

IR Assignment-5

- SUBHANI SHAIK (MT18117)

Dataset: [20newsgroups dataset](#)

In this dataset, used **comp.graphics, sci.med, talk.politics.misc, rec.sport.hockey, sci.space** [5 classes] which have 1000 documents each i.e., total 5000 documents.

1. K-Means Clustering using Bag-of-words and Word2Vec features:

Pre-Processing the dataset:

- Initially read all the class document names from each folder and stored to a list.
- Read those files and done the pre-processing for both train and test data.
- For pre-processing the text, used NLTK library and done following steps:
 - Conversion to lowercase
 - Contractions
 - Removed unnecessary characters and punctuations and tokenized using RegexpTokenizer.
 - Lemmatization
 - Stop words removal.
- **Feature Extraction:**
 - For **Bag-of-words** features: used term-frequency (here our vector dimension is 65309).
 - For **Word2Vector**: extracted the 300 dimension vector for each word.

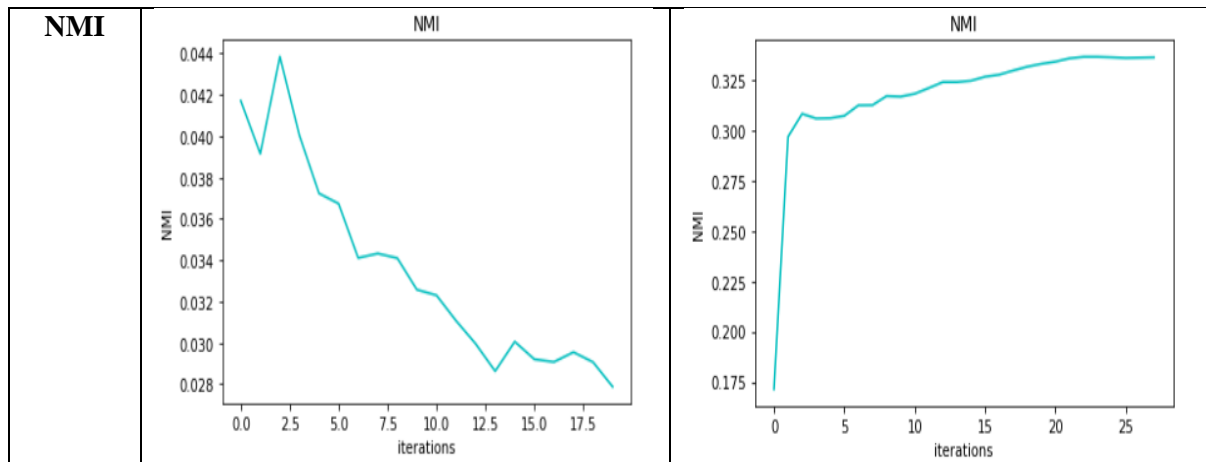
Procedure:

- Prepare the 'terms' list, 'document_ids', Vector space model(VSM) vector and saved them as '.sav'(serializable) files.
- Done the K-means clustering for k=5 classes by taking seed by random initialization.
- Repeated the K-means algorithm and restricted it by setting maximum iteration as 20.
- Computed the metrics like Purity, ARI(Adjusted Rand Index), NMI(Normalized Mutual Information), RSS(Residual Sum of Squares), analysed and plotted them.

Results and Analysis:

- Ran the algorithm for maximum iteration=20. In each iteration calculated the metrics.
- For each iteration 'Purity' was fluctuating, 'RSS' was decreasing, 'ARI' was fluctuating, 'NMI' was fluctuating.

Metric	Plot and Analysis																																																																																															
	Bag-of-Words	Word2Vec																																																																																														
Purity	<p>Purity</p> <table><tr><th>iterations</th><th>Purity</th></tr><tr><td>0.0</td><td>0.2920</td></tr><tr><td>1.0</td><td>0.2895</td></tr><tr><td>2.0</td><td>0.2925</td></tr><tr><td>3.0</td><td>0.2860</td></tr><tr><td>4.0</td><td>0.2830</td></tr><tr><td>5.0</td><td>0.2860</td></tr><tr><td>6.0</td><td>0.2850</td></tr><tr><td>7.0</td><td>0.2850</td></tr><tr><td>8.0</td><td>0.2870</td></tr><tr><td>9.0</td><td>0.2850</td></tr><tr><td>10.0</td><td>0.2860</td></tr><tr><td>11.0</td><td>0.2830</td></tr><tr><td>12.0</td><td>0.2760</td></tr><tr><td>13.0</td><td>0.2740</td></tr><tr><td>14.0</td><td>0.2770</td></tr><tr><td>15.0</td><td>0.2760</td></tr><tr><td>16.0</td><td>0.2750</td></tr><tr><td>17.0</td><td>0.2760</td></tr><tr><td>18.0</td><td>0.2740</td></tr></table>	iterations	Purity	0.0	0.2920	1.0	0.2895	2.0	0.2925	3.0	0.2860	4.0	0.2830	5.0	0.2860	6.0	0.2850	7.0	0.2850	8.0	0.2870	9.0	0.2850	10.0	0.2860	11.0	0.2830	12.0	0.2760	13.0	0.2740	14.0	0.2770	15.0	0.2760	16.0	0.2750	17.0	0.2760	18.0	0.2740	<p>Purity</p> <table><tr><th>iterations</th><th>Purity</th></tr><tr><td>0</td><td>0.375</td></tr><tr><td>1</td><td>0.480</td></tr><tr><td>2</td><td>0.480</td></tr><tr><td>3</td><td>0.480</td></tr><tr><td>4</td><td>0.480</td></tr><tr><td>5</td><td>0.480</td></tr><tr><td>6</td><td>0.482</td></tr><tr><td>7</td><td>0.482</td></tr><tr><td>8</td><td>0.485</td></tr><tr><td>9</td><td>0.485</td></tr><tr><td>10</td><td>0.485</td></tr><tr><td>11</td><td>0.488</td></tr><tr><td>12</td><td>0.490</td></tr><tr><td>13</td><td>0.492</td></tr><tr><td>14</td><td>0.495</td></tr><tr><td>15</td><td>0.498</td></tr><tr><td>16</td><td>0.500</td></tr><tr><td>17</td><td>0.502</td></tr><tr><td>18</td><td>0.503</td></tr><tr><td>19</td><td>0.504</td></tr><tr><td>20</td><td>0.505</td></tr><tr><td>21</td><td>0.506</td></tr><tr><td>22</td><td>0.507</td></tr><tr><td>23</td><td>0.508</td></tr><tr><td>24</td><td>0.508</td></tr><tr><td>25</td><td>0.508</td></tr></table>	iterations	Purity	0	0.375	1	0.480	2	0.480	3	0.480	4	0.480	5	0.480	6	0.482	7	0.482	8	0.485	9	0.485	10	0.485	11	0.488	12	0.490	13	0.492	14	0.495	15	0.498	16	0.500	17	0.502	18	0.503	19	0.504	20	0.505	21	0.506	22	0.507	23	0.508	24	0.508	25	0.508
	iterations	Purity																																																																																														
0.0	0.2920																																																																																															
1.0	0.2895																																																																																															
2.0	0.2925																																																																																															
3.0	0.2860																																																																																															
4.0	0.2830																																																																																															
5.0	0.2860																																																																																															
6.0	0.2850																																																																																															
7.0	0.2850																																																																																															
8.0	0.2870																																																																																															
9.0	0.2850																																																																																															
10.0	0.2860																																																																																															
11.0	0.2830																																																																																															
12.0	0.2760																																																																																															
13.0	0.2740																																																																																															
14.0	0.2770																																																																																															
15.0	0.2760																																																																																															
16.0	0.2750																																																																																															
17.0	0.2760																																																																																															
18.0	0.2740																																																																																															
iterations	Purity																																																																																															
0	0.375																																																																																															
1	0.480																																																																																															
2	0.480																																																																																															
3	0.480																																																																																															
4	0.480																																																																																															
5	0.480																																																																																															
6	0.482																																																																																															
7	0.482																																																																																															
8	0.485																																																																																															
9	0.485																																																																																															
10	0.485																																																																																															
11	0.488																																																																																															
12	0.490																																																																																															
13	0.492																																																																																															
14	0.495																																																																																															
15	0.498																																																																																															
16	0.500																																																																																															
17	0.502																																																																																															
18	0.503																																																																																															
19	0.504																																																																																															
20	0.505																																																																																															
21	0.506																																																																																															
22	0.507																																																																																															
23	0.508																																																																																															
24	0.508																																																																																															
25	0.508																																																																																															
RSS	<p>RSS</p> <table><tr><th>iterations</th><th>RSS</th></tr><tr><td>0.0</td><td>700</td></tr><tr><td>1.0</td><td>535</td></tr><tr><td>2.0</td><td>530</td></tr><tr><td>3.0</td><td>528</td></tr><tr><td>4.0</td><td>528</td></tr><tr><td>5.0</td><td>528</td></tr><tr><td>6.0</td><td>528</td></tr><tr><td>7.0</td><td>528</td></tr><tr><td>8.0</td><td>528</td></tr><tr><td>9.0</td><td>528</td></tr><tr><td>10.0</td><td>528</td></tr><tr><td>11.0</td><td>528</td></tr><tr><td>12.0</td><td>528</td></tr><tr><td>13.0</td><td>528</td></tr><tr><td>14.0</td><td>528</td></tr><tr><td>15.0</td><td>528</td></tr><tr><td>16.0</td><td>528</td></tr><tr><td>17.0</td><td>528</td></tr><tr><td>18.0</td><td>528</td></tr></table>	iterations	RSS	0.0	700	1.0	535	2.0	530	3.0	528	4.0	528	5.0	528	6.0	528	7.0	528	8.0	528	9.0	528	10.0	528	11.0	528	12.0	528	13.0	528	14.0	528	15.0	528	16.0	528	17.0	528	18.0	528	<p>RSS</p> <table><tr><th>iterations</th><th>RSS</th></tr><tr><td>0</td><td>2620</td></tr><tr><td>1</td><td>2100</td></tr><tr><td>2</td><td>2095</td></tr><tr><td>3</td><td>2095</td></tr><tr><td>4</td><td>2095</td></tr><tr><td>5</td><td>2095</td></tr><tr><td>6</td><td>2095</td></tr><tr><td>7</td><td>2095</td></tr><tr><td>8</td><td>2095</td></tr><tr><td>9</td><td>2095</td></tr><tr><td>10</td><td>2095</td></tr><tr><td>11</td><td>2095</td></tr><tr><td>12</td><td>2095</td></tr><tr><td>13</td><td>2095</td></tr><tr><td>14</td><td>2095</td></tr><tr><td>15</td><td>2095</td></tr><tr><td>16</td><td>2095</td></tr><tr><td>17</td><td>2095</td></tr><tr><td>18</td><td>2095</td></tr><tr><td>19</td><td>2095</td></tr><tr><td>20</td><td>2095</td></tr><tr><td>21</td><td>2095</td></tr><tr><td>22</td><td>2095</td></tr><tr><td>23</td><td>2095</td></tr><tr><td>24</td><td>2095</td></tr><tr><td>25</td><td>2095</td></tr></table>	iterations	RSS	0	2620	1	2100	2	2095	3	2095	4	2095	5	2095	6	2095	7	2095	8	2095	9	2095	10	2095	11	2095	12	2095	13	2095	14	2095	15	2095	16	2095	17	2095	18	2095	19	2095	20	2095	21	2095	22	2095	23	2095	24	2095	25	2095
	iterations	RSS																																																																																														
0.0	700																																																																																															
1.0	535																																																																																															
2.0	530																																																																																															
3.0	528																																																																																															
4.0	528																																																																																															
5.0	528																																																																																															
6.0	528																																																																																															
7.0	528																																																																																															
8.0	528																																																																																															
9.0	528																																																																																															
10.0	528																																																																																															
11.0	528																																																																																															
12.0	528																																																																																															
13.0	528																																																																																															
14.0	528																																																																																															
15.0	528																																																																																															
16.0	528																																																																																															
17.0	528																																																																																															
18.0	528																																																																																															
iterations	RSS																																																																																															
0	2620																																																																																															
1	2100																																																																																															
2	2095																																																																																															
3	2095																																																																																															
4	2095																																																																																															
5	2095																																																																																															
6	2095																																																																																															
7	2095																																																																																															
8	2095																																																																																															
9	2095																																																																																															
10	2095																																																																																															
11	2095																																																																																															
12	2095																																																																																															
13	2095																																																																																															
14	2095																																																																																															
15	2095																																																																																															
16	2095																																																																																															
17	2095																																																																																															
18	2095																																																																																															
19	2095																																																																																															
20	2095																																																																																															
21	2095																																																																																															
22	2095																																																																																															
23	2095																																																																																															
24	2095																																																																																															
25	2095																																																																																															
ARI	<p>ARI</p> <table><tr><th>iterations</th><th>ARI</th></tr><tr><td>0.0</td><td>0.031</td></tr><tr><td>1.0</td><td>0.025</td></tr><tr><td>2.0</td><td>0.026</td></tr><tr><td>3.0</td><td>0.027</td></tr><tr><td>4.0</td><td>0.027</td></tr><tr><td>5.0</td><td>0.027</td></tr><tr><td>6.0</td><td>0.024</td></tr><tr><td>7.0</td><td>0.023</td></tr><tr><td>8.0</td><td>0.022</td></tr><tr><td>9.0</td><td>0.021</td></tr><tr><td>10.0</td><td>0.020</td></tr><tr><td>11.0</td><td>0.019</td></tr><tr><td>12.0</td><td>0.018</td></tr><tr><td>13.0</td><td>0.017</td></tr><tr><td>14.0</td><td>0.018</td></tr><tr><td>15.0</td><td>0.018</td></tr><tr><td>16.0</td><td>0.019</td></tr><tr><td>17.0</td><td>0.018</td></tr><tr><td>18.0</td><td>0.018</td></tr></table>	iterations	ARI	0.0	0.031	1.0	0.025	2.0	0.026	3.0	0.027	4.0	0.027	5.0	0.027	6.0	0.024	7.0	0.023	8.0	0.022	9.0	0.021	10.0	0.020	11.0	0.019	12.0	0.018	13.0	0.017	14.0	0.018	15.0	0.018	16.0	0.019	17.0	0.018	18.0	0.018	<p>ARI</p> <table><tr><th>iterations</th><th>ARI</th></tr><tr><td>0</td><td>0.08</td></tr><tr><td>1</td><td>0.24</td></tr><tr><td>2</td><td>0.25</td></tr><tr><td>3</td><td>0.255</td></tr><tr><td>4</td><td>0.258</td></tr><tr><td>5</td><td>0.260</td></tr><tr><td>6</td><td>0.262</td></tr><tr><td>7</td><td>0.265</td></tr><tr><td>8</td><td>0.268</td></tr><tr><td>9</td><td>0.270</td></tr><tr><td>10</td><td>0.272</td></tr><tr><td>11</td><td>0.275</td></tr><tr><td>12</td><td>0.278</td></tr><tr><td>13</td><td>0.280</td></tr><tr><td>14</td><td>0.282</td></tr><tr><td>15</td><td>0.285</td></tr><tr><td>16</td><td>0.288</td></tr><tr><td>17</td><td>0.290</td></tr><tr><td>18</td><td>0.292</td></tr><tr><td>19</td><td>0.293</td></tr><tr><td>20</td><td>0.294</td></tr><tr><td>21</td><td>0.295</td></tr><tr><td>22</td><td>0.296</td></tr><tr><td>23</td><td>0.297</td></tr><tr><td>24</td><td>0.298</td></tr><tr><td>25</td><td>0.298</td></tr></table>	iterations	ARI	0	0.08	1	0.24	2	0.25	3	0.255	4	0.258	5	0.260	6	0.262	7	0.265	8	0.268	9	0.270	10	0.272	11	0.275	12	0.278	13	0.280	14	0.282	15	0.285	16	0.288	17	0.290	18	0.292	19	0.293	20	0.294	21	0.295	22	0.296	23	0.297	24	0.298	25	0.298
	iterations	ARI																																																																																														
0.0	0.031																																																																																															
1.0	0.025																																																																																															
2.0	0.026																																																																																															
3.0	0.027																																																																																															
4.0	0.027																																																																																															
5.0	0.027																																																																																															
6.0	0.024																																																																																															
7.0	0.023																																																																																															
8.0	0.022																																																																																															
9.0	0.021																																																																																															
10.0	0.020																																																																																															
11.0	0.019																																																																																															
12.0	0.018																																																																																															
13.0	0.017																																																																																															
14.0	0.018																																																																																															
15.0	0.018																																																																																															
16.0	0.019																																																																																															
17.0	0.018																																																																																															
18.0	0.018																																																																																															
iterations	ARI																																																																																															
0	0.08																																																																																															
1	0.24																																																																																															
2	0.25																																																																																															
3	0.255																																																																																															
4	0.258																																																																																															
5	0.260																																																																																															
6	0.262																																																																																															
7	0.265																																																																																															
8	0.268																																																																																															
9	0.270																																																																																															
10	0.272																																																																																															
11	0.275																																																																																															
12	0.278																																																																																															
13	0.280																																																																																															
14	0.282																																																																																															
15	0.285																																																																																															
16	0.288																																																																																															
17	0.290																																																																																															
18	0.292																																																																																															
19	0.293																																																																																															
20	0.294																																																																																															
21	0.295																																																																																															
22	0.296																																																																																															
23	0.297																																																																																															
24	0.298																																																																																															
25	0.298																																																																																															



2. KNN Classification for different splits:

Pre-Processing the dataset:

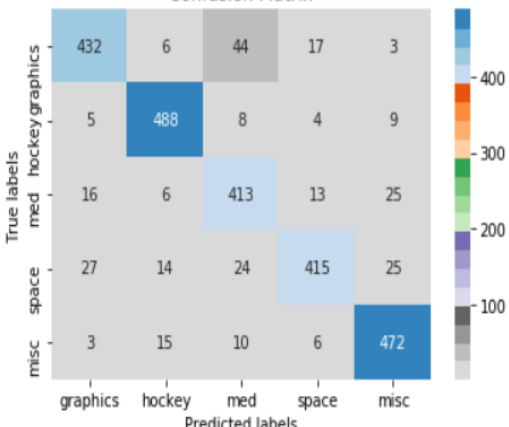
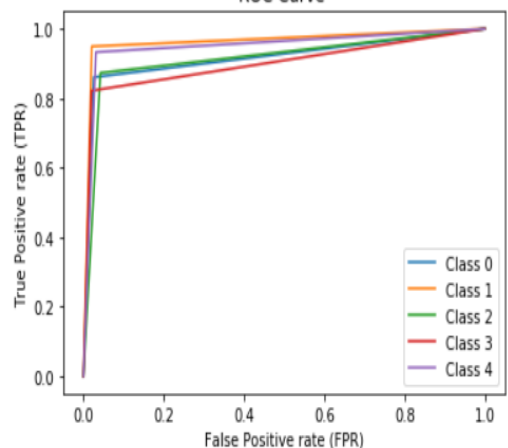
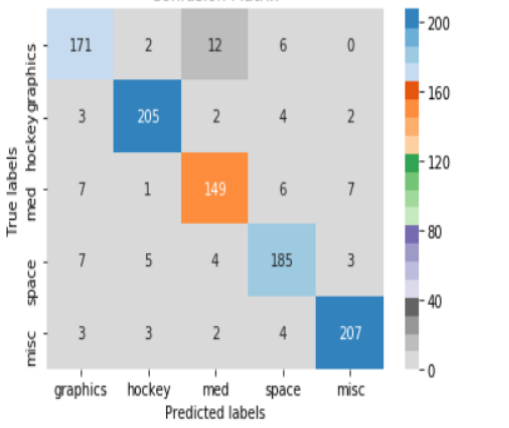
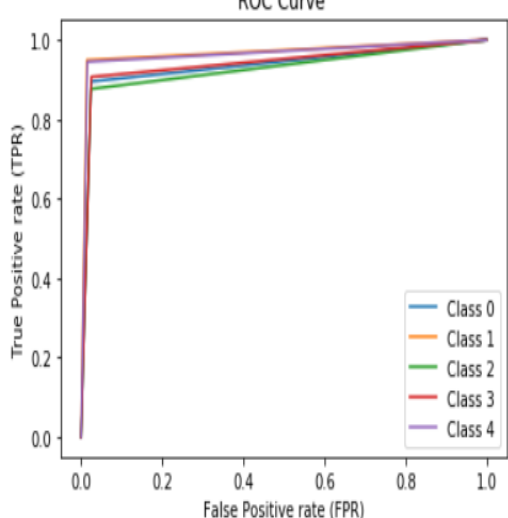
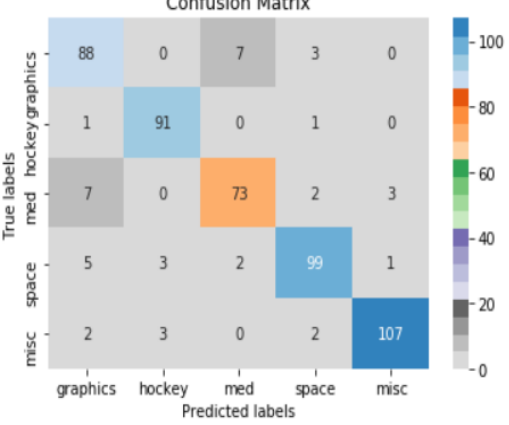
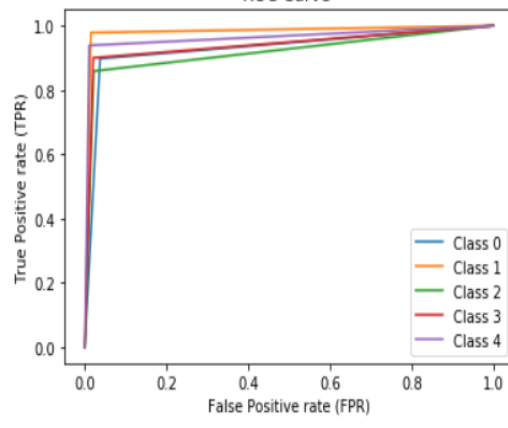
- Initially read all the class document names from each folder and stored to a list.
- Read those files and done the pre-processing for both train and test data.
- For pre-processing the text, used NLTK library and done following steps:
 - Conversion to lowercase
 - Contractions
 - Removed unnecessary characters and punctuations and tokenized using RegexpTokenizer.
 - Lemmatization
 - Stop words removal.
- **Feature Extraction:**
 - For **Word2Vector**: extracted the 300 dimension vector for each word.

Procedure:

- Initially split the dataset according to our need as training and test data as 50:50, 80:20, 90:10 ratios.
- Done the KNN for k=1,3,5 values for each splitted data and computed the metrics like Accuracy, Confusion matrix and ROC curves.
- Save those variables into files by using 'Joblib' and load them for testing.
- Later after prediciting for all the docs, calculate the 'Confusion matrix' and 'Accuracy' for that split.
- Repeated the same process for other splits.

Results:

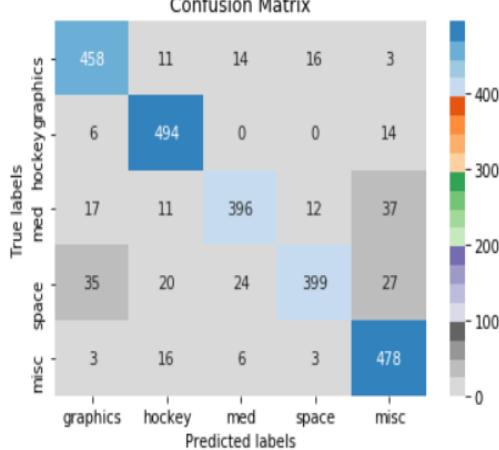
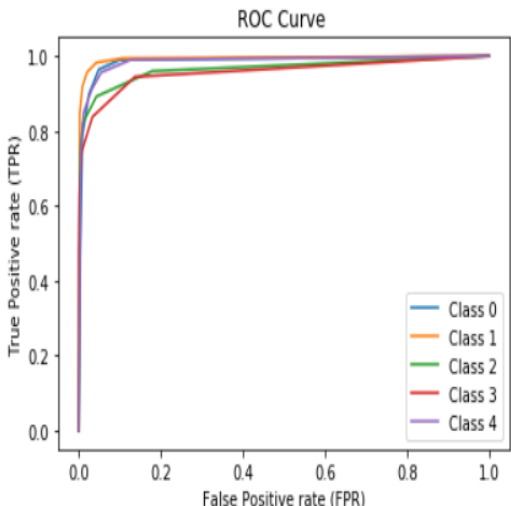
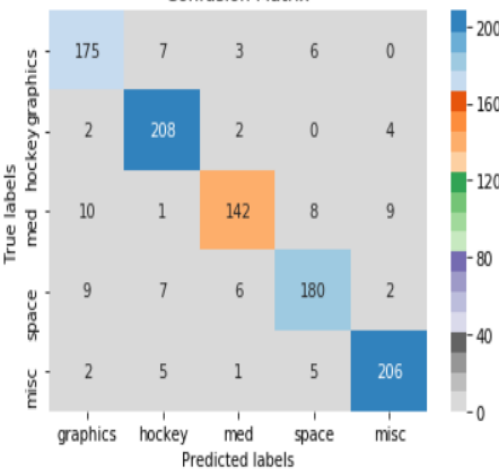
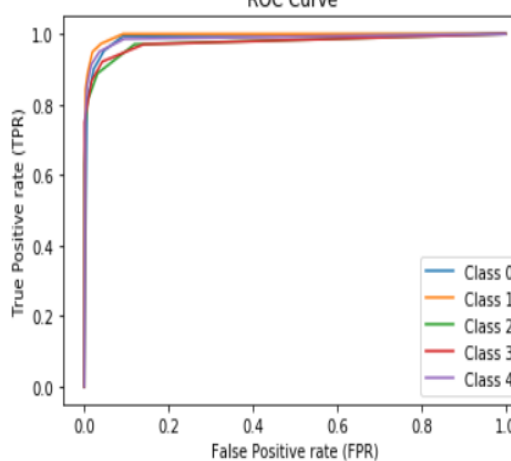
For k=1:

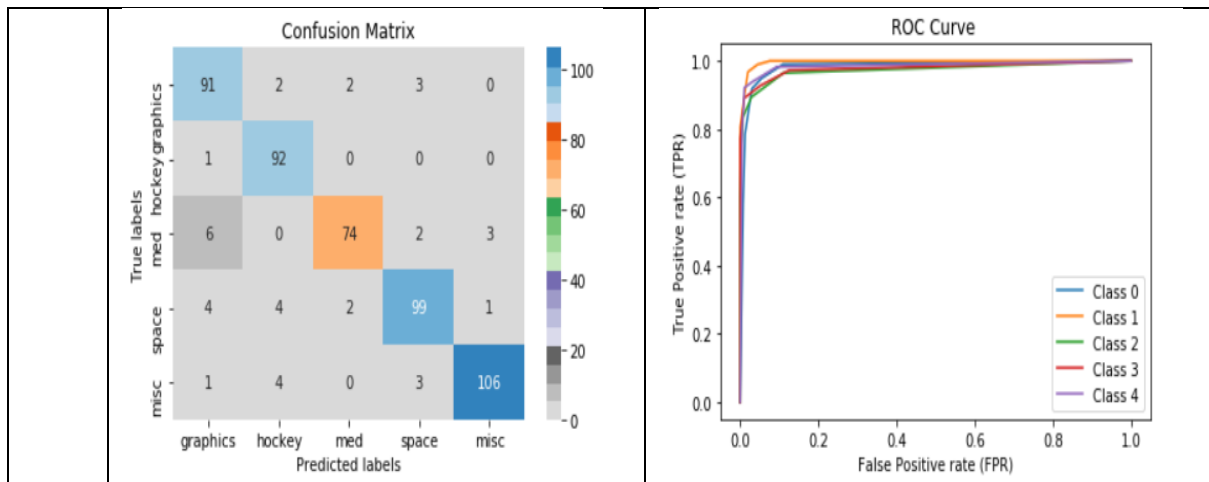
Split	Accuracy and Confusion matrix	ROC
50:50	Accuracy=0.88 	
80:20	Accuracy=0.917 	
90:10	Accuracy=0.916 	

For k=3:

Split	Accuracy and Confusion matrix	ROC																																				
50:50	<p>Accuracy=0.8952</p> <p>Confusion Matrix</p> <table><tr><th></th><th>graphics</th><th>hockey</th><th>med</th><th>space</th><th>misc</th></tr><tr><th>graphics</th><td>463</td><td>9</td><td>18</td><td>11</td><td>1</td></tr><tr><th>hockey</th><td>6</td><td>490</td><td>4</td><td>3</td><td>11</td></tr><tr><th>med</th><td>16</td><td>9</td><td>408</td><td>9</td><td>31</td></tr><tr><th>space</th><td>34</td><td>17</td><td>27</td><td>403</td><td>24</td></tr><tr><th>misc</th><td>5</td><td>14</td><td>7</td><td>6</td><td>474</td></tr></table>		graphics	hockey	med	space	misc	graphics	463	9	18	11	1	hockey	6	490	4	3	11	med	16	9	408	9	31	space	34	17	27	403	24	misc	5	14	7	6	474	<p>ROC Curve</p>
	graphics	hockey	med	space	misc																																	
graphics	463	9	18	11	1																																	
hockey	6	490	4	3	11																																	
med	16	9	408	9	31																																	
space	34	17	27	403	24																																	
misc	5	14	7	6	474																																	
80:20	<p>Accuracy=0.91</p> <p>Confusion Matrix</p> <table><tr><th></th><th>graphics</th><th>hockey</th><th>med</th><th>space</th><th>misc</th></tr><tr><th>graphics</th><td>174</td><td>7</td><td>4</td><td>5</td><td>1</td></tr><tr><th>hockey</th><td>5</td><td>205</td><td>2</td><td>1</td><td>3</td></tr><tr><th>med</th><td>5</td><td>1</td><td>147</td><td>6</td><td>11</td></tr><tr><th>space</th><td>8</td><td>6</td><td>5</td><td>181</td><td>4</td></tr><tr><th>misc</th><td>4</td><td>5</td><td>0</td><td>7</td><td>203</td></tr></table>		graphics	hockey	med	space	misc	graphics	174	7	4	5	1	hockey	5	205	2	1	3	med	5	1	147	6	11	space	8	6	5	181	4	misc	4	5	0	7	203	<p>ROC Curve</p>
	graphics	hockey	med	space	misc																																	
graphics	174	7	4	5	1																																	
hockey	5	205	2	1	3																																	
med	5	1	147	6	11																																	
space	8	6	5	181	4																																	
misc	4	5	0	7	203																																	
90:10	<p>Accuracy=0.914</p> <p>Confusion Matrix</p> <table><tr><th></th><th>graphics</th><th>hockey</th><th>med</th><th>space</th><th>misc</th></tr><tr><th>graphics</th><td>90</td><td>2</td><td>3</td><td>3</td><td>0</td></tr><tr><th>hockey</th><td>2</td><td>91</td><td>0</td><td>0</td><td>0</td></tr><tr><th>med</th><td>4</td><td>0</td><td>74</td><td>3</td><td>4</td></tr><tr><th>space</th><td>4</td><td>3</td><td>3</td><td>97</td><td>3</td></tr><tr><th>misc</th><td>2</td><td>4</td><td>0</td><td>3</td><td>105</td></tr></table>		graphics	hockey	med	space	misc	graphics	90	2	3	3	0	hockey	2	91	0	0	0	med	4	0	74	3	4	space	4	3	3	97	3	misc	2	4	0	3	105	<p>ROC Curve</p>
	graphics	hockey	med	space	misc																																	
graphics	90	2	3	3	0																																	
hockey	2	91	0	0	0																																	
med	4	0	74	3	4																																	
space	4	3	3	97	3																																	
misc	2	4	0	3	105																																	

For k=5:

Split	Accuracy and Confusion matrix	ROC																																				
50:50	<p>Accuracy=0.89</p> <p>Confusion Matrix</p>  <table><tr><th></th><th>graphics</th><th>hockey</th><th>med</th><th>space</th><th>misc</th></tr><tr><th>graphics</th><td>458</td><td>11</td><td>14</td><td>16</td><td>3</td></tr><tr><th>hockey</th><td>6</td><td>494</td><td>0</td><td>0</td><td>14</td></tr><tr><th>med</th><td>17</td><td>11</td><td>396</td><td>12</td><td>37</td></tr><tr><th>space</th><td>35</td><td>20</td><td>24</td><td>399</td><td>27</td></tr><tr><th>misc</th><td>3</td><td>16</td><td>6</td><td>3</td><td>478</td></tr></table>		graphics	hockey	med	space	misc	graphics	458	11	14	16	3	hockey	6	494	0	0	14	med	17	11	396	12	37	space	35	20	24	399	27	misc	3	16	6	3	478	<p>ROC Curve</p> 
	graphics	hockey	med	space	misc																																	
graphics	458	11	14	16	3																																	
hockey	6	494	0	0	14																																	
med	17	11	396	12	37																																	
space	35	20	24	399	27																																	
misc	3	16	6	3	478																																	
80:20	<p>Accuracy=0.911</p> <p>Confusion Matrix</p>  <table><tr><th></th><th>graphics</th><th>hockey</th><th>med</th><th>space</th><th>misc</th></tr><tr><th>graphics</th><td>175</td><td>7</td><td>3</td><td>6</td><td>0</td></tr><tr><th>hockey</th><td>2</td><td>208</td><td>2</td><td>0</td><td>4</td></tr><tr><th>med</th><td>10</td><td>1</td><td>142</td><td>8</td><td>9</td></tr><tr><th>space</th><td>9</td><td>7</td><td>6</td><td>180</td><td>2</td></tr><tr><th>misc</th><td>2</td><td>5</td><td>1</td><td>5</td><td>206</td></tr></table>		graphics	hockey	med	space	misc	graphics	175	7	3	6	0	hockey	2	208	2	0	4	med	10	1	142	8	9	space	9	7	6	180	2	misc	2	5	1	5	206	<p>ROC Curve</p> 
	graphics	hockey	med	space	misc																																	
graphics	175	7	3	6	0																																	
hockey	2	208	2	0	4																																	
med	10	1	142	8	9																																	
space	9	7	6	180	2																																	
misc	2	5	1	5	206																																	
90:10	<p>Accuracy=0.924</p>																																					



- In KNN, it is giving accuracy of max of 91% in a split.
- But in Naïve Bayes, it is giving accuracy of >90% in almost all splits.
- So Naïve bayes is better than KNN, because it was calculated depending on joint probability, whereas KNN is just predicts according nearest neighbour which is not robust.