Assignment 5

Information Retrieval

- Mayank Chauhan, MT18008

Methodology

Preprocessing

- Collect the documents.
- Apply tokenization - remove punctuation, accent marks and other non alphanumeric characters like (, >, $, \, \\n, /t, _ etc.
- Apply Normalization and Lemmatization.
- Remove non english words.
- Now, we have the final list of tokens.
- Vocabulary size = 19k (approx.)

K-Means

- Using Bag of Words and Word2Vec models for feature extraction.
- For k = 5 # of classes.
- Stopping criteria is whenever difference between RSS (Residual Sum of Squares) between two iterations is less than 0.001 .

	Bag of Words	**word2vec**
Purity	0.6048	0.6982
ARI	0.3381	0.5064
RSS (1st iteration)	5970.40	2769.62
RSS (At stopping criteria)	4409.96	2052.26

Analysis

The purity and ARI have higher values for word2vec, as the wod2vec features vectors of similar words are closer. The vectors captures very precisely the syntactic and semantic relationship between the words.

K Nearest Neighbours (KNN)

Training & Test Split	K	Accuracy (%)	Log loss
50:50	**1**	93.08	2.3900
	3	93.36	0.8389
	5	93.16	0.5298
80:20	**1**	93.89	2.1068
	3	93.40	1.0292
	5	94.30	0.4919
90:10	**1**	93.60	2.2104
	3	93.60	1.0016
	5	92.60	0.6431

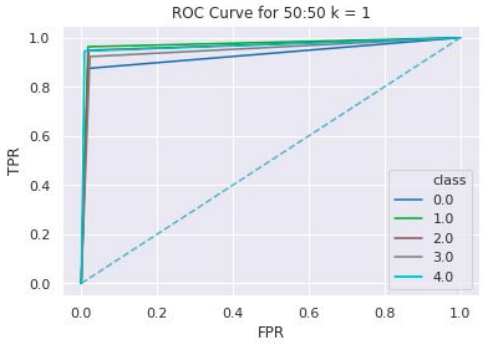
Naive Bayes

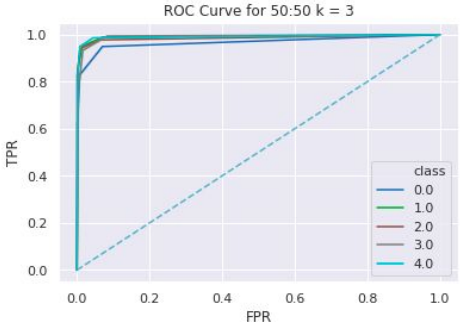
Train - Test Split	Accuracy
50:50	97.8
80:20	98.3
90:10	98.6

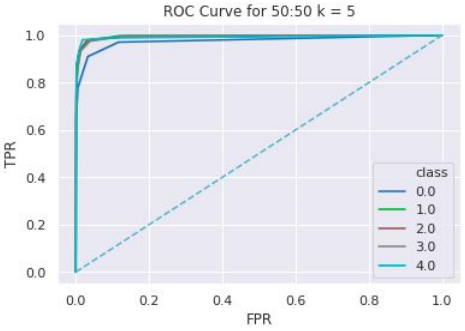
Analysis

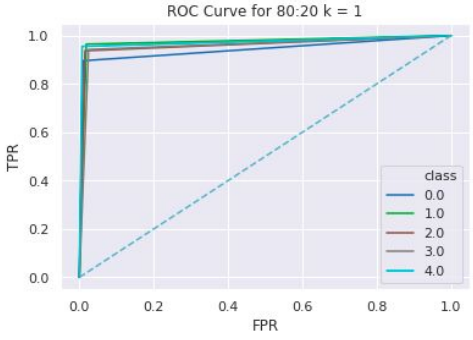
At k = 5, we can see the log loss is minimum in all the splits. That means k = 5 is the best Choice in this case. Using bag of words model, we may have tf = 0 for unknown words in a document , which don't have any effect in the vector representation of the document. Whereas, the naive bayes have is considering giving some probabilities to those unseen words, by using laplace smoothing.

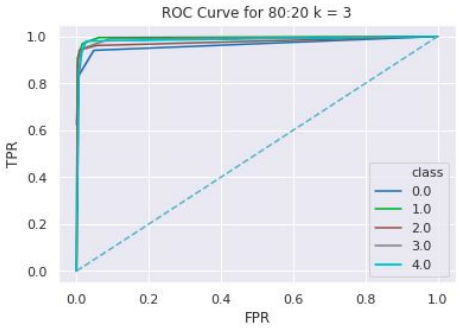
ROC

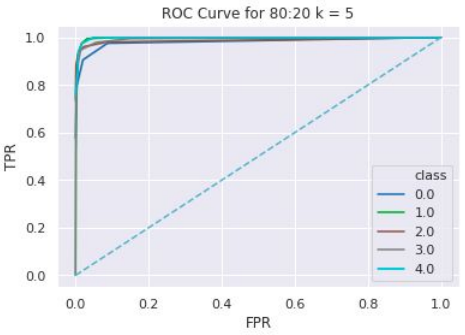

ROC Curve for 50:50 k = 1
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0

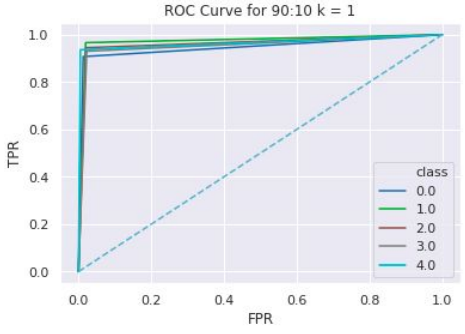

ROC Curve for 50:50 k = 3
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0

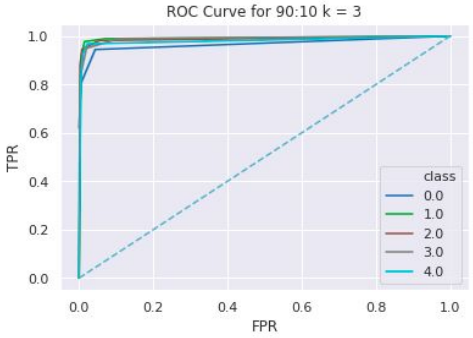

ROC Curve for 50:50 k = 5
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0

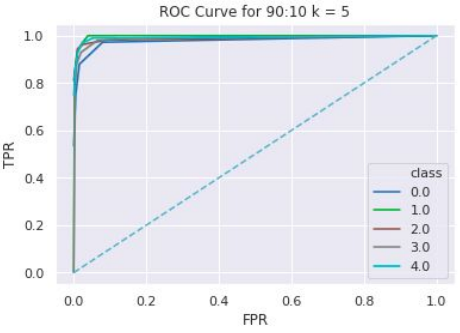

ROC Curve for 80:20 k = 1
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0


ROC Curve for 80:20 k = 3
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0

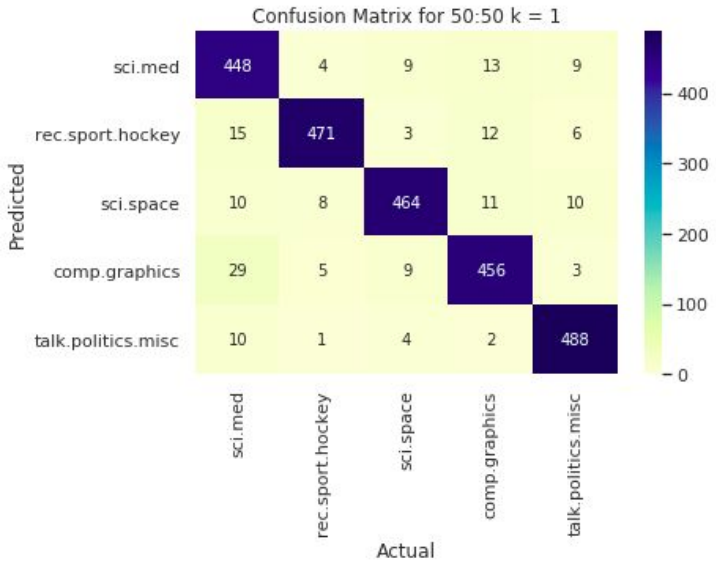

ROC Curve for 80:20 k = 5
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0

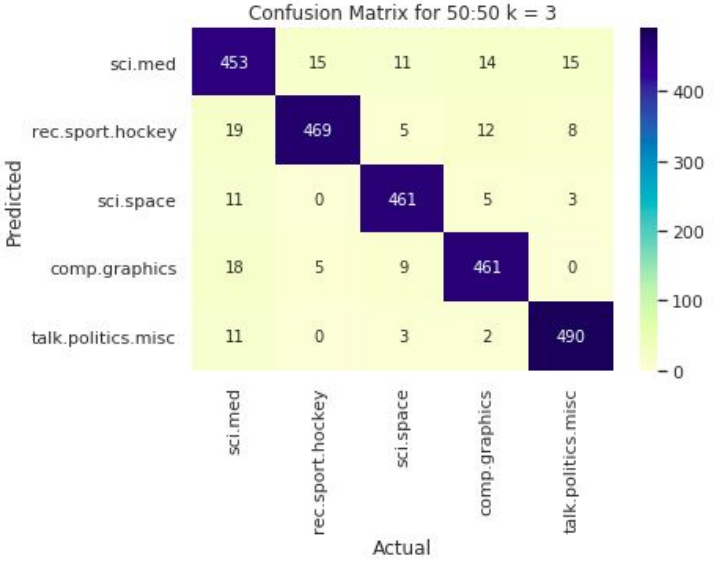

ROC Curve for 90:10 k = 1
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0

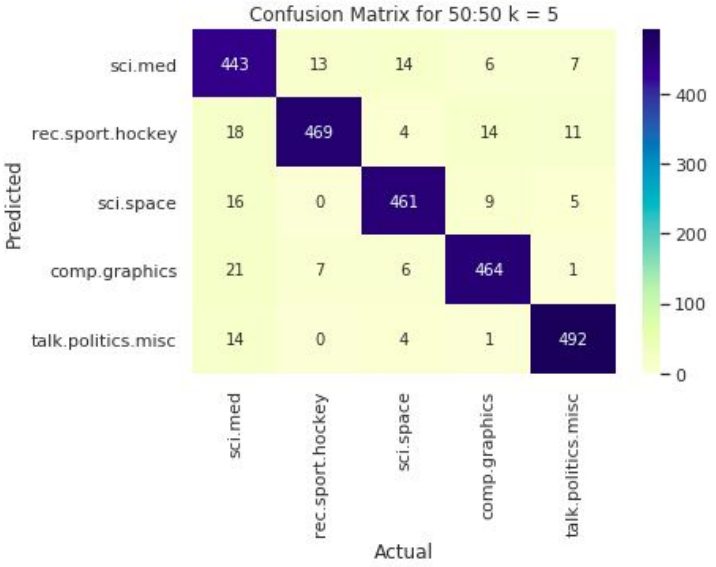

ROC Curve for 90:10 k = 3
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0

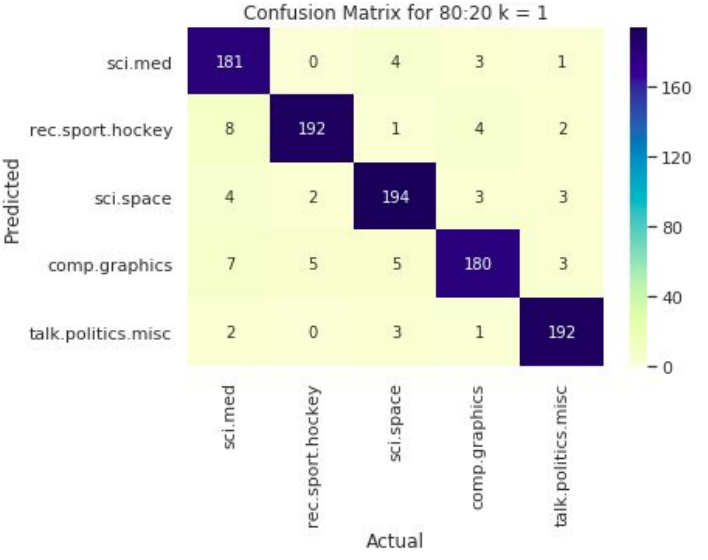

ROC Curve for 90:10 k = 5
TPR
FPR
class
0.0
1.0
2.0
3.0
4.0

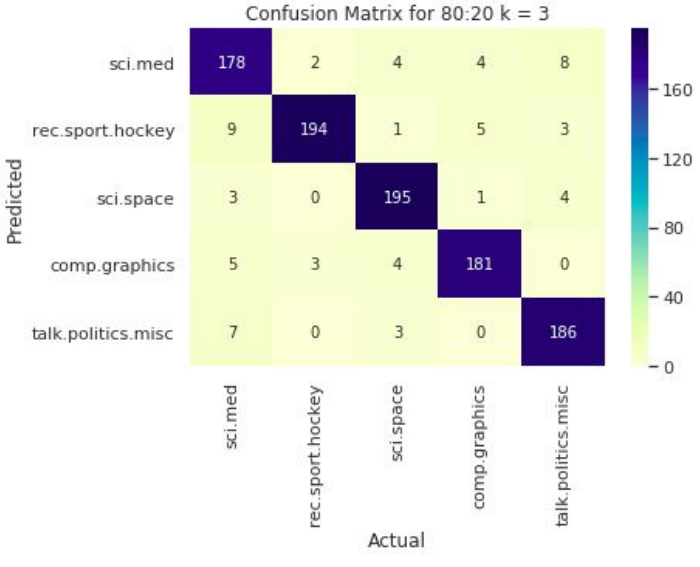
Confusion Matrices

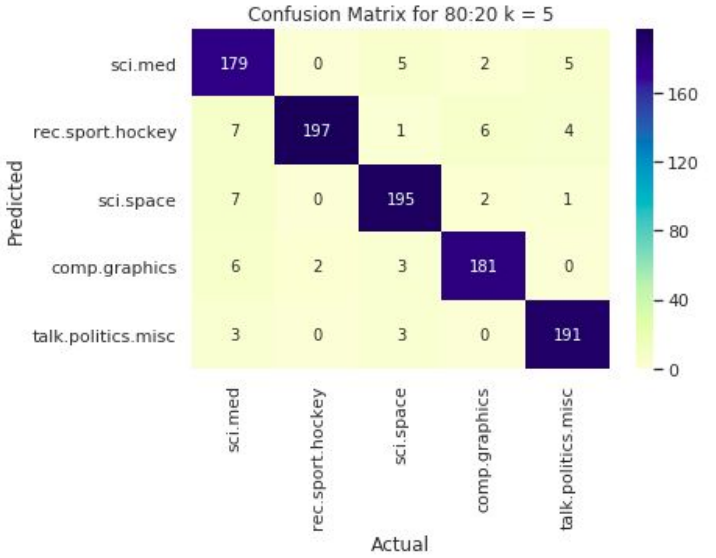


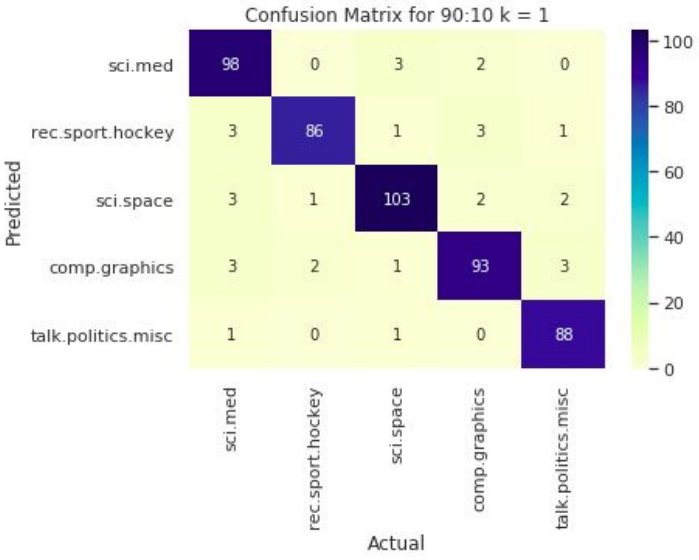


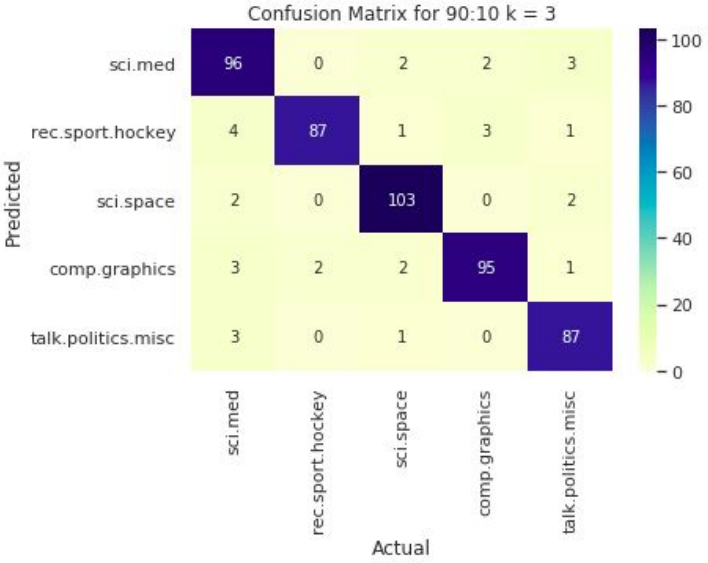


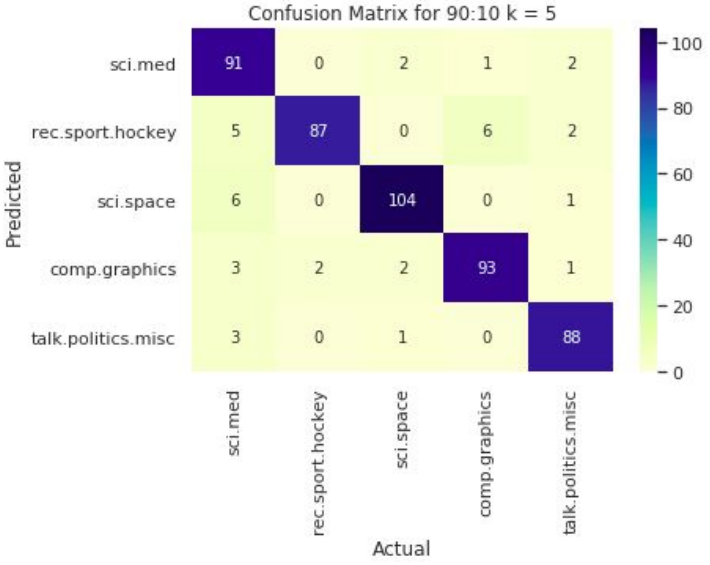







Confusion Matrix for 90:10 k = 1
sci.med
rec.sport.hockey
sci.space
comp.graphics
talk.politics.misc
Predicted
Actual
98 0 3 2 0
3 86 1 3 1
3 1 103 2 2
3 2 1 93 3
1 0 1 0 88
100
80
60
40
20
0


Confusion Matrix for 90:10 k = 3
sci.med
rec.sport.hockey
sci.space
comp.graphics
talk.politics.misc
Predicted
Actual
96 0 2 2 3
4 87 1 3 1
2 0 103 0 2
3 2 2 95 1
3 0 1 0 87
100
80
60
40
20
0


Confusion Matrix for 90:10 k = 5
sci.med
rec.sport.hockey
sci.space
comp.graphics
talk.politics.misc
Predicted
Actual
91 0 2 1 2
5 87 0 6 2
6 0 104 0 1
3 2 2 93 1
3 0 1 0 88
100
80
60
40
20
0