README
ASSIGNMENT-5

- **Data Preprocessing Steps:**

- Stop Word Removal
- Lemmatization
- Removal of words of length less than 3.
- Removing empty lines
- Removing multiple spaces between words
- Removing Special Characters
- Lower case conversion
- Removal of alpha-numeric words and numbers

- **Methodology For Question-1 (NAIVE BAYES):**

1. **TF-IDF Based Feature Selection:**

- Firstly, I read all the names of the documents from the corpus (mentioned folders) and performed a train-test split with a random state of value 0 (a particular random state is used since we obtain the same train-test splits of the data every time we run the code which eases our job in performance comparisons).

- Now I read the data corresponding to training documents and merged the data of documents belonging to the same class. Then **I have used the log-normal variate of TF (tf = 1+$\log_{10}$(tf)) and inverse variation of DF (df = $\log_{10}$(N/df)) where N is the total number of classes considered for the corpus.** Then i had computed tf-idf values by multiplying corresponding term frequencies by their document frequencies. Now **i sorted the tf-idf values of the terms in every class in descending order** to ease the feature selection. I also stored the raw term frequencies corresponding to every class which are used during probability calculations.

- Then I selected the top k% of the features from every class (k is given by the user) **based on their TF-IDF scores.**

- At the run time when we encounter the test document (query), if the word/feature in the test document is present in the top K% features of that class then we proceed to calculate the log probability value of that term with respect to that particular class.if the word is absent i used the technique of add-1 smoothing (laplace smoothing). The final log probability of the test document belonging to that class is the sum of log-probabilities of all words in that test document w.r.t that class summed up with the log of prior probability. This process is repeated for all the classes for a test document.

- **Probability of a word(feature) given the class (if feature is existing in the class) is calculated as follows,**

$$\hat{P}(w_i \mid c) = \frac{count(w_i,c)+1}{\sum_{w \in V} \left(count(w,c)+1\right)}$$

$$= \frac{count(w_i,c)+1}{\left(\sum_{w \in V} count(w,c)\right) + |V|}$$

- **Probability of a word(feature) given the class (if feature is absent in the class) is calculated as follows,**

P(w|c) = $\frac{1}{\Sigma(count(w,c))+(vocab\_length)}$;

- **The reason for using log-probabilities is that when we multiply many numbers ranging between 0 to 1 the result will be zero as the compiler data-types cannot hold such very small values.**

- count(w,c) is nothing but the term-frequency of the term "t" w.r.t class "c". Σ(count(w,c)) is defined as the count of all the words in that class, and can also be said as the sum of all term frequencies.

- Then we sort the obtained log-probabilities for the test document w.r.t every class in descending order and assign the class label with the highest log-probability value as predicted class for the test document. The same procedure is carried out for all the test documents.

- Above procedure is carried out for various splits and feature selection counts.

- **RESULTS (PERFORMANCE AND PLOTS):**
- **The Overall Picture Of Performance:**

% OF TRAINING DATA CONSIDERED	% OF FEATURES(TF-IDF BASED) SELECTED/CLASS	ACCURACY
50 %	10 %	72.24000000000001 %
50 %	20 %	79.60000000000001 %
50 %	40 %	80.92 %
50 %	60 %	87.24 %
70 %	10 %	75.93333333333334 %
70 %	20 %	81.26666666666667 %
70 %	40 %	82.33333333333334 %
70 %	60 %	89.86666666666666 %
80 %	10 %	76.5 %
80 %	20 %	81.6 %
80 %	40 %	82.3 %
80 %	60 %	87.9 %

- **Confusion Matrices:**

  Train-Test split: 50:50

  Features selected : Top 10% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	423	84	2	2	0
rec.sport.hockey	0	499	0	1	0
sci.med	11	114	367	3	0
sci.space	19	107	2	359	0
talk.politics.misc	9	339	0	1	158

  Train-Test split: 50:50

  Features selected : Top 20% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	437	68	2	4	0
rec.sport.hockey	0	499	0	1	0
sci.med	6	42	444	3	0
sci.space	16	61	2	408	0
talk.politics.misc	9	278	13	5	202

  Train-Test split: 50:50

  Features selected : Top 40% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	464	43	0	4	0
rec.sport.hockey	0	499	0	1	0
sci.med	14	40	439	2	0
sci.space	29	48	2	408	0
talk.politics.misc	16	258	14	6	213

Train-Test split: 50:50
Features selected : Top 60% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	481	22	5	3	0
rec.sport.hockey	0	498	1	1	0
sci.med	8	23	456	7	1
sci.space	25	25	2	434	1
talk.politics.misc	15	150	22	8	312

Train-Test split: 70:30
Features selected : Top 10% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	223	60	0	1	0
rec.sport.hockey	1	298	0	0	0
sci.med	7	40	249	0	0
sci.space	5	53	1	238	0
talk.politics.misc	0	191	0	2	131

Train-Test split: 70:30
Features selected : Top 20% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	232	45	3	4	0
rec.sport.hockey	1	298	0	0	0
sci.med	7	28	261	0	0
sci.space	5	28	3	261	0
talk.politics.misc	0	146	8	3	167

Train-Test split: 70:30
Features selected : Top 40% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	252	28	0	4	0
rec.sport.hockey	1	298	0	0	0
sci.med	8	28	260	0	0
sci.space	8	30	3	256	0
talk.politics.misc	1	143	8	3	169

Train-Test split: 70:30
Features selected : Top 60% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	266	17	0	1	0
rec.sport.hockey	1	298	0	0	0
sci.med	7	12	274	3	0
sci.space	8	13	3	273	0
talk.politics.misc	4	63	12	8	237

Train-Test split: 80:20
Features selected : Top 10% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	204	0	0	1	0
rec.sport.hockey	2	192	0	0	0
sci.med	37	5	157	0	0
sci.space	39	3	1	153	0
talk.politics.misc	142	5	0	0	59

Train-Test split: 80:20
Features selected : Top 20% from Each Class

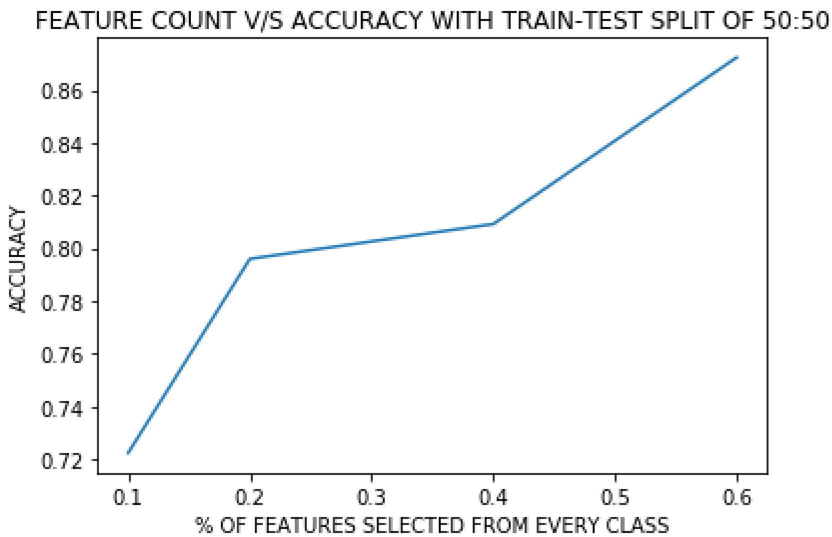
True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	201	0	1	3	0
rec.sport.hockey	2	192	0	0	0
sci.med	25	5	169	0	0
sci.space	25	3	1	167	0
talk.politics.misc	108	5	5	1	87

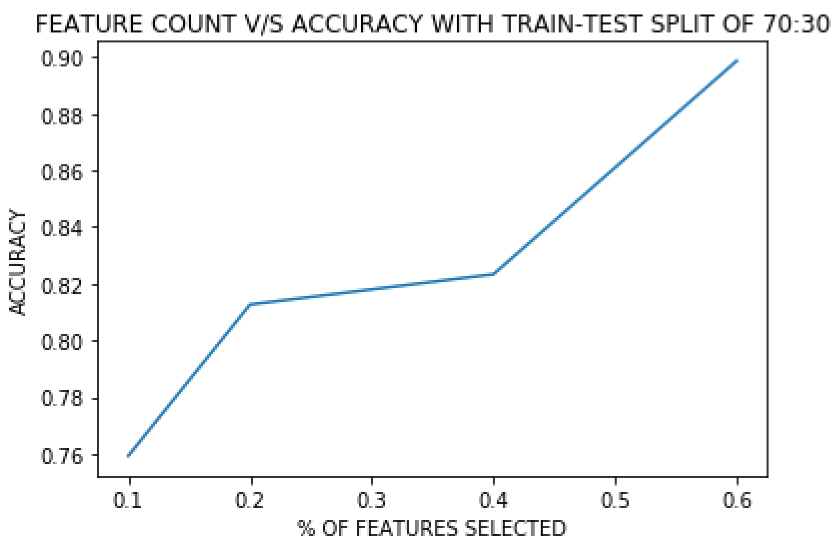
Train-Test split: 80:20
Features selected : Top 40% from Each Class

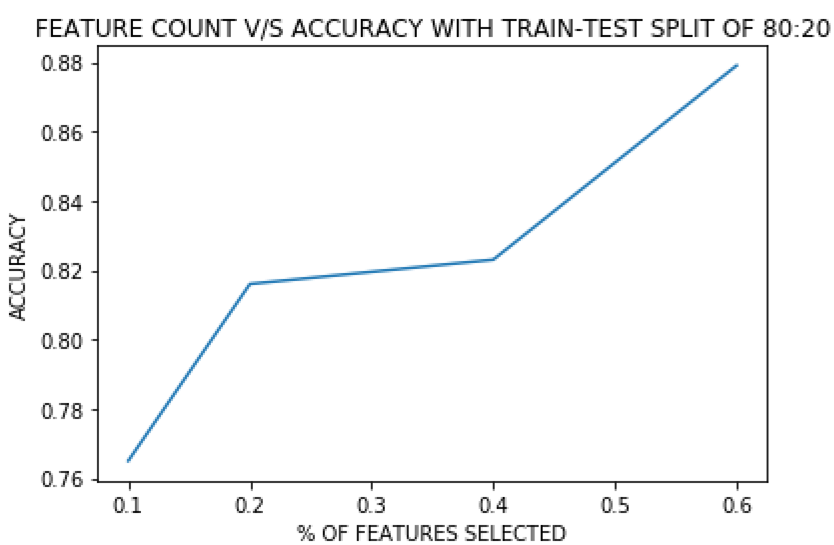
True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	203	1	0	1	0
rec.sport.hockey	0	194	0	0	0
sci.med	22	6	171	0	0
sci.space	25	4	1	166	0
talk.politics.misc	94	17	3	3	89

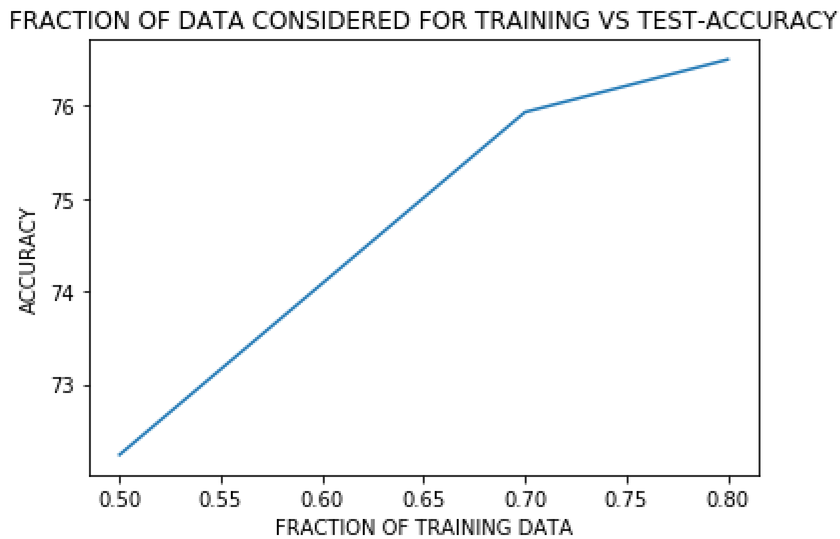
Train-Test split: 80:20
Features selected : Top 60% from Each Class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	202	3	0	0	0
rec.sport.hockey	0	194	0	0	0
sci.med	9	6	184	0	0
sci.space	15	5	2	174	0
talk.politics.misc	38	36	4	3	125


FEATURE COUNT V/S ACCURACY WITH TRAIN-TEST SPLIT OF 50:50
ACCURACY
0.72
0.74
0.76
0.78
0.80
0.82
0.84
0.86
0.1
0.2
0.3
0.4
0.5
0.6
% OF FEATURES SELECTED FROM EVERY CLASS


FEATURE COUNT V/S ACCURACY WITH TRAIN-TEST SPLIT OF 70:30
ACCURACY
0.76
0.78
0.80
0.82
0.84
0.86
0.88
0.90
0.1
0.2
0.3
0.4
0.5
0.6
% OF FEATURES SELECTED


FEATURE COUNT V/S ACCURACY WITH TRAIN-TEST SPLIT OF 80:20
ACCURACY
0.76
0.78
0.80
0.82
0.84
0.86
0.88
0.1
0.2
0.3
0.4
0.5
0.6
% OF FEATURES SELECTED



2. <u>MI Based Feature Selection</u>:

The splitting,reading of the documents is done in the same way as it was done for TF-IDF based feature selection. Then I extracted the vocabulary of words/features from the training documents and had built the MI table in which every row corresponds to a feature and every column corresponds to a class. **The entry in mi_table(t,c) indicates the count of documents belonging to class "c" and containing term "t"**. A mi table built by me for train-test split ratio of 50-50 is as follows (just displaying the first few rows),

	↓ term ; class ->	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
0	call	29	37	23	25	57
1	presentation	10	0	3	4	5
2	navy	2	0	1	3	3
3	scientific	9	0	33	29	5
4	visualization	13	0	0	0	0
5	virtual	14	1	2	2	0
6	reality	13	4	6	7	12
7	seminar	5	0	4	1	1
8	tuesday	4	16	5	6	5
9	june	5	1	12	8	5
10	carderock	3	0	1	0	0
11	division	9	51	9	6	9
12	naval	4	0	2	3	4
13	surface	22	2	10	32	1
14	warfare	3	0	0	1	4
15	center	35	20	18	37	16

- Now i calculated the MI for every word in the vocab w.r.t to every class using the following formula, the values of **N_{00} , N_{01} , N_{10} , N_{11} are calculated by extracting needful values from the table above.**

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

N_{10} : count of documents that contain t and are not in c
N_{11}: count of documents that contain t and are in c
N_{01} : count of documents that do not contain t and are in c
N_{00} : count of documents that do not contain t and are not in c
$N = N_{00} + N_{01} + N_{10} + N_{11}$.

- Now i selected the top k features from every class (k is entered by the user) and calculated the log-probability of test document w.r.t every class as, if the word/feature in the test document is present in the top K features (selected by mi score) of that class then we proceed to calculate the log probability value of that term with respect to that particular class. if the word is absent i used the technique of add-1 smoothing (laplace smoothing). The final log probability of the test document belonging to that class is the sum of log-probabilities of all words in that test document w.r.t that class summed up with the log of prior probability. This process is repeated for all the classes for a test document.

- **Probability of a word(feature) given the class (if feature is existing in the class) is calculated as follows,**

$$\hat{P}(w_i \mid c) = \frac{count(w_i,c)+1}{\sum_{w\in V}\left(count(w,c)+1\right)} = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right) + |V|}$$

- **Probability of a word(feature) given the class (if feature is absent in the class) is calculated as follows,**

$$\mathbf{P(w|c) = \frac{1}{\Sigma(count(w,c))+(vocab\_length)};}$$

- **RESULTS (PERFORMANCE & PLOTS):**

- **The Overall Picture Of Performance:**

% OF TRAINING DATA CONSIDERED	FEATURE SELECTION (MI)	ACCURACY
50%	5000	76.92%
50%	15000	83.16%
50%	25000	87.24%
70%	5000	78.60000000000001%
70%	15000	81.26666666666667%
70%	25000	88.6%
80%	5000	77.60000000000001%
80%	15000	82.1%
80%	25000	88.7%

- **Confusion Matrices:**

Train-Test split: 50:50

Features selected : Top 5000/class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	478	8	3	5	17
rec.sport.hockey	7	483	0	2	8
sci.med	81	45	217	8	144
sci.space	103	25	2	252	105
talk.politics.misc	8	5	0	1	493

Train-Test split: 50:50

Features selected : Top 15000/class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	489	7	1	5	9
rec.sport.hockey	4	486	0	3	7
sci.med	68	22	296	8	101
sci.space	79	19	1	317	71
talk.politics.misc	7	4	3	2	491

Train-Test split: 50:50
Features selected : Top 25000/class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	488	5	2	8	8
rec.sport.hockey	4	487	0	2	7
sci.med	47	18	348	10	72
sci.space	53	11	3	368	52
talk.politics.misc	6	6	2	3	490

Train-Test split: 70:30
Features selected : Top 5000/class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	253	2	4	10	15
rec.sport.hockey	11	265	0	1	22
sci.med	52	3	165	3	73
sci.space	44	2	0	180	71
talk.politics.misc	3	2	0	3	316

Train-Test split: 70:30
Features selected : Top 15000/class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	269	2	0	7	6
rec.sport.hockey	4	290	0	1	4
sci.med	44	9	167	1	75
sci.space	53	5	0	179	60
talk.politics.misc	7	1	1	1	314

Train-Test split: 70:30
Features selected : Top 25000/class

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	274	3	0	2	5
rec.sport.hockey	5	292	0	1	1
sci.med	22	4	230	3	37
sci.space	28	5	2	221	41
talk.politics.misc	4	4	2	2	312

Train-Test split: 80:20
Features selected : Top 5000/class

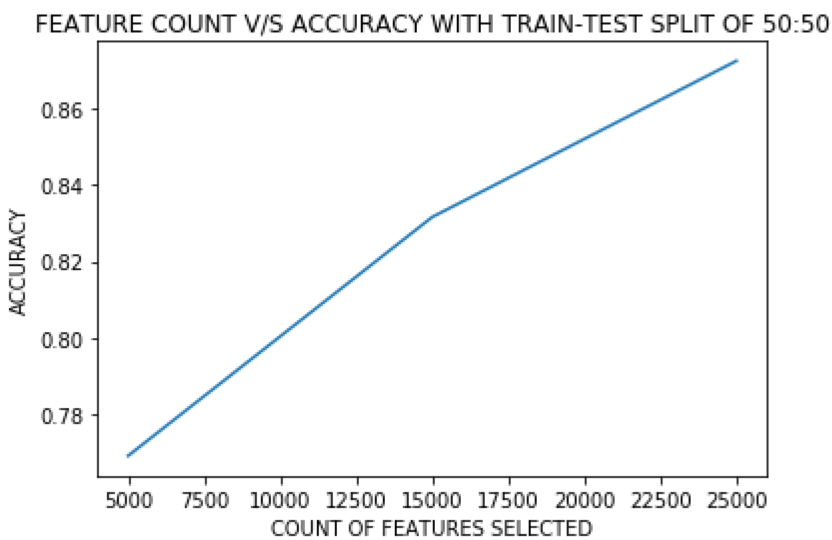
True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	181	3	2	7	12
rec.sport.hockey	6	179	0	1	8
sci.med	46	8	97	3	45
sci.space	33	2	0	119	42
talk.politics.misc	5	1	0	0	200

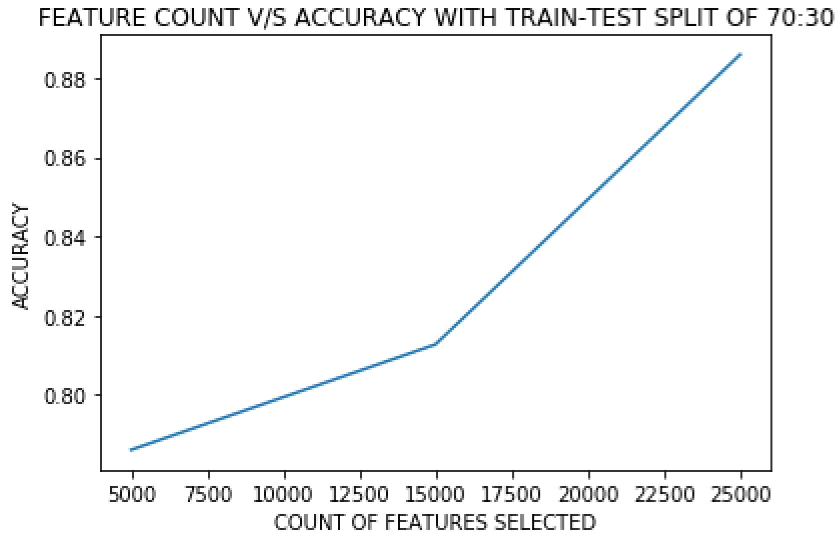
Train-Test split: 80:20
Features selected : Top 15000/class

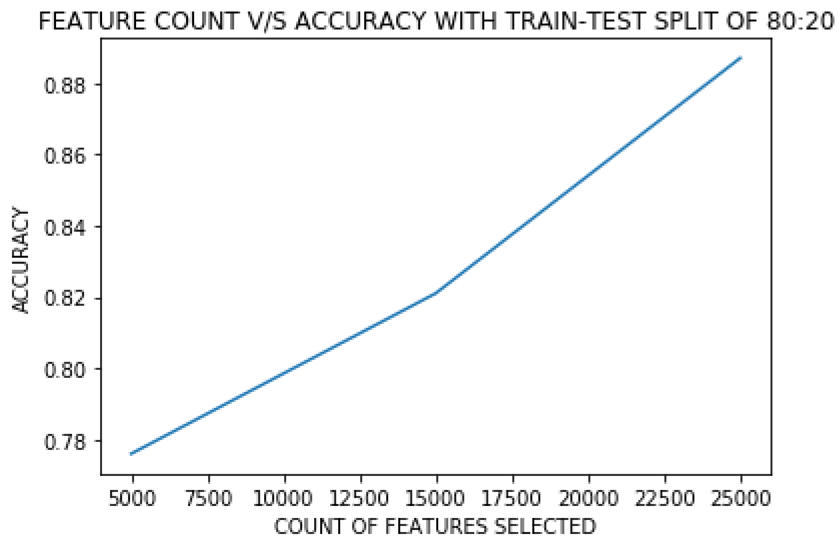
True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	196	3	0	3	3
rec.sport.hockey	2	191	0	0	1
sci.med	41	7	114	3	34
sci.space	37	2	0	123	34
talk.politics.misc	5	1	2	1	197

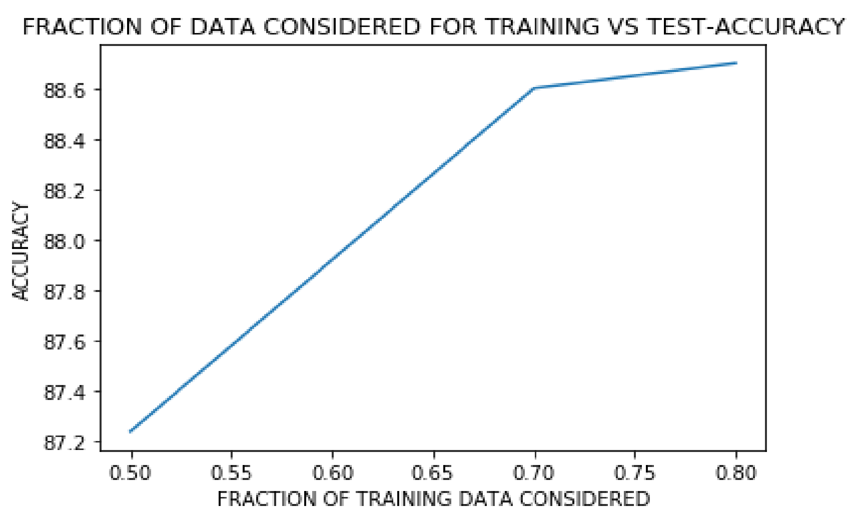
Train-Test split: 80:20
Features selected : Top 25000/class

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
True labels					
comp.graphics	190	5	1	5	4
rec.sport.hockey	1	191	0	0	2
sci.med	10	10	147	7	25
sci.space	12	4	1	162	17
talk.politics.misc	5	2	1	1	197




FEATURE COUNT V/S ACCURACY WITH TRAIN-TEST SPLIT OF 70:30
ACCURACY
0.88
0.86
0.84
0.82
0.80
5000
7500
10000
12500
15000
17500
20000
22500
25000
COUNT OF FEATURES SELECTED


FEATURE COUNT V/S ACCURACY WITH TRAIN-TEST SPLIT OF 80:20
ACCURACY
0.88
0.86
0.84
0.82
0.80
0.78
5000
7500
10000
12500
15000
17500
20000
22500
25000
COUNT OF FEATURES SELECTED


FRACTION OF DATA CONSIDERED FOR TRAINING VS TEST-ACCURACY
ACCURACY
88.6
88.4
88.2
88.0
87.8
87.6
87.4
87.2
0.50
0.55
0.60
0.65
0.70
0.75
0.80
FRACTION OF TRAINING DATA CONSIDERED

- **Methodology For Question-2 (K-NN; K-NEAREST NEIGHBOURS):**

- **TF-IDF Based Feature Selection:**

- Firstly, I read all the names of the documents from the corpus (mentioned folders) and performed a train-test split with a random state of value 0 (a particular random state is used since we obtain the same train-test splits of the data every time we run the code which eases our job in performance comparisons).

- Now I read the data corresponding to training documents "**doc-wise**" as well as "**class-wise**". In class-wise reading, I merged the data of documents belonging to the same class. **I have used the log-normal variate of TF (tf = 1+$\log_{10}$(tf)) and inverse variation of DF (df = $\log_{10}$(N/df)) where N is the total number of documents/classes considered for the corpus.** Then i had computed tf-idf values for both document-wise and also class-wise by multiplying corresponding term frequencies by their document frequencies. Now **i sorted the tf-idf values which were calculated class-wise descending order** to ease the feature selection.

- I selected the top k features from every class (k is entered by the user) and merged them into one named as "top_featrures". Now when building the training vectors I considered only the features/words present in the top_features and constructed a vector by extracting tf-idf values for every feature in the top_features, which were earlier calculated doc-wise. Similarly, in the test document for building test vectors only the words present in the top_features are considered with their respective term-frequency in the test document and IDF value calculated by doc-wise The ordering of the features is maintained in both the vectors (train and test).

- Now given the test vector and training vectors **i computed the cosine-similarity between the test vector and every training vector and sorted them in descending order**. Based on the value of K (in Knn) I considered the nearest K file's classes and took a majority vote amongst them to predict a class label for the given test vector. This process is repeated for every test vector.

- Cosine-Similarity between two vectors A & B is as follows,

$$\textbf{cosine-sim(A,B)} = \frac{\textbf{Dot-product(A,B)}}{|\textbf{A}| * |\textbf{B}|}$$; A and B are vectors of the same size.

- **RESULTS (PERFORMANCE & PLOTS) with TF-IDF BASED Feature Selection:**

- **The Overall view of performance:**

% OF TRAINING DATA CONSIDERED	K (in K-NN)	COUNT OF FEATURES(TF-IDF BASED)SELECTED/CLASS	TOTAL FEATURES	ACCURACY
50 %	1	1000	5000	91.04 %
50 %	3	1000	5000	91.4 %
50 %	5	1000	5000	91.24 %
70 %	1	1000	5000	91.93 %
70 %	3	1000	5000	92.2 %
70 %	5	1000	5000	92.0 %
80 %	1	1000	5000	91.6 %
80 %	3	1000	5000	91.8 %
80 %	5	1000	5000	91.8 %

- **Confusion Matrices:**
  Train-Test split: 50:50
  Features selected : Top 1000 features from each class ; total features considered:5000

CONFUSION MATRIX ON TEST DATA WITH K=1::

True labels	sci.space	sci.med	rec.sport.hockey	talk.politics.misc	comp.graphics
sci.space	478	7	6	14	6
sci.med	14	477	2	0	7
rec.sport.hockey	41	0	431	15	8
talk.politics.misc	48	2	6	427	4
comp.graphics	30	4	7	3	463

CONFUSION MATRIX ON TEST DATA WITH K=3::

True labels	sci.space	sci.med	rec.sport.hockey	talk.politics.misc	comp.graphics
sci.space	478	7	4	14	8
sci.med	17	477	1	2	3
rec.sport.hockey	42	1	434	13	5
talk.politics.misc	40	2	4	436	5
comp.graphics	30	8	4	5	460

CONFUSION MATRIX ON TEST DATA WITH K=5::

True labels	sci.space	sci.med	rec.sport.hockey	talk.politics.misc	comp.graphics
sci.space	487	7	5	8	4
sci.med	15	484	0	0	1
rec.sport.hockey	59	2	418	10	6
talk.politics.misc	47	2	2	432	4
comp.graphics	34	8	1	4	460

Train-Test split: 70:30

Features selected : Top 1000 features from each class ; total features considered:5000

CONFUSION MATRIX ON TEST DATA WITH K=1::

True labels	comp.graphics	sci.med	sci.space	rec.sport.hockey	talk.politics.misc
comp.graphics	276	1	1	2	4
sci.med	11	284	3	0	1
sci.space	31	1	256	5	3
rec.sport.hockey	26	1	3	265	2
talk.politics.misc	18	0	4	4	298

CONFUSION MATRIX ON TEST DATA WITH K=3::

True labels	comp.graphics	sci.med	sci.space	rec.sport.hockey	talk.politics.misc
comp.graphics	276	1	1	0	6
sci.med	11	286	1	0	1
sci.space	34	0	256	4	2
rec.sport.hockey	23	1	4	267	2
talk.politics.misc	18	0	4	4	298

CONFUSION MATRIX ON TEST DATA WITH K=5::

True labels	comp.graphics	sci.med	sci.space	rec.sport.hockey	talk.politics.misc
comp.graphics	277	1	1	0	5
sci.med	11	287	0	0	1
sci.space	37	0	254	2	3
rec.sport.hockey	24	1	5	265	2
talk.politics.misc	21	0	2	4	297

Train-Test split: 80:20
Features selected : Top 1000 features from each class ; total features considered:5000

CONFUSION MATRIX ON TEST DATA WITH K=1::

True labels	rec.sport.hockey	sci.space	sci.med	comp.graphics	talk.politics.misc
rec.sport.hockey	199	0	0	3	3
sci.space	4	187	3	0	0
sci.med	21	1	169	4	4
comp.graphics	16	1	1	176	2
talk.politics.misc	16	0	2	3	185

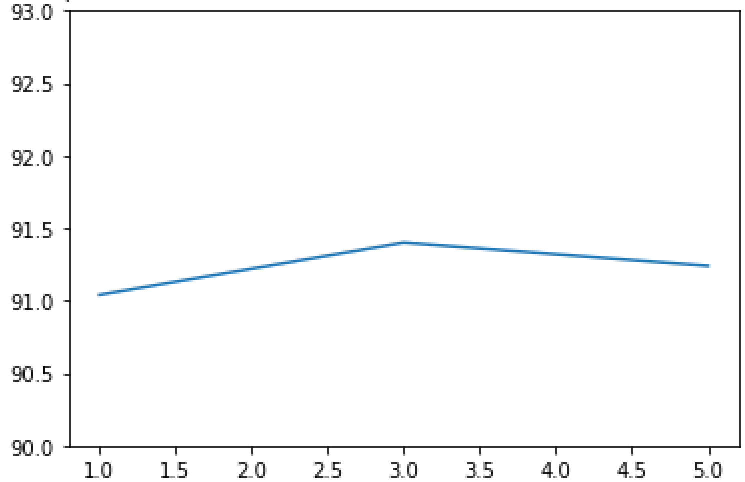
CONFUSION MATRIX ON TEST DATA WITH K=3::

True labels	rec.sport.hockey	sci.space	sci.med	comp.graphics	talk.politics.misc
rec.sport.hockey	199	0	0	2	4
sci.space	4	188	2	0	0
sci.med	21	3	168	4	3
comp.graphics	13	1	2	178	2
talk.politics.misc	16	0	2	3	185

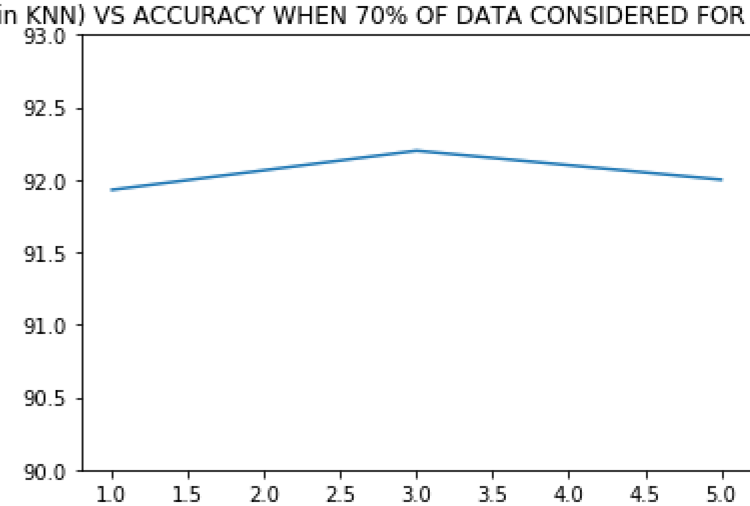
CONFUSION MATRIX ON TEST DATA WITH K=5::

True labels	rec.sport.hockey	sci.space	sci.med	comp.graphics	talk.politics.misc
rec.sport.hockey	199	1	0	1	4
sci.space	4	190	0	0	0
sci.med	25	1	168	2	3
comp.graphics	15	1	2	176	2
talk.politics.misc	17	0	1	3	185

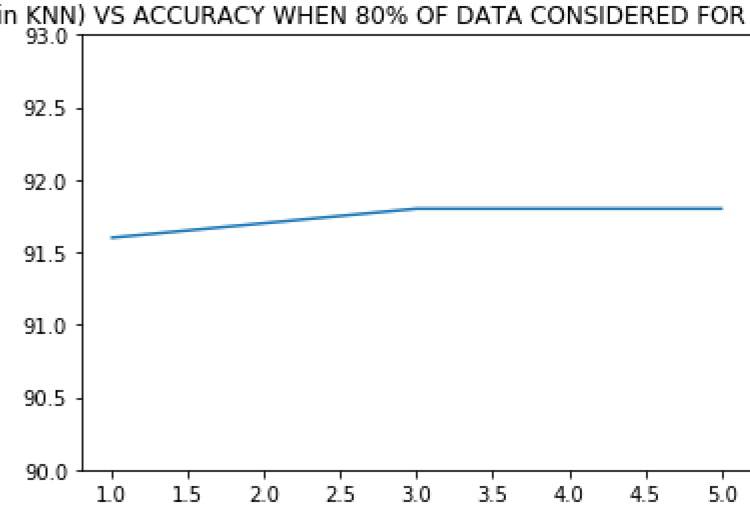
K (in KNN) VS ACCURACY WHEN 50% OF DATA CONSIDERED FOR TRAINING



K (in KNN) VS ACCURACY WHEN 70% OF DATA CONSIDERED FOR TRAINING



K (in KNN) VS ACCURACY WHEN 80% OF DATA CONSIDERED FOR TRAINING



- **<u>MI (MUTUAL INFORMATION) Based Feature Selection:</u>**

- In the MI based feature selection I used the mutual information based feature selection **(detailed explanation is written above in naive bayes section)** and followed the same procedure as above in building the training and test vectors.

- The same similarity metric "cosine-similarity" is used in identifying the closer documents related to the test document and thereby predicting the labels of the class by using majority vote.

- Mi based feature selection clearly outperforms TF-IDF based feature selection. Even with very few features/class i achieved a good performance because Mi considers the features globally whereas in TF-IDF we consider the features local to the class.

- **<u>RESULTS (PERFORMANCE AND PLOTS)</u>**

% OF TRAINING DATA CONSIDERED	K (in K-NN)	COUNT OF FEATURES(MI BASED)SELECTED/CLASS	TOTAL FEATURES	ACCURACY
50 %	1	300	1500	91.84 %
50 %	3	300	1500	91.92 %
50 %	5	300	1500	92.88 %
70 %	1	300	1500	92.53 %
70 %	3	300	1500	93.06 %
70 %	5	300	1500	93.2 %
80 %	1	300	1500	93.8 %
80 %	3	300	1500	93.0 %
80 %	5	300	1500	92.7 %

- **Confusion Matrices:**

Train-Test Split: 50:50
Features selected : Top 300 features from each class ; total features considered:1500

CONFUSION MATRIX ON TEST DATA WITH K=1::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	458	3	21	21	8
rec.sport.hockey	6	483	4	4	3
sci.med	27	6	451	5	6
sci.space	25	5	18	431	8
talk.politics.misc	9	1	18	6	473

CONFUSION MATRIX ON TEST DATA WITH K=3::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	455	3	15	32	6
rec.sport.hockey	2	485	1	11	1
sci.med	24	3	438	22	8
sci.space	21	4	10	444	8
talk.politics.misc	7	3	7	14	476

CONFUSION MATRIX ON TEST DATA WITH K=5::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	464	3	10	27	7
rec.sport.hockey	3	489	3	4	1
sci.med	19	3	448	18	7
sci.space	18	5	9	446	9
talk.politics.misc	4	7	8	13	475

Train-Test Split: 70:30

Features selected : Top 300 features from each class ; total features considered:1500

CONFUSION MATRIX ON TEST DATA WITH K=1::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	258	0	12	13	1
rec.sport.hockey	4	287	4	0	4
sci.med	14	0	274	3	5
sci.space	10	2	9	272	4
talk.politics.misc	11	0	10	6	297

CONFUSION MATRIX ON TEST DATA WITH K=3::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	270	0	5	6	3
rec.sport.hockey	8	288	1	1	1
sci.med	17	2	275	1	1
sci.space	19	2	4	266	6
talk.politics.misc	15	2	6	4	297

CONFUSION MATRIX ON TEST DATA WITH K=5::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	269	1	8	4	2
rec.sport.hockey	3	291	2	1	2
sci.med	19	2	273	1	1
sci.space	20	1	2	266	8
talk.politics.misc	12	2	7	4	299

Train-Test Split: 80:20

Features selected : Top 300 features from each class ; total features considered:1500

CONFUSION MATRIX ON TEST DATA WITH K=1::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	189	0	5	10	1
rec.sport.hockey	2	190	1	0	1
sci.med	10	1	185	2	1
sci.space	4	2	4	186	0
talk.politics.misc	7	1	6	4	188

CONFUSION MATRIX ON TEST DATA WITH K=3::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	190	7	4	3	1
rec.sport.hockey	1	191	2	0	0
sci.med	8	6	183	1	1
sci.space	10	5	2	178	1
talk.politics.misc	5	9	2	2	188

CONFUSION MATRIX ON TEST DATA WITH K=5::

True labels	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	192	5	3	3	2
rec.sport.hockey	1	192	1	0	0
sci.med	11	8	178	1	1
sci.space	10	5	2	177	2
talk.politics.misc	7	7	2	2	188