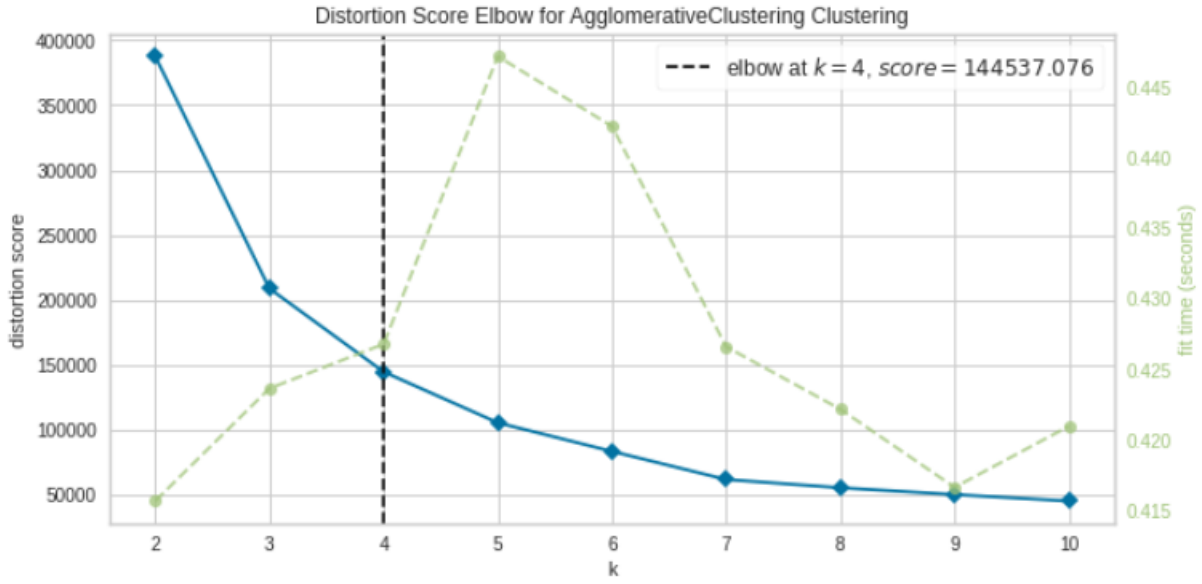
```
For n_clusters = 2 The average silhouette_score is : 0.4627344691352979
For n_clusters = 3 The average silhouette_score is : 0.6587246729774663
For n_clusters = 4 The average silhouette_score is : 0.8342889098771163
For n_clusters = 5 The average silhouette_score is : 0.8426680846639687
For n_clusters = 6 The average silhouette_score is : 0.8398227639053918
For n_clusters = 7 The average silhouette_score is : 0.8482309492589816
For n_clusters = 8 The average silhouette_score is : 0.8507207639093265
For n_clusters = 9 The average silhouette_score is : 0.8418144856378025
For n_clusters = 10 The average silhouette_score is : 0.8468194656087428
```

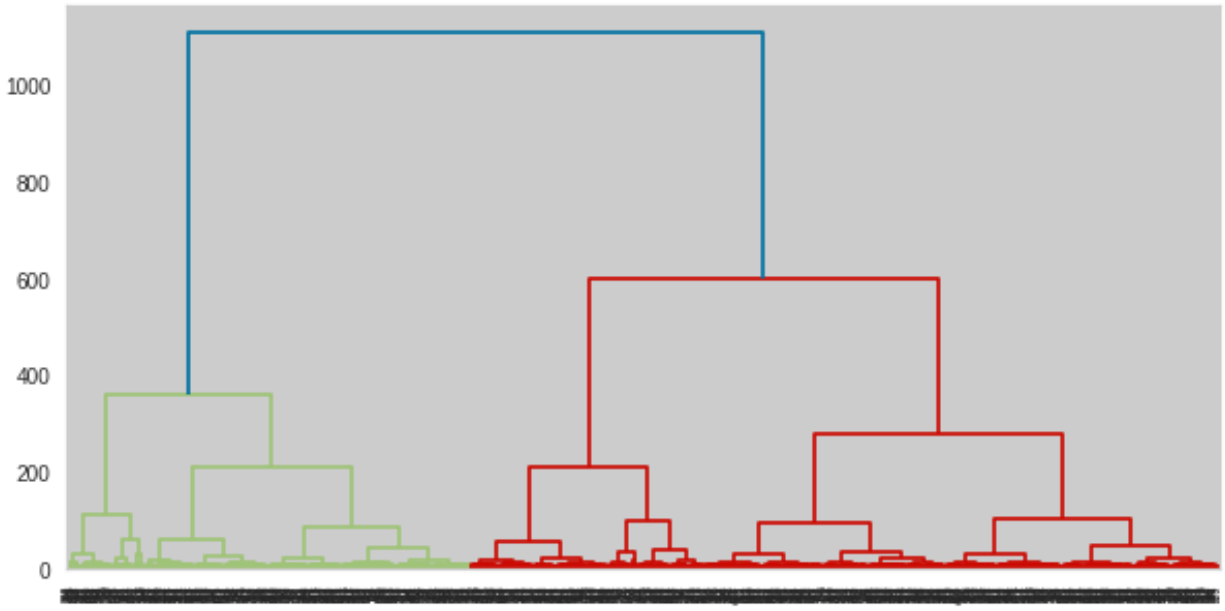




15) Hierarchical Clustering on Word2Vec

```
For n_clusters = 2 The average silhouette_score is : 0.51545924
For n_clusters = 3 The average silhouette_score is : 0.47239295
For n_clusters = 4 The average silhouette_score is : 0.4452574
For n_clusters = 5 The average silhouette_score is : 0.3849801
For n_clusters = 6 The average silhouette_score is : 0.38101
For n_clusters = 7 The average silhouette_score is : 0.37134388
For n_clusters = 8 The average silhouette_score is : 0.36309132
For n_clusters = 9 The average silhouette_score is : 0.30589318
For n_clusters = 10 The average silhouette_score is : 0.3002851
```





Display silhouette and kappa for All models to choose the champion model.

```
1- Expectation maximization(Em) With BOW by N_clusters= 5
silhouette: 0.04418761397569847, kappa: 0.26125
2- Expectation maximization(Em) With TFiDF by N_clusters= 7
silhouette: 0.04565883686655526, kappa: 0.02303120356612187
3- Expectation maximization(Em) With LDA-BOW by N_clusters= 5
silhouette: 0.8488551357697011, kappa: 0.25
4- Expectation maximization(Em) With LDA-IFiDF by N_clusters= 4
silhouette: 0.7691073760466554, kappa: 0.11375000000000002
5- Expectation maximization(Em) With Word2Vec by N_clusters= 5
silhouette: 0.421756774187088, kappa: -0.03249999999999997
----------------------------------------------------------------
1- Kmeans With BOW by N_clusters= 5
silhouette: 0.05027128239892194, kappa: -0.14874999999999994
2- Kmeans With TFiDF by N_clusters= 8
silhouette: 0.03457015297009799, kappa: -0.03295454545454546
3- Kmeans With LDA-BOW by N_clusters= 5
silhouette: 0.9106970812339511, kappa: 0.25
4- Kmeans With LDA-IFiDF by N_clusters= 4
silhouette: 0.8673328508347292, kappa: -0.17999999999999994
5- Kmeans With Word2Vec by N_clusters= 4
silhouette: 0.4469189941883087, kappa: -0.016250000000000098
----------------------------------------------------------------
1- hierarchical cluster With BOW by N_clusters= 5
silhouette: 0.046599535438943886, kappa: 0.16500000000000004
2- hierarchical cluster With TFiDF by N_clusters= 5
silhouette: 0.009101909264157364, kappa: -0.2450000000000001
3- hierarchical cluster With LDA-BOW by N_clusters= 5
silhouette: 0.9106970812339511, kappa: -0.25
4- hierarchical cluster With LDA-IFiDF by N_clusters= 4
silhouette: 0.8342889098771163, kappa: -0.25
5- hierarchical cluster With Word2Vec by N_clusters= 4
silhouette: 0.44525739550590515, kappa: -0.008750000000000036
```

Champion Model

So our Champion Model is K-Means With LDA-BOW

Error-Analysis on K-Means With LDA-BOW

This is our actual Class

```
[Counter({0: 200}),
 Counter({3: 200}),
 Counter({2: 200}),
 Counter({4: 200}),
 Counter({1: 200})]
```

This is Predict Class by our Champion model

```
[Counter({0: 200}),
 Counter({1: 200}),
 Counter({2: 200}),
 Counter({3: 200}),
 Counter({2: 1, 4: 199})]
```

Example of List of common words in Topics 1 and 3

```
['no',
 'need',
 'thou',
 'already',
 'and',
 'number',
 'though',
 'one',
 'but',
 'man']
```

this function takes the above list and sees how many frequent the words are used in topic 1

```
# getting the count frequent words in Topic 1
lst1,lstcount1 = numofTop10(frequentwords_Topic1_3_unq,words_topic1)
```

	Top10 words of topic1	Top10 words of topic3	The number per word of topic1	The number per word of topic3	The best selected word for topic
0	no	no	20	54	Topic3
1	need	need	20	5	Topic1
2	thou	thou	257	267	Topic3
3	already	already	1	10	Topic3
4	and	and	332	541	Topic3
5	number	number	8	5	Topic1
6	though	though	130	14	Topic1
7	one	one	96	83	Topic1
8	but	but	127	164	Topic3
9	man	man	87	155	Topic3

The table analysis for words repetition in the first and the third Topic and identifying where the word belongs to each topic and this led to the misclassification of some of the words

Conclusion

We have trained and evaluated 15 models using K-Means, EM, and hierarchal Clustering using BOW, TF-IDF, LDA_BOW, LDA-TFIDF, and Word2vec. and The best scores that we have got were with LDA-BOW and gave us the highest scores

So, here's our silhouette graph which is showing us that our number of clusters is 5 clusters.

```
For n_clusters = 2 The average silhouette_score is : 0.3944806755479702
For n_clusters = 3 The average silhouette_score is : 0.5817413878021643
For n_clusters = 4 The average silhouette_score is : 0.750110704336471
For n_clusters = 5 The average silhouette_score is : 0.9106970812339511
For n_clusters = 6 The average silhouette_score is : 0.8665047030079577
For n_clusters = 7 The average silhouette_score is : 0.8134924249990083
For n_clusters = 8 The average silhouette_score is : 0.7971160490303818
For n_clusters = 9 The average silhouette_score is : 0.7993143122638098
For n_clusters = 10 The average silhouette_score is : 0.7918492701848875
```