Assignment 5

Submitted by,
Swati Verma
MT19073

Note:

1. **All the necessary comments are written in the code you can refer it.**
2. **How the code is working i.e. methodology is written in Readme file.**
3. **Analysis is written at the end.**

Question 1. Naïve Bayes

1. Training and testing documents are split in **80:20 ratio.**

```
Enter the percent of documents you want in training set:
80
```

```
print("Number of documents in training set is: ",len(train1))
print("Number of documents in testing set is: ",len(test1))
```

```
Number of documents in training set is:  3996
Number of documents in testing set is:  999
```

2. Top 40 % tokens for class **comp.graphics** using mutual information as feature selection.

```
print("Top",k,"% terms using MI for feature selection for comp.graphics are: \n ",mi_final['comp.graphics'])
```

```
Top 40 % terms using MI for feature selection for comp.graphics are:
  {'line': 1, 'subject': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1, 'date': 1, 'organization': 1, 'apr': 1, 'gmt': 1, 'co
mpgraphics': 1, 'one thousand, nine hundred and ninety-three': 1, 'reference': 1, 'sender': 1, 'nntppostinghost': 1, 'univers
ity': 1, 'one': 1, 'ninety-three': 1, 'writes': 1, 'graphic': 1, 'would': 1, 'know': 1, 'xref': 1, 'cantaloupesrvcscmuedu':
1, 'article': 1, 'like': 1, 'thanks': 1, 'anyone': 1, 'file': 1, 'two': 1, 'image': 1, 'email': 1, 'help': 1, 'need': 1, 'ple
ase': 1, 'get': 1, 'computer': 1, 'may': 1, 'program': 1, 'system': 1, 'dont': 1, 'use': 1, 'im': 1, 'news': 1, 'also': 1, 'c
ould': 1, 'looking': 1, 'distribution': 1, 'replyto': 1, 'version': 1, 'time': 1, 'keywords': 1, 'think': 1, 'sixteen': 1, 'f
ormat': 1, 'eleven': 1, 'find': 1, 'world': 1, 'using': 1, '3d': 1, 'good': 1, 'work': 1, 'twenty': 1, 'software': 1, 'new':
1, 'problem': 1, 'much': 1, 'c': 1, 'way': 1, 'fifteen': 1, 'available': 1, 'information': 1, 'ive': 1, 'three': 1, 'scienc
e': 1, 'many': 1, 'want': 1, 'make': 1, 'usenet': 1, 'six': 1, 'color': 1, 'bit': 1, 'point': 1, 'code': 1, 'see': 1, 'hi':
1, 'well': 1, 'thing': 1, 'twenty-one': 1, 'window': 1, 'used': 1, 'group': 1, 'something': 1, 'number': 1, 'ftp': 1, 'got':
1, 'four': 1, 'etc': 1, 'go': 1, 'question': 1, 'people': 1, 'take': 1, 'post': 1, 'read': 1, 'look': 1, 'sun': 1, 'say': 1,
'advance': 1, 'ten': 1, 'tue': 1, 'even': 1, 'five': 1, 'wed': 1, 'algorithm': 1, 'package': 1, 'eight': 1, 'nineteen': 1, 't
wenty-four': 1, 'since': 1, 'fax': 1, 'thirty': 1, 'site': 1, 'interested': 1, 'first': 1, 'seventeen': 1, 'lot': 1, 'suppor
t': 1, 'twenty-two': 1, 'pc': 1, 'convert': 1, 'xnewsreader': 1, 'u': 1, 'tin': 1, 'year': 1, 'address': 1, 'twelve': 1, 'bes
t': 1, 'gif': 1, 'cant': 1, 'write': 1, 'library': 1, 'source': 1, 'fourteen': 1, 'display': 1, 'book': 1, 'someone': 1, 'x':
1, 'another': 1, 'run': 1, 'better': 1, 'thu': 1, 'twenty-seven': 1, 'give': 1, 'video': 1, 'center': 1, 'info': 1, 'card':
1, 'data': 1, 'twenty-three': 1, 'sure': 1, 'able': 1, 'fri': 1, 'still': 1, 'else': 1, 'twenty-eight': 1, 'seven': 1, 'diffe
rent': 1, 'thirteen': 1, 'user': 1, 'let': 1, 'come': 1, 'state': 1, 'wrote': 1, 'internet': 1, 'might': 1, 'dept': 1, 'try':
```

3. Top 40 % tokens for class **sci.med** using mutual information as feature selection.

```
print("Top",k,"5 terms using MI for feature selection for sci.med are: \n ",mi_final['sci.med'])
```

```
Top 40 5 terms using MI for feature selection for sci.med are:
  {'one': 1, 'subject': 1, 'date': 1, 'organization': 1, 'line': 1, 'newsgroups': 1, 'apr': 1, 'path': 1, 'messageid': 1, 'gm
t': 1, 'one thousand, nine hundred and ninety-three': 1, 'reference': 1, 'article': 1, 'writes': 1, 'would': 1, 'medical': 1,
'also': 1, 'sender': 1, 'university': 1, 'get': 1, 'two': 1, 'know': 1, 'dont': 1, 'use': 1, 'like': 1, 'people': 1, 'time':
1, 'ninety-three': 1, 'may': 1, 'food': 1, 'study': 1, 'year': 1, 'cancer': 1, 'new': 1, 'problem': 1, 'im': 1, 'nntppostingh
ost': 1, 'good': 1, 'system': 1, 'think': 1, 'effect': 1, 'science': 1, 'many': 1, 'six': 1, 'information': 1, 'research': 1,
'drug': 1, 'work': 1, 'well': 1, 'day': 1, 'cause': 1, 'much': 1, 'case': 1, 'three': 1, 'help': 1, 'number': 1, 'gordon': 1,
'make': 1, 'could': 1, 'u': 1, 'news': 1, 'medicine': 1, 'question': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'even': 1, 'wa
y': 1, 'bank': 1, 'used': 1, 'anyone': 1, 'ive': 1, 'say': 1, 'twenty': 1, 'since': 1, 'thing': 1, 'state': 1, 'take': 1, 'se
e': 1, 'want': 1, 'aid': 1, 'need': 1, 'center': 1, 'ten': 1, 'distribution': 1, 'first': 1, 'computer': 1, 'said': 1, 'reply
to': 1, 'page': 1, 'test': 1, 'program': 1, 'without': 1, 'result': 1, 'person': 1, 'four': 1, 'back': 1, 'might': 1, 'mont
h': 1, 'group': 1, 'five': 1, 'david': 1, 'really': 1, 'something': 1, 'go': 1, 'april': 1, 'world': 1, 'risk': 1, 'human':
1, 'body': 1, 'evidence': 1, 'part': 1, 'find': 1, 'lot': 1, 'level': 1, 'eleven': 1, 'available': 1, 'email': 1, 'read': 1,
'dr': 1, 'service': 1, 'seems': 1, 'national': 1, 'school': 1, 'long': 1, 'week': 1, 'newsletter': 1, 'using': 1, 'still': 1,
'steve': 1, 'anything': 1, 'believe': 1, 'high': 1, 'found': 1, 'c': 1, 'fifteen': 1, 'course': 1, 'disease': 1, 'pittsburg
h': 1, 'public': 1, 'fact': 1, 'try': 1, 'please': 1, 'every': 1, 'sure': 1, 'never': 1, 'volume': 1, 'thirty': 1, 'thats':
1, 'post': 1, 'point': 1, 'better': 1, 'thanks': 1, 'le': 1, 'woman': 1, 'probably': 1, 'however': 1, 'enough': 1, 'blood':
1, 'twenty-five': 1, 'done': 1, 'general': 1, 'etc': 1, 'right': 1, 'give': 1, 'different': 1, 'come': 1, 'cant': 1, 'among':
1, 'info': 1, 'someone': 1, 'twenty-one': 1, 'usenet': 1, 'type': 1, 'product': 1, 'j': 1, 'best': 1, 'weight': 1, 'tell': 1,
```

4. Top 40 % tokens for class **talk.politics.misc** using mutual information as feature selection.

```
print("Top",k,"% terms using MI for feature selection for talk.politics.misc are: \n ",mi_final['talk.politics.misc'])
```

```
Top 40 % terms using MI for feature selection for talk.politics.misc are:
  {'apr': 1, 'line': 1, 'subject': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1, 'date': 1, 'organization': 1, 'reference':
1, 'gmt': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'writes': 1, 'article': 1, 'one thousand, nine hundred and ninety-three':
1, 'one': 1, 'people': 1, 'sender': 1, 'would': 1, 'nntppostinghost': 1, 'ninety-three': 1, 'dont': 1, 'university': 1, 'lik
e': 1, 'u': 1, 'think': 1, 'time': 1, 'say': 1, 'make': 1, 'two': 1, 'know': 1, 'government': 1, 'get': 1, 'state': 1, 'new
s': 1, 'new': 1, 'right': 1, 'even': 1, 'way': 1, 'many': 1, 'much': 1, 'well': 1, 'see': 1, 'also': 1, 'im': 1, 'opinion':
1, 'twenty': 1, 'go': 1, 'want': 1, 'talkpoliticsmisc': 1, 'could': 1, 'good': 1, 'law': 1, 'thing': 1, 'case': 1, 'system':
1, 'fact': 1, 'believe': 1, 'usa': 1, 'distribution': 1, 'year': 1, 'child': 1, 'take': 1, 'really': 1, 'first': 1, 'may': 1,
'said': 1, 'fifteen': 1, 'world': 1, 'still': 1, 'let': 1, 'mean': 1, 'since': 1, 'american': 1, 'part': 1, 'point': 1, 'us
e': 1, 'sure': 1, 'sixteen': 1, 'day': 1, 'clinton': 1, 'going': 1, 'six': 1, 'look': 1, 'cant': 1, 'country': 1, 'gay': 1,
'last': 1, 'might': 1, 'another': 1, 'ten': 1, 'anything': 1, 'number': 1, 'never': 1, 'back': 1, 'need': 1, 'twenty-one': 1,
'replyto': 1, 'problem': 1, 'five': 1, 'made': 1, 'come': 1, 'question': 1, 'place': 1, 'something': 1, 'public': 1, 'reaso
n': 1, 'tue': 1, 'three': 1, 'usenet': 1, 'got': 1, 'president': 1, 'clayton': 1, 'life': 1, 'support': 1, 'someone': 1, 'any
one': 1, 'cramer': 1, 'actually': 1, 'person': 1, 'group': 1, 'issue': 1, 'give': 1, 'work': 1, 'four': 1, 'money': 1, 'nex
t': 1, 'c': 1, 'tax': 1, 'show': 1, 'without': 1, 'tell': 1, 'must': 1, 'kind': 1, 'every': 1, 'doesnt': 1, 'far': 1, 'care':
1, 'find': 1, 'force': 1, 'enough': 1, 'put': 1, 'study': 1, 'fbi': 1, 'thats': 1, 'crameroptilinkcom': 1, 'long': 1, 'lot':
1, 'service': 1, 'ive': 1, 'didnt': 1, 'used': 1, 'call': 1, 'nothing': 1, 'free': 1, 'others': 1, 'national': 1, 'thought':
1, 'course': 1, 'men': 1, 'pay': 1, 'idea': 1, 'mine': 1, 'please': 1, 'david': 1, 'better': 1, 'using': 1, 'try': 1, 'bill':
1, 'little': 1, 'homosexual': 1, 'followupto': 1, 'start': 1, 'different': 1, 'true': 1, 'twenty-three': 1, 'isnt': 1, 'mr':
```

5. Top 40 % tokens for class **rec.sport.hockey** using mutual information as feature selection.

```
print("Top",k,"% terms using MI for feature selection for rec.sport.hockey are: \n ",mi_final['rec.sport.hockey'])
```

```
Top 40 % terms using MI for feature selection for rec.sport.hockey are:
  {'recsporthockey': 1, 'apr': 1, 'subject': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1, 'date': 1, 'line': 1, 'organizati
on': 1, 'gmt': 1, 'one thousand, nine hundred and ninety-three': 1, 'reference': 1, 'game': 1, 'sender': 1, 'university': 1,
'writes': 1, 'team': 1, 'one': 1, 'nntppostinghost': 1, 'article': 1, 'hockey': 1, 'ninety-three': 1, 'would': 1, 'go': 1, 't
wo': 1, 'playoff': 1, 'get': 1, 'year': 1, 'player': 1, 'nhl': 1, 'like': 1, 'play': 1, 'time': 1, 'think': 1, 'dont': 1, 'la
st': 1, 'fan': 1, 'know': 1, 'good': 1, 'see': 1, 'three': 1, 'win': 1, 'season': 1, 'first': 1, 'six': 1, 'well': 1, 'twent
y': 1, 'twenty-one': 1, 'im': 1, 'news': 1, 'even': 1, 'five': 1, 'goal': 1, 'four': 1, 'twenty-three': 1, 'sixteen': 1, 'cu
p': 1, 'new': 1, 'could': 1, 'also': 1, 'way': 1, 'fifteen': 1, 'say': 1, 'back': 1, 'going': 1, 'let': 1, 'really': 1, 'nex
t': 1, 'wing': 1, 'make': 1, 'many': 1, 'league': 1, 'night': 1, 'much': 1, 'pittsburgh': 1, 'played': 1, 'people': 1, 'canad
a': 1, 'right': 1, 'take': 1, 'since': 1, 'best': 1, 'toronto': 1, 'world': 1, 'ten': 1, 'leaf': 1, 'distribution': 1, 'poin
t': 1, 'eleven': 1, 'got': 1, 'didnt': 1, 'fri': 1, 'bruin': 1, 'thing': 1, 'twenty-six': 1, 'cant': 1, 'great': 1, 'seven':
1, 'twenty-two': 1, 'better': 1, 'anyone': 1, 'eighteen': 1, 'come': 1, 'twenty-four': 1, 'look': 1, 'detroit': 1, 'give': 1,
'little': 1, 'replyto': 1, 'ranger': 1, 'mike': 1, 'usenet': 1, 'playing': 1, 'state': 1, 'second': 1, 'tue': 1, 'want': 1,
'may': 1, 'final': 1, 'another': 1, 'division': 1, 'guy': 1, 'u': 1, 'stanley': 1, 'still': 1, 'put': 1, 'john': 1, 'boston':
1, 'made': 1, 'goalie': 1, 'sure': 1, 'devil': 1, 'series': 1, 'need': 1, 'lot': 1, 'said': 1, 'never': 1, 'getting': 1, 'h
e': 1, 'bad': 1, 'science': 1, 'eight': 1, 'doesnt': 1, 'ice': 1, 'post': 1, 'show': 1, 'system': 1, 'blue': 1, 'espn': 1, 's
hot': 1, 'twelve': 1, 'thirty': 1, 'mon': 1, 'penguin': 1, 'coach': 1, 'buffalo': 1, 'usa': 1, 'might': 1, 'mean': 1, 'end':
1, 'something': 1, 'id': 1, 'seventeen': 1, 'pen': 1, 'nineteen': 1, 'thought': 1, 'wed': 1, 'probably': 1, 'score': 1, 'ques
tion': 1, 'thats': 1, 'contact': 1, 'red': 1, 'islander': 1, 'patrick': 1, 'computer': 1, 'name': 1, 'least': 1, 'twenty-fiv
```

6. Top 40 % tokens for class **sci.space** using mutual information as feature selection.

```
print("Top",k,"% terms using MI for feature selection for sci.space are: \n ",mi_final['sci.space'])
```

```
Top 40 % terms using MI for feature selection for sci.space are:
  {'subject': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1, 'date': 1, 'line': 1, 'organization': 1, 'apr': 1, 'scispace':
1, 'gmt': 1, 'one thousand, nine hundred and ninety-three': 1, 'reference': 1, 'writes': 1, 'space': 1, 'one': 1, 'article':
1, 'sender': 1, 'nntppostinghost': 1, 'would': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'university': 1, 'like': 1, 'may':
1, 'ninety-three': 1, 'get': 1, 'system': 1, 'two': 1, 'u': 1, 'know': 1, 'time': 1, 'distribution': 1, 'think': 1, 'dont':
1, 'also': 1, 'could': 1, 'year': 1, 'much': 1, 'thing': 1, 'orbit': 1, 'new': 1, 'make': 1, 'news': 1, 'see': 1, 'nasa': 1,
'world': 1, 'earth': 1, 'way': 1, 'well': 1, 'im': 1, 'first': 1, 'people': 1, 'day': 1, 'science': 1, 'might': 1, 'use': 1,
'go': 1, 'mission': 1, 'shuttle': 1, 'work': 1, 'idea': 1, 'launch': 1, 'even': 1, 'three': 1, 'something': 1, 'good': 1, 'si
x': 1, 'program': 1, 'need': 1, 'question': 1, 'moon': 1, 'high': 1, 'anyone': 1, 'since': 1, 'twenty': 1, 'many': 1, 'cost':
1, 'find': 1, 'sixteen': 1, 'better': 1, 'want': 1, 'long': 1, 'say': 1, 'pat': 1, 'going': 1, 'around': 1, 'sun': 1, 'proble
m': 1, 'project': 1, 'right': 1, 'back': 1, 'technology': 1, 'satellite': 1, 'henry': 1, 'tue': 1, 'usa': 1, 'look': 1, 'twen
ty-three': 1, 'thirty': 1, 'part': 1, 'come': 1, 'usenet': 1, 'give': 1, 'research': 1, 'put': 1, 'five': 1, 'large': 1, 'eno
ugh': 1, 'sky': 1, 'flight': 1, 'used': 1, 'still': 1, 'point': 1, 'fifteen': 1, 'using': 1, 'eleven': 1, 'thats': 1, 'spence
r': 1, 'spacecraft': 1, 'fri': 1, 'information': 1, 'lot': 1, 'ten': 1, 'access': 1, 'data': 1, 'solar': 1, 'twenty-two': 1,
'probably': 1, 'real': 1, 'someone': 1, 'sure': 1, 'henryzootorontoedu': 1, 'really': 1, 'twenty-one': 1, 'communication': 1,
'mean': 1, 'center': 1, 'state': 1, 'please': 1, 'power': 1, 'keywords': 1, 'take': 1, 'twenty-seven': 1, 'sci': 1, 'cant':
1, 'four': 1, 'computer': 1, 'course': 1, 'international': 1, 'different': 1, 'every': 1, 'said': 1, 'big': 1, 'actually': 1,
'however': 1, 'far': 1, 'last': 1, 'available': 1, 'read': 1, 'another': 1, 'help': 1, 'replyto': 1, 'id': 1, 'maybe': 1, 'ei
ght': 1, 'least': 1, 'believe': 1, 'twenty-six': 1, 'via': 1, 'made': 1, 'number': 1, 'done': 1, 'little': 1, 'possible': 1,
```

7. Top 40 % tokens for class **comp.graphics** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for comp.graphics are: \n ",tf_idf_final['comp.graphics'])
```

```
'function': 1, 'surface': 1, 'stuff': 1, 'twenty-one': 1, 'name': 1, 'include': 1, 'something': 1, 'output': 1, 'take': 1, 'p
rogramming': 1, 'archive': 1, 'year': 1, 'give': 1, 'ca': 1, 'got': 1, 'product': 1, 'fast': 1, 'change': 1, 'say': 1, 'hi':
1, 'colour': 1, 'called': 1, 'interested': 1, 'try': 1, 'since': 1, 'simple': 1, 'either': 1, 'access': 1, 'cant': 1, 'best':
1, 'phone': 1, 'view': 1, 'mean': 1, 'full': 1, 'following': 1, 'check': 1, 'call': 1, 'written': 1, 'space': 1, 'seventeen':
1, 'fourteen': 1, 'p': 1, 'utility': 1, 'edge': 1, 'advance': 1, 'value': 1, 'service': 1, 'original': 1, 'twelve': 1, 'ninet
een': 1, 'box': 1, 'around': 1, 'able': 1, 'wed': 1, 'text': 1, 'still': 1, 'steve': 1, 'let': 1, 'language': 1, 'area': 1,
'tue': 1, 'must': 1, 'conference': 1, 'usa': 1, 'network': 1, 'form': 1, 'disk': 1, 'current': 1, 'author': 1, 'twenty-seve
n': 1, 'come': 1, 'various': 1, 'thirteen': 1, 'however': 1, 'wrote': 1, 'size': 1, 'part': 1, 'might': 1, 'back': 1, 'tin':
1, 'someone': 1, 'online': 1, 'copy': 1, 'volume': 1, 'twenty-two': 1, 'several': 1, 'per': 1, 'else': 1, 'david': 1, 'anythi
ng': 1, 'xnewsreader': 1, 'robert': 1, 'resource': 1, 'least': 1, 'computing': 1, 'b': 1, 'sure': 1, 'public': 1, 'includes':
1, 'design': 1, 'there': 1, 'state': 1, 'place': 1, 'handle': 1, 'found': 1, 'nine': 1, 'twenty-three': 1, 'including': 1, 'i
d': 1, 'e': 1, 'technical': 1, 'real': 1, 'twenty-eight': 1, 'thu': 1, 'page': 1, 'order': 1, 'method': 1, 'level': 1, 'doesn
t': 1, 'case': 1, 'result': 1, 'speed': 1, 'section': 1, 'posting': 1, 'mark': 1, 'map': 1, 'last': 1, 'currently': 1, 'yes':
1, 'setting': 1, 'right': 1, 'provides': 1, 'memory': 1, 'device': 1, 'dept': 1, 'twenty-nine': 1, 'tel': 1, 'small': 1, 'req
uires': 1, 'possible': 1, 'national': 1, 'large': 1, 'general': 1, 'far': 1, 'company': 1, 'bb': 1, 'twenty-six': 1, 'rathe
r': 1, 'lab': 1, 'input': 1, 'answer': 1, 'useful': 1, 'running': 1, 'one hundred': 1, 'john': 1, 'fri': 1, 'cost': 1, 'anybo
dy': 1, 'processing': 1, 'twenty-five': 1, 'tell': 1, 'operation': 1, 'mon': 1, 'commercial': 1, 'working': 1, 'really': 1,
'probably': 1, 'n': 1, 'going': 1, 'canada': 1, 'paper': 1, 'others': 1, 'net': 1, 'listing': 1, 'eighteen': 1, 'appreciate
d': 1, 'v': 1, 'create': 1, 'needed': 1, 'ill': 1, 'draw': 1, 'summary': 1, 'purpose': 1, 'process': 1, 'plus': 1, 'institut
```

8. Top 40 % tokens for class **sci.med** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for sci.med are: \n ",tf_idf_final['sci.med'])
```

```
e': 1, 'got': 1, 'old': 1, 'robert': 1, 'put': 1, 'prevent': 1, 'stone': 1, 'response': 1, 'real': 1, 'eye': 1, 'college': 1,
'bad': 1, 'amount': 1, 'age': 1, 'rather': 1, 'quite': 1, 'either': 1, 'dept': 1, 'univ': 1, 'twenty-six': 1, 'technology':
1, 'several': 1, 'seen': 1, 'else': 1, 'tue': 1, 'twenty-three': 1, 'show': 1, 'low': 1, 'id': 1, 'around': 1, 'told': 1, 'me
thod': 1, 'heard': 1, 'twenty-nine': 1, 'twelve': 1, 'support': 1, 'picture': 1, 'interested': 1, 'increase': 1, 'fri': 1, 'a
lways': 1, 'natural': 1, 'let': 1, 'l': 1, 'heart': 1, 'far': 1, 'michael': 1, 'mark': 1, 'ill': 1, 'brain': 1, 'address': 1,
'word': 1, 'summary': 1, 'followupto': 1, 'following': 1, 'department': 1, 'york': 1, 'name': 1, 'minute': 1, 'known': 1, 'gi
ven': 1, 'end': 1, 'eighteen': 1, 'mon': 1, 'list': 1, 'isnt': 1, 'contact': 1, 'able': 1, 'trying': 1, 'practice': 1, 'exerc
ise': 1, 'claim': 1, 'software': 1, 'place': 1, 'city': 1, 'already': 1, 'advice': 1, 'united': 1, 'student': 1, 'process':
1, 'mind': 1, 'feel': 1, 'eight': 1, 'couple': 1, 'taken': 1, 'provide': 1, 'others': 1, 'life': 1, 'current': 1, 'thirteen':
1, 'term': 1, 'sound': 1, 'form': 1, 'certain': 1, 'call': 1, 'although': 1, 'yet': 1, 'wife': 1, 'wed': 1, 'understand': 1,
'twenty-four': 1, 'sleep': 1, 'seven': 1, 'looking': 1, 'large': 1, 'internet': 1, 'hour': 1, 'home': 1, 'company': 1, 'cause
d': 1, 'approach': 1, 'pm': 1, 'office': 1, 'next': 1, 'major': 1, 'knowledge': 1, 'including': 1, 'difference': 1, 'certainl
y': 1, 'ca': 1, 'based': 1, 'water': 1, 'version': 1, 'unless': 1, 'sort': 1, 'similar': 1, 'related': 1, 'procedure': 1, 'in
creased': 1, 'fourteen': 1, 'especially': 1, 'early': 1, 'didnt': 1, 'stuff': 1, 'small': 1, 'send': 1, 'sat': 1, 'left': 1,
'keywords': 1, 'hand': 1, 'factor': 1, 'cost': 1, 'user': 1, 'one thousand, nine hundred and ninety-two': 1, 'matter': 1, 'la
b': 1, 'check': 1, 'associated': 1, 'pain': 1, 'wrote': 1, 'washington': 1, 'tried': 1, 'set': 1, 'nine': 1, 'field': 1, 'sta
ndard': 1, 'specific': 1, 'family': 1, 'chemical': 1, 'california': 1, 'child': 1, 'thu': 1, 'thought': 1, 'r': 1, 'pretty':
1, 'fax': 1, 'due': 1, 'answer': 1, 'within': 1, 'sometimes': 1, 'n': 1, 'communication': 1, 'bill': 1, 'almost': 1, 'sense':
1, 'run': 1, 'remember': 1, 'recently': 1, 'poster': 1, 'network': 1, 'machine': 1, 'likely': 1, 'discussion': 1, 'communit
```

9. Top 40 % tokens for class **talk.politics.misc** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for talk.politics.misc are: \n ",tf_idf_final['talk.politics.misc'])
```

```
ree': 1, 'tell': 1, 'never': 1, 'white': 1, 'give': 1, 'united': 1, 'party': 1, 'free': 1, 'put': 1, 'got': 1, 'bill': 1, 'fi
ve': 1, 'idea': 1, 'distribution': 1, 'without': 1, 'yes': 1, 'another': 1, 'situation': 1, 'human': 1, 'someone': 1, 'ever
y': 1, 'show': 1, 'next': 1, 'six': 1, 'package': 1, 'working': 1, 'whether': 1, 'sixteen': 1, 'try': 1, 'four': 1, 'didnt':
1, 'administration': 1, 'press': 1, 'member': 1, 'actually': 1, 'report': 1, 'plan': 1, 'anyone': 1, 'used': 1, 'statement':
1, 'doesnt': 1, 'must': 1, 'done': 1, 'nothing': 1, 'he': 1, 'claim': 1, 'c': 1, 'thought': 1, 'enough': 1, 'others': 1, 'fa
r': 1, 'david': 1, 'twenty-one': 1, 'long': 1, 'evidence': 1, 'different': 1, 'already': 1, 'argument': 1, 'youre': 1, 'star
t': 1, 'important': 1, 'find': 1, 'word': 1, 'there': 1, 'department': 1, 'option': 1, 'million': 1, 'information': 1, 'littl
e': 1, 'ive': 1, 'true': 1, 'le': 1, 'call': 1, 'course': 1, 'control': 1, 'woman': 1, 'yet': 1, 'usenet': 1, 'replyto': 1,
'policy': 1, 'keep': 1, 'death': 1, 'trying': 1, 'community': 1, 'russia': 1, 'example': 1, 'business': 1, 'area': 1, 'view':
1, 'tue': 1, 'steve': 1, 'society': 1, 'help': 1, 'wrong': 1, 'using': 1, 'post': 1, 'better': 1, 'order': 1, 'certainly': 1,
'probably': 1, 'least': 1, 'ever': 1, 'answer': 1, 'secretary': 1, 'school': 1, 'official': 1, 'isnt': 1, 'interest': 1, 'tho
ugh': 1, 'saying': 1, 'however': 1, 'coverage': 1, 'read': 1, 'april': 1, 'sort': 1, 'set': 1, 'week': 1, 'please': 1, 'cos
t': 1, 'seems': 1, 'mark': 1, 'man': 1, 'george': 1, 'around': 1, 'child': 1, 'considered': 1, 'best': 1, 'always': 1, 'see
n': 1, 'month': 1, 'end': 1, 'education': 1, 'ask': 1, 'yesterday': 1, 'ill': 1, 'either': 1, 'able': 1, 'tax': 1, 'mine': 1,
'else': 1, 'america': 1, 'fund': 1, 'feel': 1, 'whole': 1, 'percentage': 1, 'hand': 1, 'theyre': 1, 'term': 1, 'south': 1, 'r
eal': 1, 'id': 1, 'continue': 1, 'black': 1, 'started': 1, 'population': 1, 'twenty-three': 1, 'rather': 1, 'live': 1, 'leve
l': 1, 'economic': 1, 'company': 1, 'given': 1, 'getting': 1, 'etc': 1, 'young': 1, 'office': 1, 'individual': 1, 'agree': 1,
'wanted': 1, 'simply': 1, 'possible': 1, 'nation': 1, 'happened': 1, 'act': 1, 'washington': 1, 'talk': 1, 'followupto': 1,
'ago': 1, 'myers': 1, 'side': 1, 'quite': 1, 'obviously': 1, 'medium': 1, 'matter': 1, 'making': 1, 'effort': 1, 'corporatio
```

10. Top 40 % tokens for class **rec.sport.hockey** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for rec.sport.hockey are: \n ",tf_idf_final['rec.sport.hockey'])
```

```
Top 40 % terms using tf_idf for feature selection for rec.sport.hockey are:
  {'one': 1, 'zero': 1, 'two': 1, 'game': 1, 'three': 1, 'team': 1, 'line': 1, 'subject': 1, 'apr': 1, 'date': 1, 'organizati
on': 1, 'newsgroups': 1, 'messageid': 1, 'path': 1, 'four': 1, 'gmt': 1, 'one thousand, nine hundred and ninety-three': 1, 'f
ive': 1, 'player': 1, 'six': 1, 'reference': 1, 'university': 1, 'writes': 1, 'twenty-five': 1, 'go': 1, 'would': 1, 'seven':
1, 'play': 1, 'year': 1, 'sender': 1, 'get': 1, 'article': 1, 'think': 1, 'goal': 1, 'nntppostinghost': 1, 'time': 1, 'like':
1, 'win': 1, 'fan': 1, 'first': 1, 'dont': 1, 'ninety-three': 1, 'period': 1, 'new': 1, 'ten': 1, 'good': 1, 'know': 1, 'five
hundred and fifty': 1, 'last': 1, 'see': 1, 'twenty': 1, 'v': 1, 'twenty-one': 1, 'league': 1, 'leaf': 1, 'eleven': 1, 'sho
t': 1, 'eight': 1, 'wing': 1, 'la': 1, 'im': 1, 'blue': 1, 'well': 1, 'pittsburgh': 1, 'nine': 1, 'division': 1, 'point': 1,
'boston': 1, 'twenty-three': 1, 'sixteen': 1, 'fifteen': 1, 'news': 1, 'even': 1, 'canada': 1, 'back': 1, 'second': 1, 'reall
y': 1, 'way': 1, 'toronto': 1, 'also': 1, 'say': 1, 'john': 1, 'right': 1, 'people': 1, 'next': 1, 'could': 1, 'l': 1, 'eight
een': 1, 'best': 1, 'make': 1, 'let': 1, 'mike': 1, 'twelve': 1, 'twenty-six': 1, 'night': 1, 'series': 1, 'going': 1, 'apri
l': 1, 'played': 1, 'st': 1, 'twenty-four': 1, 'take': 1, 'final': 1, 'much': 1, 'ice': 1, 'twenty-two': 1, 'thirty': 1, 'man
y': 1, 'got': 1, 'better': 1, 'great': 1, 'world': 1, 'didnt': 1, 'mon': 1, 'guy': 1, 'thirteen': 1, 'name': 1, 'said': 1, 'n
ineteen': 1, 'since': 1, 'fourteen': 1, 'van': 1, 'jet': 1, 'roger': 1, 'lead': 1, 'chicago': 1, 'made': 1, 'cant': 1, 'sho
w': 1, 'usa': 1, 'seventeen': 1, 'power': 1, 'may': 1, 'thing': 1, 'state': 1, 'pick': 1, 'never': 1, 'he': 1, 'come': 1, 'pl
aying': 1, 'look': 1, 'city': 1, 'save': 1, 'draft': 1, 'vancouver': 1, 'san': 1, 'give': 1, 'coverage': 1, 'thirty-three':
1, 'thirty-one': 1, 'louis': 1, 'little': 1, 'twenty-eight': 1, 'still': 1, 'red': 1, 'nj': 1, 'season': 1, 'usenet': 1, 'pos
t': 1, 'put': 1, 'fri': 1, 'bad': 1, 'u': 1, 'end': 1, 'distribution': 1, 'third': 1, 'list': 1, 'king': 1, 'doesnt': 1, 'can
adian': 1, 'another': 1, 'bob': 1, 'anyone': 1, 'net': 1, 'twenty-nine': 1, 'tue': 1, 'round': 1, 'mark': 1, 'sure': 1, 'sta
```

11. Top 40 % tokens for class **sci.space** using TF-IDF as feature selection.

```
: print("Top",k,"% terms using tf_idf for feature selection for sci.space are: \n ",tf_idf_final['sci.space'])
```

```
Top 40 % terms using tf_idf for feature selection for sci.space are:
  {'space': 1, 'one': 1, 'line': 1, 'date': 1, 'subject': 1, 'organization': 1, 'would': 1, 'newsgroups': 1, 'path': 1, 'mess
ageid': 1, 'one thousand, nine hundred and ninety-three': 1, 'apr': 1, 'gmt': 1, 'reference': 1, 'writes': 1, 'system': 1, 'a
rticle': 1, 'two': 1, 'like': 1, 'u': 1, 'launch': 1, 'mission': 1, 'may': 1, 'nasa': 1, 'time': 1, 'earth': 1, 'sender': 1,
'year': 1, 'nntppostinghost': 1, 'also': 1, 'get': 1, 'satellite': 1, 'university': 1, 'dont': 1, 'first': 1, 'could': 1, 'ne
w': 1, 'three': 1, 'think': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'program': 1, 'data': 1, 'planet': 1, 'ninety-three':
1, 'know': 1, 'much': 1, 'thing': 1, 'science': 1, 'cost': 1, 'distribution': 1, 'see': 1, 'make': 1, 'people': 1, 'well': 1,
'world': 1, 'day': 1, 'work': 1, 'use': 1, 'pat': 1, 'way': 1, 'station': 1, 'project': 1, 'news': 1, 'center': 1, 'technolog
y': 1, 'flight': 1, 'right': 1, 'idea': 1, 'im': 1, 'even': 1, 'part': 1, 'problem': 1, 'sun': 1, 'six': 1, 'five': 1, 'goo
d': 1, 'information': 1, 'might': 1, 'go': 1, 'research': 1, 'question': 1, 'need': 1, 'since': 1, 'around': 1, 'high': 1, 'b
ack': 1, 'power': 1, 'something': 1, 'want': 1, 'many': 1, 'available': 1, 'long': 1, 'large': 1, 'surface': 1, 'four': 1, 't
en': 1, 'light': 1, 'used': 1, 'astronomy': 1, 'star': 1, 'better': 1, 'find': 1, 'going': 1, 'still': 1, 'enough': 1, 'desig
n': 1, 'thirty': 1, 'say': 1, 'look': 1, 'small': 1, 'commercial': 1, 'april': 1, 'twenty': 1, 'said': 1, 'point': 1, 'anyon
e': 1, 'national': 1, 'life': 1, 'twenty-three': 1, 'object': 1, 'put': 1, 'give': 1, 'sixteen': 1, 'mass': 1, 'fifteen': 1,
'degree': 1, 'come': 1, 'international': 1, 'thats': 1, 'software': 1, 'human': 1, 'eight': 1, 'book': 1, 'using': 1, 'sourc
e': 1, 'really': 1, 'real': 1, 'money': 1, 'made': 1, 'group': 1, 'access': 1, 'usa': 1, 'support': 1, 'test': 1, 'number':
1, 'eleven': 1, 'usenet': 1, 'communication': 1, 'field': 1, 'different': 1, 'tue': 1, 'service': 1, 'lot': 1, 'material': 1,
'last': 1, 'institute': 1, 'development': 1, 'moon': 1, 'twenty-two': 1, 'toronto': 1, 'post': 1, 'far': 1, 'computer': 1, 'n
ext': 1, 'big': 1, 'take': 1, 'ray': 1, 'probably': 1, 'possible': 1, 'option': 1, 'air': 1, 'please': 1, 'etc': 1, 'set': 1,
```

12. Prior probability for each class.

```
print("Prior probability class wise are: \n" ,prior_prob)
```

```
Prior probability class wise are:
 {'comp.graphics': 0.1961961961961962, 'sci.med': 0.1981981981981982, 'talk.politics.misc': 0.20245245245245244, 'rec.sport.hoc
key': 0.2002002002002002, 'sci.space': 0.20295295295295296}
```

13. Actual classes for testing documents.

```
print("Actual classes for test documents are: \n" ,actual)
actual1=[]
for i in actual:
    actual1.append(actual[i])
# print(actual1)
```

```
Actual classes for test documents are:
 {'61026': 'sci.space', '38784': 'comp.graphics', '176952': 'talk.politics.misc', '59263': 'sci.med', '58864': 'sci.med', '54
704': 'rec.sport.hockey', '178425': 'talk.politics.misc', '54013': 'rec.sport.hockey', '59341': 'sci.med', '178628': 'talk.po
litics.misc', '53654': 'rec.sport.hockey', '53748': 'rec.sport.hockey', '53722': 'rec.sport.hockey', '38762': 'comp.graphic
s', '59491': 'sci.med', '54510': 'rec.sport.hockey', '39006': 'comp.graphics', '38738': 'comp.graphics', '61507': 'sci.spac
e', '53797': 'rec.sport.hockey', '179115': 'talk.politics.misc', '53655': 'rec.sport.hockey', '53715': 'rec.sport.hockey', '5
8884': 'sci.med', '61215': 'sci.space', '59372': 'sci.med', '54756': 'rec.sport.hockey', '178883': 'talk.politics.misc', '379
38': 'comp.graphics', '52566': 'rec.sport.hockey', '179008': 'talk.politics.misc', '54254': 'rec.sport.hockey', '58139': 'sc
i.med', '61316': 'sci.space', '59204': 'sci.med', '58811': 'sci.med', '54231': 'rec.sport.hockey', '60810': 'sci.space', '601
91': 'sci.space', '53798': 'rec.sport.hockey', '39627': 'comp.graphics', '59644': 'sci.med', '179004': 'talk.politics.misc',
'59430': 'sci.med', '61086': 'sci.space', '59287': 'sci.med', '61179': 'sci.space', '58066': 'sci.med', '58147': 'sci.med',
'178601': 'talk.politics.misc', '38534': 'comp.graphics', '53530': 'rec.sport.hockey', '59500': 'sci.med', '52583': 'rec.spor
t.hockey', '53626': 'rec.sport.hockey', '59361': 'sci.med', '58118': 'sci.med', '38605': 'comp.graphics', '39615': 'comp.grap
hics', '178895': 'talk.politics.misc', '38588': 'comp.graphics', '38602': 'comp.graphics', '59849': 'sci.space', '60941': 'sc
i.space', '54773': 'rec.sport.hockey', '38314': 'comp.graphics', '61147': 'sci.space', '54089': 'rec.sport.hockey', '61117':
'sci.space', '60969': 'sci.space', '39079': 'comp.graphics', '38911': 'comp.graphics', '178964': 'talk.politics.misc', '5365
8': 'rec.sport.hockey', '60181': 'sci.space', '38815': 'comp.graphics', '179100': 'talk.politics.misc', '61071': 'sci.space',
'59459': 'sci.med', '59041': 'sci.med', '54724': 'rec.sport.hockey', '53581': 'rec.sport.hockey', '53950': 'rec.sport.hocke
y', '60921': 'sci.space', '38986': 'comp.graphics', '178364': 'talk.politics.misc', '52570': 'rec.sport.hockey', '59278': 'sc
```

14. Predicted classes using **mutual information** as feature selection.

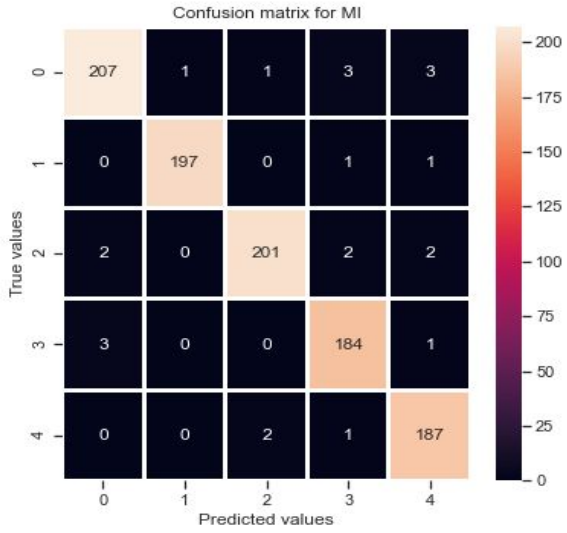
```
Predicted classes for test files using mutual information as feature selection are:
 {'61026': 'sci.space', '38784': 'comp.graphics', '176952': 'talk.politics.misc', '59263': 'sci.med', '58864': 'sci.med', '54
704': 'rec.sport.hockey', '178425': 'talk.politics.misc', '54013': 'rec.sport.hockey', '59341': 'sci.med', '178628': 'talk.po
litics.misc', '53654': 'rec.sport.hockey', '53748': 'rec.sport.hockey', '53722': 'rec.sport.hockey', '38762': 'comp.graphic
s', '59491': 'sci.med', '54510': 'rec.sport.hockey', '39006': 'comp.graphics', '38738': 'comp.graphics', '61507': 'sci.spac
e', '53797': 'rec.sport.hockey', '179115': 'talk.politics.misc', '53655': 'rec.sport.hockey', '53715': 'rec.sport.hockey', '5
8884': 'sci.med', '61215': 'sci.space', '59372': 'sci.med', '54756': 'rec.sport.hockey', '178883': 'talk.politics.misc', '379
38': 'comp.graphics', '52566': 'rec.sport.hockey', '179008': 'talk.politics.misc', '54254': 'rec.sport.hockey', '58139': 'sc
i.med', '61316': 'sci.space', '59204': 'sci.med', '58811': 'sci.med', '54231': 'rec.sport.hockey', '60810': 'sci.space', '601
91': 'sci.space', '53798': 'rec.sport.hockey', '39627': 'sci.space', '59644': 'sci.med', '179004': 'talk.politics.misc', '594
30': 'sci.med', '61086': 'sci.space', '59287': 'sci.med', '61179': 'sci.space', '58066': 'sci.med', '58147': 'sci.med', '1786
01': 'talk.politics.misc', '38534': 'comp.graphics', '53530': 'rec.sport.hockey', '59500': 'sci.med', '52583': 'rec.sport.hoc
key', '53626': 'rec.sport.hockey', '59361': 'sci.med', '58118': 'sci.med', '38605': 'comp.graphics', '39615': 'comp.graphic
s', '178895': 'talk.politics.misc', '38588': 'comp.graphics', '38602': 'comp.graphics', '59849': 'sci.space', '60941': 'sci.s
pace', '54773': 'rec.sport.hockey', '38314': 'comp.graphics', '61147': 'sci.space', '54089': 'rec.sport.hockey', '61117': 'sc
i.space', '60969': 'sci.space', '39079': 'comp.graphics', '38911': 'comp.graphics', '178964': 'talk.politics.misc', '53658':
'rec.sport.hockey', '60181': 'sci.space', '38815': 'comp.graphics', '179100': 'talk.politics.misc', '61071': 'sci.space', '59
459': 'sci.med', '59041': 'sci.med', '54724': 'rec.sport.hockey', '53581': 'rec.sport.hockey', '53950': 'rec.sport.hockey',
'60921': 'sci.space', '38986': 'comp.graphics', '178364': 'talk.politics.misc', '52570': 'rec.sport.hockey', '59278': 'sci.me
```

15. Predicted classes using **TF-IDF** as feature selection.

```
Predicted classes for test files using tf_idf as feature selection are:
 {'61026': 'sci.space', '38784': 'comp.graphics', '176952': 'talk.politics.misc', '59263': 'sci.med', '58864': 'sci.med', '54
704': 'rec.sport.hockey', '178425': 'talk.politics.misc', '54013': 'rec.sport.hockey', '59341': 'sci.med', '178628': 'talk.po
litics.misc', '53654': 'rec.sport.hockey', '53748': 'rec.sport.hockey', '53722': 'rec.sport.hockey', '38762': 'comp.graphic
s', '59491': 'sci.med', '54510': 'rec.sport.hockey', '39006': 'comp.graphics', '38738': 'comp.graphics', '61507': 'sci.spac
e', '53797': 'rec.sport.hockey', '179115': 'talk.politics.misc', '53655': 'rec.sport.hockey', '53715': 'rec.sport.hockey', '5
8884': 'sci.med', '61215': 'sci.space', '59372': 'sci.med', '54756': 'rec.sport.hockey', '178883': 'talk.politics.misc', '379
38': 'comp.graphics', '52566': 'rec.sport.hockey', '179008': 'talk.politics.misc', '54254': 'rec.sport.hockey', '58139': 'sc
i.med', '61316': 'sci.space', '59204': 'sci.med', '58811': 'sci.med', '54231': 'rec.sport.hockey', '60810': 'sci.space', '601
91': 'sci.space', '53798': 'rec.sport.hockey', '39627': 'comp.graphics', '59644': 'sci.med', '179004': 'talk.politics.misc',
'59430': 'sci.med', '61086': 'sci.space', '59287': 'sci.med', '61179': 'sci.space', '58066': 'sci.med', '58147': 'sci.med',
'178601': 'talk.politics.misc', '38534': 'comp.graphics', '53530': 'rec.sport.hockey', '59500': 'sci.med', '52583': 'rec.spor
t.hockey', '53626': 'rec.sport.hockey', '59361': 'sci.med', '58118': 'sci.med', '38605': 'comp.graphics', '39615': 'comp.grap
hics', '178895': 'talk.politics.misc', '38588': 'comp.graphics', '38602': 'comp.graphics', '59849': 'sci.space', '60941': 'sc
i.space', '54773': 'rec.sport.hockey', '38314': 'comp.graphics', '61147': 'sci.space', '54089': 'rec.sport.hockey', '61117':
'sci.space', '60969': 'sci.space', '39079': 'comp.graphics', '38911': 'comp.graphics', '178964': 'talk.politics.misc', '5365
8': 'rec.sport.hockey', '60181': 'sci.space', '38815': 'comp.graphics', '179100': 'talk.politics.misc', '61071': 'sci.space',
'59459': 'sci.med', '59041': 'sci.med', '54724': 'rec.sport.hockey', '53581': 'rec.sport.hockey', '53950': 'rec.sport.hocke
y', '60921': 'sci.space', '38986': 'comp.graphics', '178364': 'talk.politics.misc', '52570': 'rec.sport.hockey', '59278': 'sc
```

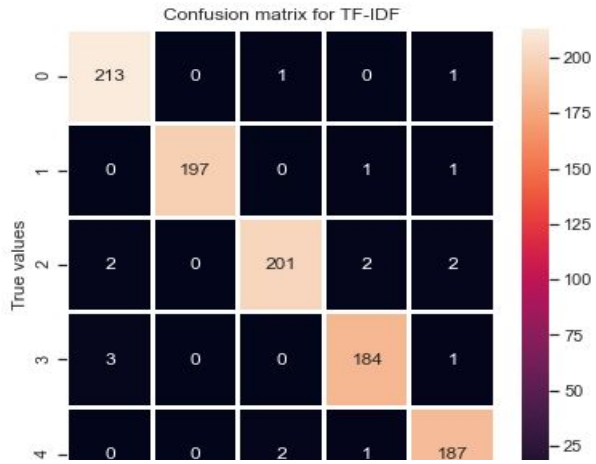
16. Accuracy and confusion matrix for **mutual information**.

```
accuracy after choosing mutual information as feature selection in naive bayes is:  0.9769769769769769
Confusion matrix for MI is:
 [[207   1   1   3   3]
 [  0 197   0   1   1]
 [  2   0 201   2   2]
 [  3   0   0 184   1]
 [  0   0   2   1 187]]
```



17. Accuracy and confusion matrix for **TF-IDF.**

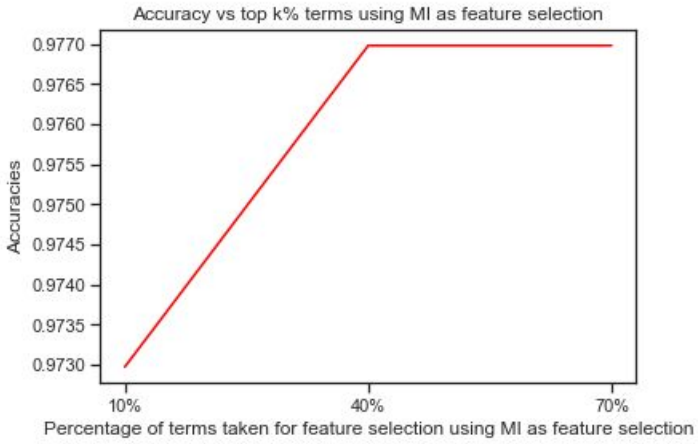
```
accuracy after choosing tf_idf as feature selection in naive bayes is:  0.982982982982983
Confusion matrix for MI is:
 [[213   0   1   0   1]
 [  0 197   0   1   1]
 [  2   0 201   2   2]
 [  3   0   0 184   1]
 [  0   0   2   1 187]]
```



18. Graph between different top k% features and accuracy using **mutual information** as feature selection.

```
k_val=["10%", "40%", "70%" ]
a_mi=[0.972972972972973,0.9769769769769769,0.9769769769769869]
a_tf=[0.975975975975976,0.982982982982983, 0.982982982983983]
plt.plot(k_val, a_mi,color="red")
plt.xlabel("Percentage of terms taken for feature selection using MI as feature selection")
plt.ylabel("Accuracies")
plt.title("Accuracy vs top k% terms using MI as feature selection")
```

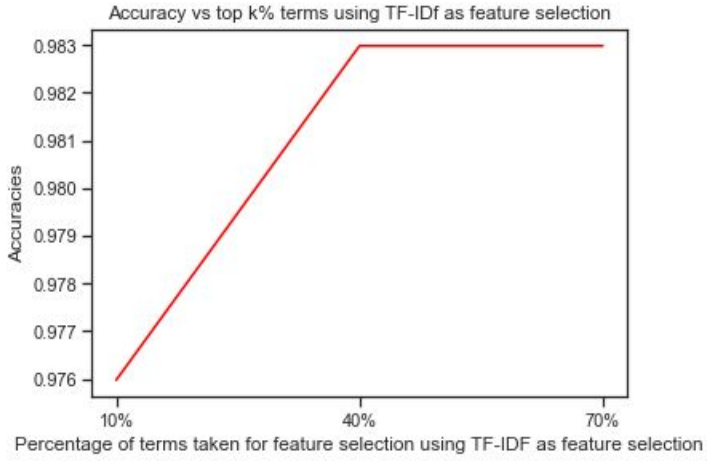
```
Text(0.5, 1.0, 'Accuracy vs top k% terms using MI as feature selection')
```



19. Graph between different top k% features and accuracy using **TF-IDF** as feature selection.

```
plt.plot(k_val, a_tf,color="red")
plt.xlabel("Percentage of terms taken for feature selection using TF-IDF as feature selection")
plt.ylabel("Accuracies")
plt.title("Accuracy vs top k% terms using TF-IDf as feature selection")
```

```
Text(0.5, 1.0, 'Accuracy vs top k% terms using TF-IDf as feature selection')
```



20. Training and testing documents are split in **70:30 ratio.**

```
Enter the percent of documents you want in training set:
70
```

```
print("Number of documents in training set is: ",len(train1))
print("Number of documents in testing set is: ",len(test1))
```

```
Number of documents in training set is:  3497
Number of documents in testing set is:  1498
```

21. Prior probability for each class.

```
print("Prior probability class wise are: \n" ,prior_prob)
```

```
Prior probability class wise are:
 {'comp.graphics': 0.20331712896768658, 'sci.med': 0.19988561624249357, 'talk.politics.misc': 0.19330855018587362, 'rec.sport.h
ockey': 0.19816985987989705, 'sci.space': 0.20531884472404918}
```

22. Actual classes for testing documents.

```
Actual classes for test documents are:
 {'178654': 'talk.politics.misc', '38572': 'comp.graphics', '38248': 'comp.graphics', '59296': 'sci.med', '38884': 'comp.grap
hics', '38326': 'comp.graphics', '52596': 'rec.sport.hockey', '58797': 'sci.med', '178658': 'talk.politics.misc', '52583': 'r
ec.sport.hockey', '58858': 'sci.med', '59055': 'sci.med', '38983': 'comp.graphics', '59323': 'sci.med', '38965': 'comp.graphi
cs', '178758': 'talk.politics.misc', '59011': 'sci.med', '54164': 'rec.sport.hockey', '176949': 'talk.politics.misc', '17908
7': 'talk.politics.misc', '60238': 'sci.space', '176901': 'talk.politics.misc', '39009': 'comp.graphics', '61401': 'sci.spac
e', '61373': 'sci.space', '38926': 'comp.graphics', '178419': 'talk.politics.misc', '53638': 'rec.sport.hockey', '60887': 'sc
i.space', '59490': 'sci.med', '61510': 'sci.space', '60205': 'sci.space', '61166': 'sci.space', '178960': 'talk.politics.mis
c', '38403': 'comp.graphics', '61518': 'sci.space', '179096': 'talk.politics.misc', '54092': 'rec.sport.hockey', '179028': 't
alk.politics.misc', '179100': 'talk.politics.misc', '54134': 'rec.sport.hockey', '61473': 'sci.space', '38279': 'comp.graphic
s', '59643': 'sci.med', '59065': 'sci.med', '38240': 'comp.graphics', '53594': 'rec.sport.hockey', '60863': 'sci.space', '383
44': 'comp.graphics', '177000': 'talk.politics.misc', '60914': 'sci.space', '59326': 'sci.med', '54284': 'rec.sport.hockey',
'58925': 'sci.med', '53920': 'rec.sport.hockey', '60787': 'sci.space', '53848': 'rec.sport.hockey', '178617': 'talk.politics.
misc', '58822': 'sci.med', '38863': 'comp.graphics', '58137': 'sci.med', '59429': 'sci.med', '179077': 'talk.politics.misc',
'178915': 'talk.politics.misc', '61382': 'sci.space', '178322': 'talk.politics.misc', '179015': 'talk.politics.misc', '3872
6': 'comp.graphics', '179023': 'talk.politics.misc', '178712': 'talk.politics.misc', '38540': 'comp.graphics', '178517': 'tal
k.politics.misc', '54517': 'rec.sport.hockey', '176971': 'talk.politics.misc', '178502': 'talk.politics.misc', '178512': 'tal
k.politics.misc', '53986': 'rec.sport.hockey', '53648': 'rec.sport.hockey', '54116': 'rec.sport.hockey', '61047': 'sci.spac
e', '54761': 'rec.sport.hockey', '60814': 'sci.space', '61085': 'sci.space', '60217': 'sci.space', '54129': 'rec.sport.hocke
```

23. Predicted classes using **mutual information** as feature selection.

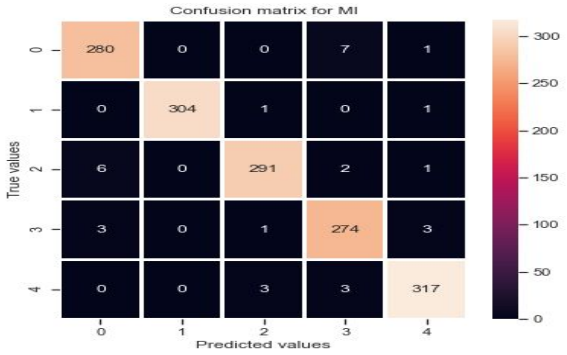
```
Predicted classes for test files using mutual information as feature selection are:
 {'178654': 'talk.politics.misc', '38572': 'comp.graphics', '38248': 'comp.graphics', '59296': 'sci.med', '38884': 'comp.grap
hics', '38326': 'comp.graphics', '52596': 'rec.sport.hockey', '58797': 'sci.med', '178658': 'talk.politics.misc', '52583': 'r
ec.sport.hockey', '58858': 'sci.med', '59055': 'sci.med', '38983': 'comp.graphics', '59323': 'sci.med', '38965': 'comp.graphi
cs', '178758': 'talk.politics.misc', '59011': 'sci.med', '54164': 'rec.sport.hockey', '176949': 'talk.politics.misc', '17908
7': 'talk.politics.misc', '60238': 'sci.space', '176901': 'talk.politics.misc', '39009': 'comp.graphics', '61401': 'sci.spac
e', '61373': 'sci.space', '38926': 'comp.graphics', '178419': 'talk.politics.misc', '53638': 'rec.sport.hockey', '60887': 'sc
i.space', '59490': 'sci.med', '61510': 'sci.space', '60205': 'sci.space', '61166': 'sci.space', '178960': 'talk.politics.mis
c', '38403': 'comp.graphics', '61518': 'sci.space', '179096': 'talk.politics.misc', '54092': 'rec.sport.hockey', '179028': 't
alk.politics.misc', '179100': 'talk.politics.misc', '54134': 'rec.sport.hockey', '61473': 'sci.space', '38279': 'comp.graphic
s', '59643': 'sci.med', '59065': 'sci.med', '38240': 'comp.graphics', '53594': 'rec.sport.hockey', '60863': 'sci.space', '383
44': 'comp.graphics', '177000': 'talk.politics.misc', '60914': 'sci.space', '59326': 'sci.med', '54284': 'rec.sport.hockey',
'58925': 'sci.med', '53920': 'rec.sport.hockey', '60787': 'sci.space', '53848': 'rec.sport.hockey', '178617': 'talk.politics.
misc', '58822': 'sci.med', '38863': 'comp.graphics', '58137': 'sci.med', '59429': 'sci.med', '179077': 'talk.politics.misc',
'178915': 'talk.politics.misc', '61382': 'sci.space', '178322': 'talk.politics.misc', '179015': 'sci.space', '38726': 'comp.g
raphics', '179023': 'talk.politics.misc', '178712': 'talk.politics.misc', '38540': 'comp.graphics', '178517': 'talk.politics.
misc', '54517': 'talk.politics.misc', '176971': 'talk.politics.misc', '178502': 'talk.politics.misc', '178512': 'talk.politic
s.misc', '53986': 'rec.sport.hockey', '53648': 'rec.sport.hockey', '54116': 'rec.sport.hockey', '61047': 'sci.space', '5476
1': 'rec.sport.hockey', '60814': 'sci.space', '61085': 'sci.space', '60217': 'sci.space', '54129': 'rec.sport.hockey', '6112
```

24. Predicted classes using **TF-IDF** as feature selection.

```
Predicted classes for test files using tf_idf as feature selection are:
 {'178654': 'talk.politics.misc', '38572': 'comp.graphics', '38248': 'comp.graphics', '59296': 'sci.med', '38884': 'comp.grap
hics', '38326': 'comp.graphics', '52596': 'rec.sport.hockey', '58797': 'sci.med', '178658': 'talk.politics.misc', '52583': 'r
ec.sport.hockey', '58858': 'sci.med', '59055': 'sci.med', '38983': 'comp.graphics', '59323': 'sci.med', '38965': 'comp.graphi
cs', '178758': 'talk.politics.misc', '59011': 'sci.med', '54164': 'rec.sport.hockey', '176949': 'talk.politics.misc', '17908
7': 'talk.politics.misc', '60238': 'sci.space', '176901': 'talk.politics.misc', '39009': 'comp.graphics', '61401': 'sci.spac
e', '61373': 'sci.space', '38926': 'comp.graphics', '178419': 'talk.politics.misc', '53638': 'rec.sport.hockey', '60887': 'sc
i.space', '59490': 'sci.med', '61510': 'sci.space', '60205': 'sci.space', '61166': 'sci.space', '178960': 'talk.politics.mis
c', '38403': 'comp.graphics', '61518': 'sci.space', '179096': 'talk.politics.misc', '54092': 'rec.sport.hockey', '179028': 't
alk.politics.misc', '179100': 'talk.politics.misc', '54134': 'rec.sport.hockey', '61473': 'sci.space', '38279': 'comp.graphic
s', '59643': 'sci.med', '59065': 'sci.med', '38240': 'comp.graphics', '53594': 'rec.sport.hockey', '60863': 'sci.space', '383
44': 'comp.graphics', '177000': 'talk.politics.misc', '60914': 'sci.space', '59326': 'sci.med', '54284': 'rec.sport.hockey',
'58925': 'sci.med', '53920': 'rec.sport.hockey', '60787': 'sci.space', '53848': 'rec.sport.hockey', '178617': 'talk.politics.
misc', '58822': 'sci.med', '38863': 'comp.graphics', '58137': 'sci.med', '59429': 'sci.med', '179077': 'talk.politics.misc',
'178915': 'talk.politics.misc', '61382': 'sci.space', '178322': 'talk.politics.misc', '179015': 'sci.space', '38726': 'comp.g
raphics', '179023': 'talk.politics.misc', '178712': 'talk.politics.misc', '38540': 'comp.graphics', '178517': 'talk.politics.
misc', '54517': 'talk.politics.misc', '176971': 'talk.politics.misc', '178502': 'talk.politics.misc', '178512': 'talk.politic
s.misc', '53986': 'rec.sport.hockey', '53648': 'rec.sport.hockey', '54116': 'rec.sport.hockey', '61047': 'sci.space', '5476
1': 'rec.sport.hockey', '60814': 'sci.space', '61085': 'sci.space', '60217': 'sci.space', '54129': 'rec.sport.hockey', '6112
```

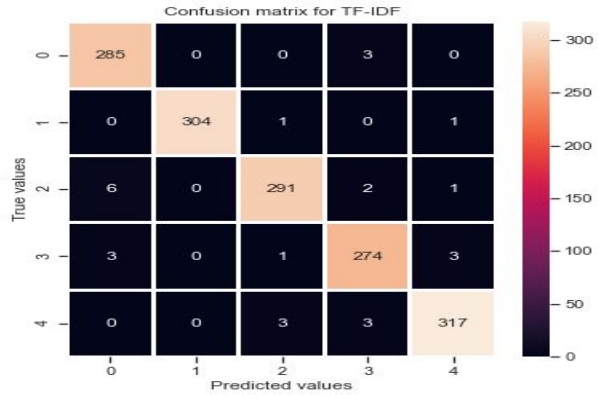
25. Accuracy and confusion matrix for **mutual information**.

```
accuracy after choosing mutual information as feature selection in naive bayes is:  0.9786381842456608
Confusion matrix for MI is:
 [[280   0   0   7   1]
 [  0 304   1   0   1]
 [  6   0 291   2   1]
 [  3   0   1 274   3]
 [  0   0   3   3 317]]
```



26. Accuracy and confusion matrix for **TF-IDF.**

```
accuracy after choosing tf_idf as feature selection in naive bayes is:  0.9819759679572764
Confusion matrix for TF-IDF is:
 [[285   0   0   3   0]
 [  0 304   1   0   1]
 [  6   0 291   2   1]
 [  3   0   1 274   3]
 [  0   0   3   3 317]]
```



27. Training and testing documents are split in **50:50 ratio.**

```
Enter the percent of documents you want in training set:
50
```

```
: print("Number of documents in training set is: ",len(train1))
  print("Number of documents in testing set is: ",len(test1))
```

```
Number of documents in training set is:  2498
Number of documents in testing set is:  2497
```

28. Prior probability for each class.

```
print("Prior probability class wise are: \n" ,prior_prob)
```

```
Prior probability class wise are:
 {'comp.graphics': 0.21377101681345076, 'sci.med': 0.18895116092874 3, 'talk.politics.misc': 0.20096076861489193, 'rec.sport.hoc
key': 0.1969575660528423, 'sci.space': 0.19935948759007205}
```

29. Actual classes for testing documents.

```
Actual classes for test documents are:
 {'176932': 'talk.politics.misc', '60896': 'sci.space', '59206': 'sci.med', '38559': 'comp.graphics', '54069': 'rec.sport.hoc
key', '58956': 'sci.med', '179049': 'talk.politics.misc', '178369': 'talk.politics.misc', '53580': 'rec.sport.hockey', '3906
1': 'comp.graphics', '179043': 'talk.politics.misc', '178719': 'talk.politics.misc', '38234': 'comp.graphics', '178486': 'tal
k.politics.misc', '178438': 'talk.politics.misc', '61297': 'sci.space', '179095': 'talk.politics.misc', '38692': 'comp.graphi
cs', '59266': 'sci.med', '59358': 'sci.med', '52599': 'rec.sport.hockey', '54087': 'rec.sport.hockey', '54738': 'rec.sport.ho
ckey', '59225': 'sci.med', '178555': 'talk.politics.misc', '60837': 'sci.space', '53738': 'rec.sport.hockey', '38703': 'comp.
graphics', '38901': 'comp.graphics', '38374': 'comp.graphics', '52556': 'rec.sport.hockey', '38723': 'comp.graphics', '5387
5': 'rec.sport.hockey', '38463': 'comp.graphics', '52594': 'rec.sport.hockey', '59146': 'sci.med', '38693': 'comp.graphics',
'53686': 'rec.sport.hockey', '53926': 'rec.sport.hockey', '61352': 'sci.space', '179060': 'talk.politics.misc', '54205': 're
c.sport.hockey', '38945': 'comp.graphics', '61363': 'sci.space', '178563': 'talk.politics.misc', '58896': 'sci.med', '53756':
'rec.sport.hockey', '62708': 'sci.space', '53796': 'rec.sport.hockey', '62392': 'sci.space', '38930': 'comp.graphics', '6077
1': 'sci.space', '53531': 'rec.sport.hockey', '60899': 'sci.space', '54192': 'rec.sport.hockey', '38775': 'comp.graphics', '5
8857': 'sci.med', '37928': 'comp.graphics', '60950': 'sci.space', '59055': 'sci.med', '176870': 'talk.politics.misc', '6148
6': 'sci.space', '37957': 'comp.graphics', '53870': 'rec.sport.hockey', '38666': 'comp.graphics', '178323': 'talk.politics.mi
sc', '54019': 'rec.sport.hockey', '59640': 'sci.med', '61051': 'sci.space', '38245': 'comp.graphics', '59111': 'sci.med', '58
796': 'sci.med', '53752': 'rec.sport.hockey', '52636': 'rec.sport.hockey', '61094': 'sci.space', '38699': 'comp.graphics', '1
79074': 'talk.politics.misc', '61479': 'sci.space', '59532': 'sci.med', '179068': 'talk.politics.misc', '62117': 'sci.space',
'61339': 'sci.space', '38683': 'comp.graphics', '59528': 'sci.med', '58780': 'sci.med', '178591': 'talk.politics.misc', '5907
```

30. Predicted classes using **mutual information** as feature selection.

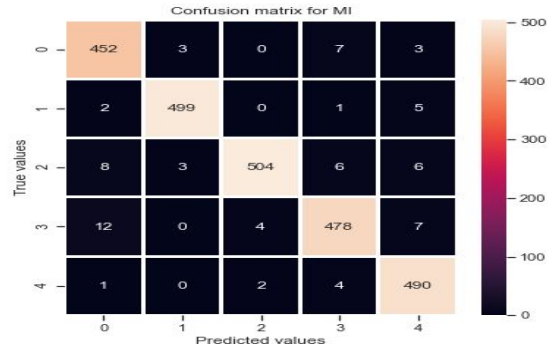
```
Predicted classes for test files using mutual information as feature selection are:
 {'176932': 'talk.politics.misc', '60896': 'sci.space', '59206': 'sci.med', '38559': 'comp.graphics', '54069': 'rec.sport.hoc
key', '58956': 'sci.med', '179049': 'talk.politics.misc', '178369': 'talk.politics.misc', '53580': 'rec.sport.hockey', '3906
1': 'comp.graphics', '179043': 'talk.politics.misc', '178719': 'talk.politics.misc', '38234': 'sci.space', '178486': 'talk.po
litics.misc', '178438': 'talk.politics.misc', '61297': 'sci.space', '179095': 'talk.politics.misc', '38692': 'comp.graphics',
'59266': 'sci.med', '59358': 'sci.med', '52599': 'rec.sport.hockey', '54087': 'rec.sport.hockey', '54738': 'rec.sport.hocke
y', '59225': 'sci.med', '178555': 'talk.politics.misc', '60837': 'sci.space', '53738': 'rec.sport.hockey', '38703': 'comp.gra
phics', '38901': 'comp.graphics', '38374': 'comp.graphics', '52556': 'rec.sport.hockey', '38723': 'comp.graphics', '53875':
'rec.sport.hockey', '38463': 'comp.graphics', '52594': 'rec.sport.hockey', '59146': 'sci.med', '38693': 'comp.graphics', '536
86': 'rec.sport.hockey', '53926': 'rec.sport.hockey', '61352': 'sci.space', '179060': 'talk.politics.misc', '54205': 'rec.spo
rt.hockey', '38945': 'comp.graphics', '61363': 'sci.space', '178563': 'talk.politics.misc', '58896': 'sci.med', '53756': 're
c.sport.hockey', '62708': 'sci.space', '53796': 'rec.sport.hockey', '62392': 'sci.space', '38930': 'sci.space', '60771': 'sc
i.space', '53531': 'rec.sport.hockey', '60899': 'sci.space', '54192': 'rec.sport.hockey', '38775': 'comp.graphics', '58857':
'sci.med', '37928': 'comp.graphics', '60950': 'sci.space', '59055': 'sci.med', '176870': 'talk.politics.misc', '61486': 'sci.
space', '37957': 'comp.graphics', '53870': 'rec.sport.hockey', '38666': 'comp.graphics', '178323': 'talk.politics.misc', '540
19': 'rec.sport.hockey', '59640': 'sci.med', '61051': 'sci.space', '38245': 'comp.graphics', '59111': 'sci.med', '58796': 'sc
i.med', '53752': 'rec.sport.hockey', '52636': 'rec.sport.hockey', '61094': 'sci.space', '38699': 'comp.graphics', '179074':
'talk.politics.misc', '61479': 'sci.space', '59532': 'sci.med', '179068': 'talk.politics.misc', '62117': 'sci.space', '6133
9': 'sci.space', '38683': 'comp.graphics', '59528': 'sci.med', '58780': 'sci.med', '178591': 'talk.politics.misc', '59071':
```

31. Predicted classes using **TF-IDF** as feature selection.

```
Predicted classes for test files using tf_idf as feature selection are:
 {'176932': 'talk.politics.misc', '60896': 'sci.space', '59206': 'sci.med', '38559': 'comp.graphics', '54069': 'rec.sport.hoc
key', '58956': 'sci.med', '179049': 'talk.politics.misc', '178369': 'talk.politics.misc', '53580': 'rec.sport.hockey', '3906
1': 'comp.graphics', '179043': 'talk.politics.misc', '178719': 'talk.politics.misc', '38234': 'comp.graphics', '178486': 'tal
k.politics.misc', '178438': 'talk.politics.misc', '61297': 'sci.space', '179095': 'talk.politics.misc', '38692': 'comp.graphi
cs', '59266': 'sci.med', '59358': 'sci.med', '52599': 'rec.sport.hockey', '54087': 'rec.sport.hockey', '54738': 'rec.sport.ho
ckey', '59225': 'sci.med', '178555': 'talk.politics.misc', '60837': 'sci.space', '53738': 'rec.sport.hockey', '38703': 'comp.
graphics', '38901': 'comp.graphics', '38374': 'comp.graphics', '52556': 'rec.sport.hockey', '38723': 'comp.graphics', '5387
5': 'rec.sport.hockey', '38463': 'comp.graphics', '52594': 'rec.sport.hockey', '59146': 'sci.med', '38693': 'comp.graphics',
'53686': 'rec.sport.hockey', '53926': 'rec.sport.hockey', '61352': 'sci.space', '179060': 'talk.politics.misc', '54205': 're
c.sport.hockey', '38945': 'comp.graphics', '61363': 'sci.space', '178563': 'talk.politics.misc', '58896': 'sci.med', '53756':
'rec.sport.hockey', '62708': 'sci.space', '53796': 'rec.sport.hockey', '62392': 'sci.space', '38930': 'comp.graphics', '6077
1': 'sci.space', '53531': 'rec.sport.hockey', '60899': 'sci.space', '54192': 'rec.sport.hockey', '38775': 'comp.graphics', '5
8857': 'sci.med', '37928': 'comp.graphics', '60950': 'sci.space', '59055': 'sci.med', '176870': 'talk.politics.misc', '6148
6': 'sci.space', '37957': 'comp.graphics', '53870': 'rec.sport.hockey', '38666': 'comp.graphics', '178323': 'talk.politics.mi
sc', '54019': 'rec.sport.hockey', '59640': 'sci.med', '61051': 'sci.space', '38245': 'comp.graphics', '59111': 'sci.med', '58
796': 'sci.med', '53752': 'rec.sport.hockey', '52636': 'rec.sport.hockey', '61094': 'sci.space', '38699': 'comp.graphics', '1
79074': 'talk.politics.misc', '61479': 'sci.space', '59532': 'sci.med', '179068': 'talk.politics.misc', '62117': 'sci.space',
'61339': 'sci.space', '38683': 'comp.graphics', '59528': 'sci.med', '58780': 'sci.med', '178591': 'talk.politics.misc', '5907
```

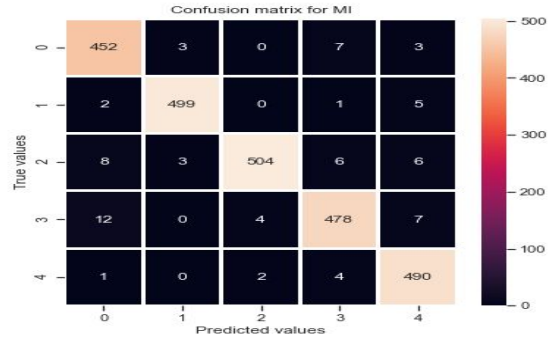
32. Accuracy and confusion matrix for **mutual information**.

```
accuracy after choosing mutual information as feature selection in naive bayes is:  0.9703644373247897
Confusion matrix for MI is:
 [[452   3   0   7   3]
 [  2 499   0   1   5]
 [  8   3 504   6   6]
 [ 12   0   4 478   7]
 [  1   0   2   4 490]]
```



33. Accuracy and confusion matrix for **TF-IDF.**

```
accuracy after choosing mutual information as feature selection in naive bayes is:  0.9703644373247897
Confusion matrix for MI is:
 [[452   3   0   7   3]
 [  2 499   0   1   5]
 [  8   3 504   6   6]
 [ 12   0   4 478   7]
 [  1   0   2   4 490]]
```



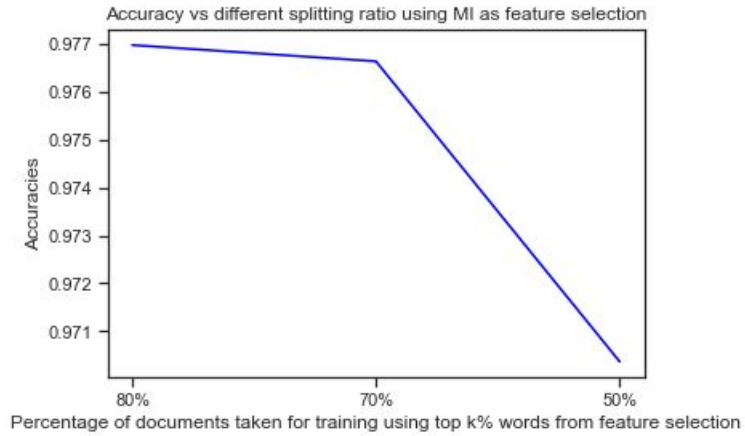
34. Accuracy vs different split ratio graph using **mutual information** as feature selection.

```
a=[0.9769769769769769,0.9766381842456608,0.9703644373247897]

p=["80%", "70%", "50%"]

plt.plot(p, a,color="blue")
plt.xlabel("Percentage of documents taken for training using top k% words from feature selection")
plt.ylabel("Accuracies")
plt.title("Accuracy vs different splitting ratio using MI as feature selection")
```

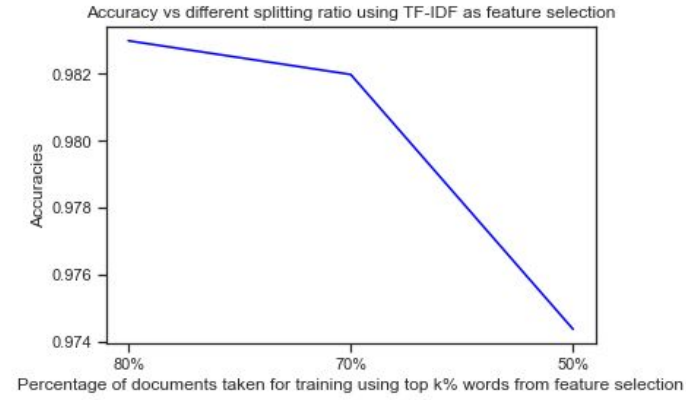
```
Text(0.5, 1.0, 'Accuracy vs different splitting ratio using MI as feature selection')
```



35. Accuracy vs different split ratio graph using **TF-IDF** as feature selection.

```
a1=[0.982982982982983,0.9819759679572764,0.97436924309171]
p=["80%", "70%", "50%"]
plt.plot(p, a1,color="blue")
plt.xlabel("Percentage of documents taken for training using top k% words from feature selection")
plt.ylabel("Accuracies")
plt.title("Accuracy vs different splitting ratio using TF-IDF as feature selection")
```

```
Text(0.5, 1.0, 'Accuracy vs different splitting ratio using TF-IDF as feature selection')
```



Question 2 KNN

1. Training and testing documents are split in **80:20 ratio.**

```
Enter the percent of documents you want in training set:
80
```

```
print("Number of documents in training set is: ",len(train1))
print("Number of documents in testing set is: ",len(test1))
```

```
Number of documents in training set is:  3996
Number of documents in testing set is:  999
```

2. Top 40 % tokens for class **comp.graphics** using mutual information as feature selection.

```
print("Top",k,"% terms using MI for feature selection for comp.graphics are: \n ",mi_final['comp.graphics'])
```

```
Top 10 % terms using MI for feature selection for comp.graphics are:
  {'line': 1, 'subject': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1, 'date': 1, 'organization': 1, 'apr': 1, 'gmt': 1, 'co
mpgraphics': 1, 'one thousand, nine hundred and ninety-three': 1, 'reference': 1, 'sender': 1, 'nntppostinghost': 1, 'univers
ity': 1, 'one': 1, 'ninety-three': 1, 'writes': 1, 'graphic': 1, 'would': 1, 'know': 1, 'xref': 1, 'cantaloupesrvcscmuedu':
1, 'thanks': 1, 'article': 1, 'like': 1, 'anyone': 1, 'file': 1, 'two': 1, 'please': 1, 'image': 1, 'email': 1, 'computer':
1, 'need': 1, 'help': 1, 'program': 1, 'may': 1, 'system': 1, 'get': 1, 'dont': 1, 'im': 1, 'use': 1, 'also': 1, 'news': 1,
'could': 1, 'replyto': 1, 'version': 1, 'distribution': 1, 'looking': 1, 'time': 1, 'keywords': 1, 'find': 1, 'think': 1, 'pr
oblem': 1, '3d': 1, 'software': 1, 'format': 1, 'twenty': 1, 'good': 1, 'eleven': 1, 'sixteen': 1, 'world': 1, 'work': 1, 'in
formation': 1, 'using': 1, 'way': 1, 'available': 1, 'three': 1, 'new': 1, 'ive': 1, 'much': 1, 'many': 1, 'well': 1, 'scienc
e': 1, 'want': 1, 'code': 1, 'c': 1, 'make': 1, 'fifteen': 1, 'color': 1, 'usenet': 1, 'bit': 1, 'six': 1, 'point': 1, 'thin
g': 1, 'see': 1, 'number': 1, 'something': 1, 'four': 1, 'group': 1, 'window': 1, 'hi': 1, 'used': 1, 'twenty-one': 1, 'rea
d': 1, 'got': 1, 'question': 1, 'first': 1, 'even': 1, 'go': 1, 'ten': 1, 'five': 1, 'eight': 1, 'say': 1, 'people': 1, 'we
d': 1, 'lot': 1, 'ftp': 1, 'pc': 1, 'advance': 1, 'take': 1, 'algorithm': 1, 'twenty-four': 1, 'etc': 1, 'since': 1, 'look':
1, 'thirty': 1, 'sun': 1, 'video': 1, 'package': 1, 'post': 1, 'tue': 1, 'nineteen': 1, 'better': 1, 'support': 1, 'seventee
n': 1, 'twenty-two': 1, 'give': 1, 'xnewsreader': 1, 'display': 1, 'gif': 1, 'card': 1, 'u': 1, 'twenty-three': 1, 'address':
1, 'run': 1, 'year': 1, 'source': 1, 'another': 1, 'convert': 1, 'fax': 1, 'mon': 1, 'info': 1, 'site': 1, 'cant': 1, 'writ
e': 1, 'tin': 1, 'book': 1, 'fourteen': 1, 'library': 1, 'wrote': 1, 'sure': 1, 'else': 1, 'interested': 1, 'come': 1, 'someo
ne': 1, 'really': 1, 'center': 1, 'thu': 1, 'best': 1, 'data': 1, 'user': 1, 'able': 1, 'twenty-eight': 1, 'right': 1, 'stil
l': 1, 'try': 1, 'x': 1, 'appreciated': 1, 'usa': 1, 'different': 1, 'twenty-six': 1, 'twelve': 1, 'fri': 1, 'vga': 1, 'anima
```

3. Top 40 % tokens for class **sci.med** using mutual information as feature selection.

```
print("Top",k,"5 terms using MI for feature selection for sci.med are: \n ",mi_final['sci.med'])
```

```
Top 10 5 terms using MI for feature selection for sci.med are:
  {'one': 1, 'subject': 1, 'organization': 1, 'date': 1, 'line': 1, 'newsgroups': 1, 'path': 1, 'apr': 1, 'messageid': 1, 'gm
t': 1, 'reference': 1, 'article': 1, 'one thousand, nine hundred and ninety-three': 1, 'writes': 1, 'would': 1, 'know': 1, 's
ender': 1, 'also': 1, 'dont': 1, 'get': 1, 'university': 1, 'like': 1, 'ninety-three': 1, 'people': 1, 'medical': 1, 'use':
1, 'time': 1, 'two': 1, 'may': 1, 'im': 1, 'food': 1, 'problem': 1, 'year': 1, 'good': 1, 'nntppostinghost': 1, 'many': 1, 't
hink': 1, 'new': 1, 'system': 1, 'science': 1, 'work': 1, 'cancer': 1, 'study': 1, 'gordon': 1, 'effect': 1, 'much': 1, 'hel
p': 1, 'bank': 1, 'day': 1, 'make': 1, 'well': 1, 'information': 1, 'cause': 1, 'could': 1, 'research': 1, 'drug': 1, 'thin
g': 1, 'way': 1, 'even': 1, 'medicine': 1, 'case': 1, 'three': 1, 'anyone': 1, 'say': 1, 'ive': 1, 'take': 1, 'u': 1, 'see':
1, 'news': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'six': 1, 'question': 1, 'used': 1, 'want': 1, 'since': 1, 'replyto': 1,
'twenty': 1, 'pain': 1, 'back': 1, 'without': 1, 'number': 1, 'state': 1, 'something': 1, 'really': 1, 'need': 1, 'computer':
1, 'first': 1, 'distribution': 1, 'center': 1, 'program': 1, 'lot': 1, 'might': 1, 'five': 1, 'email': 1, 'body': 1, 'test':
1, 'ten': 1, 'said': 1, 'go': 1, 'world': 1, 'week': 1, 'seems': 1, 'person': 1, 'long': 1, 'find': 1, 'still': 1, 'four': 1,
'evidence': 1, 'pittsburgh': 1, 'level': 1, 'david': 1, 'try': 1, 'result': 1, 'group': 1, 'fact': 1, 'steve': 1, 'anything':
1, 'believe': 1, 'enough': 1, 'sure': 1, 'course': 1, 'read': 1, 'human': 1, 'blood': 1, 'part': 1, 'thats': 1, 'school': 1,
'thanks': 1, 'etc': 1, 'post': 1, 'never': 1, 'every': 1, 'point': 1, 'please': 1, 'available': 1, 'someone': 1, 'give': 1,
'come': 1, 'using': 1, 'high': 1, 'dr': 1, 'better': 1, 'another': 1, 'right': 1, 'going': 1, 'doesnt': 1, 'probably': 1, 'ki
nd': 1, 'risk': 1, 'usenet': 1, 'cant': 1, 'doctor': 1, 'month': 1, 'c': 1, 'general': 1, 'found': 1, 'different': 1, 'usa':
1, 'type': 1, 'info': 1, 'taking': 1, 'service': 1, 'public': 1, 'done': 1, 'tell': 1, 'weight': 1, 'thirty': 1, 'little': 1,
'fifteen': 1, 'twenty-one': 1, 'mean': 1, 'le': 1, 'idea': 1, 'however': 1, 'opinion': 1, 'product': 1, 'possible': 1, 'chang
```

4. Top 40 % tokens for class **talk.politics.misc** using mutual information as feature selection.

```
print("Top",k,"% terms using MI for feature selection for talk.politics.misc are: \n ",mi_final['talk.politics.misc'])
```

```
Top 10 % terms using MI for feature selection for talk.politics.misc are:
  {'apr': 1, 'line': 1, 'subject': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1, 'date': 1, 'organization': 1, 'reference':
1, 'gmt': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'writes': 1, 'article': 1, 'one thousand, nine hundred and ninety-three':
1, 'one': 1, 'people': 1, 'would': 1, 'sender': 1, 'nntppostinghost': 1, 'dont': 1, 'ninety-three': 1, 'like': 1, 'universit
y': 1, 'u': 1, 'think': 1, 'time': 1, 'know': 1, 'say': 1, 'government': 1, 'two': 1, 'state': 1, 'make': 1, 'get': 1, 'new':
1, 'news': 1, 'even': 1, 'right': 1, 'way': 1, 'much': 1, 'many': 1, 'im': 1, 'also': 1, 'well': 1, 'see': 1, 'go': 1, 'twent
y': 1, 'want': 1, 'opinion': 1, 'could': 1, 'talkpoliticsmisc': 1, 'law': 1, 'usa': 1, 'fact': 1, 'thing': 1, 'good': 1, 'sys
tem': 1, 'case': 1, 'believe': 1, 'year': 1, 'may': 1, 'really': 1, 'child': 1, 'first': 1, 'world': 1, 'said': 1, 'take': 1,
'part': 1, 'distribution': 1, 'point': 1, 'mean': 1, 'fifteen': 1, 'still': 1, 'american': 1, 'going': 1, 'use': 1, 'cant':
1, 'sixteen': 1, 'since': 1, 'sure': 1, 'let': 1, 'gay': 1, 'clinton': 1, 'need': 1, 'ten': 1, 'day': 1, 'back': 1, 'made':
1, 'replyto': 1, 'number': 1, 'country': 1, 'last': 1, 'might': 1, 'six': 1, 'question': 1, 'look': 1, 'problem': 1, 'life':
1, 'five': 1, 'anything': 1, 'come': 1, 'twenty-one': 1, 'someone': 1, 'never': 1, 'something': 1, 'three': 1, 'anyone': 1,
'another': 1, 'reason': 1, 'president': 1, 'c': 1, 'place': 1, 'support': 1, 'got': 1, 'money': 1, 'public': 1, 'thats': 1,
'give': 1, 'usenet': 1, 'care': 1, 'show': 1, 'clayton': 1, 'four': 1, 'doesnt': 1, 'person': 1, 'group': 1, 'cramer': 1, 'se
rvice': 1, 'without': 1, 'used': 1, 'tue': 1, 'find': 1, 'enough': 1, 'health': 1, 'tell': 1, 'every': 1, 'lot': 1, 'work':
1, 'study': 1, 'put': 1, 'actually': 1, 'didnt': 1, 'must': 1, 'try': 1, 'free': 1, 'thought': 1, 'next': 1, 'issue': 1, 'fa
r': 1, 'kind': 1, 'course': 1, 'ive': 1, 'homosexual': 1, 'isnt': 1, 'nothing': 1, 'crameroptilinkcom': 1, 'national': 1, 'tw
enty-three': 1, 'little': 1, 'least': 1, 'bill': 1, 'pay': 1, 'please': 1, 'david': 1, 'better': 1, 'keep': 1, 'idea': 1, 'me
n': 1, 'however': 1, 'call': 1, 'tax': 1, 'force': 1, 'yes': 1, 'using': 1, 'fri': 1, 'start': 1, 'different': 1, 'already':
```

5. Top 40 % tokens for class **rec.sport.hockey** using mutual information as feature selection.

```
print("Top",k,"% terms using MI for feature selection for rec.sport.hockey are: \n ",mi_final['rec.sport.hockey'])
```

```
Top 10 % terms using MI for feature selection for rec.sport.hockey are:
  {'recsporthockey': 1, 'apr': 1, 'subject': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1, 'date': 1, 'line': 1, 'organizati
on': 1, 'gmt': 1, 'one thousand, nine hundred and ninety-three': 1, 'reference': 1, 'game': 1, 'university': 1, 'sender': 1,
'writes': 1, 'team': 1, 'one': 1, 'nntppostinghost': 1, 'article': 1, 'hockey': 1, 'ninety-three': 1, 'would': 1, 'go': 1, 't
wo': 1, 'playoff': 1, 'player': 1, 'year': 1, 'get': 1, 'nhl': 1, 'like': 1, 'think': 1, 'time': 1, 'know': 1, 'fan': 1, 'pla
y': 1, 'last': 1, 'dont': 1, 'good': 1, 'season': 1, 'win': 1, 'six': 1, 'three': 1, 'see': 1, 'first': 1, 'im': 1, 'well':
1, 'news': 1, 'even': 1, 'five': 1, 'twenty': 1, 'twenty-one': 1, 'goal': 1, 'cup': 1, 'four': 1, 'going': 1, 'twenty-three':
1, 'fifteen': 1, 'sixteen': 1, 'new': 1, 'also': 1, 'way': 1, 'back': 1, 'really': 1, 'could': 1, 'next': 1, 'many': 1, 'sa
y': 1, 'let': 1, 'make': 1, 'league': 1, 'wing': 1, 'pittsburgh': 1, 'night': 1, 'right': 1, 'point': 1, 'toronto': 1, 'bes
t': 1, 'played': 1, 'take': 1, 'got': 1, 'much': 1, 'since': 1, 'didnt': 1, 'eleven': 1, 'people': 1, 'bruin': 1, 'ten': 1,
'better': 1, 'canada': 1, 'fri': 1, 'leaf': 1, 'seven': 1, 'distribution': 1, 'world': 1, 'cant': 1, 'anyone': 1, 'come': 1,
'twenty-four': 1, 'detroit': 1, 'twenty-six': 1, 'great': 1, 'state': 1, 'twenty-two': 1, 'give': 1, 'usenet': 1, 'ranger':
1, 'stanley': 1, 'thing': 1, 'replyto': 1, 'division': 1, 'want': 1, 'still': 1, 'playing': 1, 'mike': 1, 'devil': 1, 'may':
1, 'eighteen': 1, 'final': 1, 'look': 1, 'u': 1, 'boston': 1, 'sure': 1, 'john': 1, 'second': 1, 'tue': 1, 'eight': 1, 'mon':
1, 'series': 1, 'another': 1, 'lot': 1, 'made': 1, 'little': 1, 'penguin': 1, 'never': 1, 'need': 1, 'guy': 1, 'put': 1, 'goa
lie': 1, 'ice': 1, 'coach': 1, 'system': 1, 'nineteen': 1, 'pen': 1, 'show': 1, 'said': 1, 'he': 1, 'might': 1, 'espn': 1, 'p
ost': 1, 'bad': 1, 'mean': 1, 'getting': 1, 'shot': 1, 'id': 1, 'end': 1, 'blue': 1, 'seventeen': 1, 'twelve': 1, 'usa': 1,
'contact': 1, 'least': 1, 'something': 1, 'science': 1, 'thats': 1, 'day': 1, 'question': 1, 'buffalo': 1, 'red': 1, 'score':
1, 'thirty': 1, 'islander': 1, 'doesnt': 1, 'around': 1, 'patrick': 1, 'probably': 1, 'round': 1, 'remember': 1, 'ever': 1,
```

6. Top 40 % tokens for class **sci.space** using mutual information as feature selection.

```
print("Top",k,"% terms using MI for feature selection for sci.space are: \n ",mi_final['sci.space'])
```

```
Top 10 % terms using MI for feature selection for sci.space are:
  {'subject': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1, 'date': 1, 'line': 1, 'organization': 1, 'apr': 1, 'gmt': 1, 'sc
ispace': 1, 'one thousand, nine hundred and ninety-three': 1, 'reference': 1, 'writes': 1, 'space': 1, 'article': 1, 'one':
1, 'sender': 1, 'nntppostinghost': 1, 'would': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'university': 1, 'like': 1, 'ninety-
three': 1, 'may': 1, 'system': 1, 'get': 1, 'know': 1, 'u': 1, 'two': 1, 'also': 1, 'could': 1, 'dont': 1, 'time': 1, 'thin
k': 1, 'distribution': 1, 'year': 1, 'much': 1, 'make': 1, 'thing': 1, 'news': 1, 'orbit': 1, 'see': 1, 'earth': 1, 'new': 1,
'well': 1, 'way': 1, 'people': 1, 'world': 1, 'first': 1, 'nasa': 1, 'science': 1, 'im': 1, 'day': 1, 'go': 1, 'six': 1, 'eve
n': 1, 'use': 1, 'three': 1, 'need': 1, 'good': 1, 'moon': 1, 'shuttle': 1, 'mission': 1, 'program': 1, 'idea': 1, 'might':
1, 'something': 1, 'question': 1, 'launch': 1, 'many': 1, 'work': 1, 'anyone': 1, 'twenty': 1, 'pat': 1, 'high': 1, 'since':
1, 'find': 1, 'long': 1, 'give': 1, 'twenty-three': 1, 'back': 1, 'usenet': 1, 'say': 1, 'technology': 1, 'henry': 1, 'cost':
1, 'better': 1, 'want': 1, 'going': 1, 'tue': 1, 'sixteen': 1, 'around': 1, 'usa': 1, 'problem': 1, 'look': 1, 'part': 1, 'pr
oject': 1, 'enough': 1, 'large': 1, 'still': 1, 'sun': 1, 'right': 1, 'put': 1, 'five': 1, 'used': 1, 'satellite': 1, 'usin
g': 1, 'sky': 1, 'thats': 1, 'spencer': 1, 'lot': 1, 'fri': 1, 'point': 1, 'flight': 1, 'twenty-two': 1, 'data': 1, 'eleven':
1, 'thirty': 1, 'information': 1, 'computer': 1, 'research': 1, 'probably': 1, 'fifteen': 1, 'keywords': 1, 'really': 1, 'fou
r': 1, 'come': 1, 'henryzootorontoedu': 1, 'take': 1, 'state': 1, 'power': 1, 'cant': 1, 'someone': 1, 'please': 1, 'access':
1, 'mean': 1, 'communication': 1, 'solar': 1, 'big': 1, 'real': 1, 'ten': 1, 'center': 1, 'twenty-one': 1, 'twenty-seven': 1,
'sci': 1, 'spacecraft': 1, 'prbaccessdigexcom': 1, 'last': 1, 'made': 1, 'replyto': 1, 'different': 1, 'said': 1, 'eight': 1,
'course': 1, 'sure': 1, 'little': 1, 'id': 1, 'maybe': 1, 'message': 1, 'available': 1, 'international': 1, 'another': 1, 'le
ast': 1, 'twenty-nine': 1, 'email': 1, 'try': 1, 'read': 1, 'number': 1, 'believe': 1, 'however': 1, 'group': 1, 'money': 1,
```

7. Top 40 % tokens for class **comp.graphics** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for comp.graphics are: \n ",tf_idf_final['comp.graphics'])
```

```
usand and twenty-eight': 1, 'one hundred and thirteen thousand, one hundred and nineteen': 1, 'one hundred and sixty-five':
1, 'one hundred and seventy-three': 1, 'one hundred and seventy-four': 1, 'one hundred and forty-two': 1, 'one hundred and fi
fty-nine': 1, 'one hundred and eighty-three thousand, six hundred and eleven': 1, 'one billion, nine hundred and twenty-three
million, five hundred and twenty-four thousand, six hundred and seventeen': 1, 'occupation': 1, 'o4fzrapdbh107heosvcrwimseybc
ca': 1, 'nz': 1, 'nuke': 1, 'nov': 1, 'notification': 1, 'northwestern': 1, 'nonetheless': 1, 'nominee': 1, 'nntppostinguse
r': 1, 'nntpaludrauscedu': 1, 'nintendo': 1, 'nine hundred and sixty-nine': 1, 'nine hundred and fifteen': 1, 'nine hundred a
nd eight': 1, 'nicer': 1, 'nfotisntuagr': 1, 'nexus': 1, 'newszolaesdsgicom': 1, 'newsworthy': 1, 'newswatsonibmcom': 1, 'new
ssfuca': 1, 'newsmastercnatekcom': 1, 'newslinusmitreorg': 1, 'newskakwaucsualbertaca': 1, 'newsjupitersuncsdunbca': 1, 'news
galileoccrochesteredu': 1, 'newsdartvaxdartmouthedu': 1, 'newsccuumanitobaca': 1, 'newsaiojscnasagov': 1, 'neighbour': 1, 'ne
gotiating': 1, 'neatly': 1, 'nationality': 1, 'nah': 1, 'murderer': 1, 'multiplier': 1, 'msg': 1, 'msdosgraphics': 1, 'mp':
1, 'motherinlaw': 1, 'mobility': 1, 'moan': 1, 'mj': 1, 'mitre': 1, 'mistakenly': 1, 'minneapolis': 1, 'minded': 1, 'micron':
1, 'micheal': 1, 'meteorology': 1, 'meteorological': 1, 'metafile': 1, 'memo': 1, 'mehl': 1, 'maureen': 1, 'matrox': 1, 'ma
t': 1, 'mario': 1, 'maria': 1, 'manifestation': 1, 'malicious': 1, 'lynch': 1, 'ltqp28innpa7pageboycsutexasedu': 1, 'lovely':
1, 'lous': 1, 'locked': 1, 'liu': 1, 'litigation': 1, 'lit': 1, 'lionessufccufledu': 1, 'lied': 1, 'librarian': 1, 'letterma
n': 1, 'lease': 1, 'lancaster': 1, 'kt': 1, 'kicker': 1, 'kenbluenjitedu': 1, 'kenaidenali': 1, 'kay': 1, 'joined': 1, 'joach
im': 1, 'jl': 1, 'jd': 1, 'jail': 1, 'irzr17inftudresdende': 1, 'irregular': 1, 'ironically': 1, 'ip': 1, 'inventing': 1, 'in
vent': 1, 'internally': 1, 'intern': 1, 'interact': 1, 'institutional': 1, 'installing': 1, 'insanity': 1, 'indicator': 1, 'i
mplicit': 1, 'illustration': 1, 'ignores': 1, 'ige': 1, 'idle': 1, 'identifying': 1, 'hurry': 1, 'hst': 1, 'horizon': 1, 'hoo
ked': 1, 'honey': 1, 'holiday': 1, 'hillary': 1}
```

8. Top 40 % tokens for class **sci.med** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for sci.med are: \n ",tf_idf_final['sci.med'])
```

```
Top 10 % terms using tf_idf for feature selection for sci.med are:
  {'one': 1, 'subject': 1, 'organization': 1, 'date': 1, 'line': 1, 'newsgroups': 1, 'path': 1, 'apr': 1, 'messageid': 1, 'gm
t': 1, 'reference': 1, 'article': 1, 'one thousand, nine hundred and ninety-three': 1, 'writes': 1, 'would': 1, 'know': 1, 's
ender': 1, 'also': 1, 'dont': 1, 'get': 1, 'university': 1, 'like': 1, 'ninety-three': 1, 'people': 1, 'medical': 1, 'use':
1, 'time': 1, 'two': 1, 'may': 1, 'im': 1, 'food': 1, 'problem': 1, 'year': 1, 'good': 1, 'nntppostinghost': 1, 'many': 1, 't
hink': 1, 'new': 1, 'system': 1, 'science': 1, 'work': 1, 'cancer': 1, 'study': 1, 'gordon': 1, 'effect': 1, 'much': 1, 'hel
p': 1, 'bank': 1, 'day': 1, 'make': 1, 'well': 1, 'information': 1, 'cause': 1, 'could': 1, 'research': 1, 'drug': 1, 'thin
g': 1, 'way': 1, 'even': 1, 'medicine': 1, 'case': 1, 'three': 1, 'anyone': 1, 'say': 1, 'ive': 1, 'take': 1, 'u': 1, 'see':
1, 'news': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'six': 1, 'question': 1, 'used': 1, 'want': 1, 'since': 1, 'replyto': 1,
'twenty': 1, 'pain': 1, 'back': 1, 'without': 1, 'number': 1, 'state': 1, 'something': 1, 'really': 1, 'need': 1, 'computer':
1, 'first': 1, 'distribution': 1, 'center': 1, 'program': 1, 'lot': 1, 'might': 1, 'five': 1, 'email': 1, 'body': 1, 'test':
1, 'ten': 1, 'said': 1, 'go': 1, 'world': 1, 'week': 1, 'seems': 1, 'person': 1, 'long': 1, 'find': 1, 'still': 1, 'four': 1,
'evidence': 1, 'pittsburgh': 1, 'level': 1, 'david': 1, 'try': 1, 'result': 1, 'group': 1, 'fact': 1, 'steve': 1, 'anything':
1, 'believe': 1, 'enough': 1, 'sure': 1, 'course': 1, 'read': 1, 'human': 1, 'blood': 1, 'part': 1, 'thats': 1, 'school': 1,
'thanks': 1, 'etc': 1, 'post': 1, 'never': 1, 'every': 1, 'point': 1, 'please': 1, 'available': 1, 'someone': 1, 'give': 1,
'come': 1, 'using': 1, 'high': 1, 'dr': 1, 'better': 1, 'another': 1, 'right': 1, 'going': 1, 'doesnt': 1, 'probably': 1, 'ki
nd': 1, 'risk': 1, 'usenet': 1, 'cant': 1, 'doctor': 1, 'month': 1, 'c': 1, 'general': 1, 'found': 1, 'different': 1, 'usa':
1, 'type': 1, 'info': 1, 'taking': 1, 'service': 1, 'public': 1, 'done': 1, 'tell': 1, 'weight': 1, 'thirty': 1, 'little': 1,
'fifteen': 1, 'twenty-one': 1, 'mean': 1, 'le': 1, 'idea': 1, 'however': 1, 'opinion': 1, 'product': 1, 'possible': 1, 'chang
```

9. Top 40 % tokens for class **talk.politics.misc** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for talk.politics.misc are: \n ",tf_idf_final['talk.politics.misc'])
```

```
Top 10 % terms using tf_idf for feature selection for talk.politics.misc are:
  {'would': 1, 'people': 1, 'one': 1, 'writes': 1, 'article': 1, 'q': 1, 'line': 1, 'subject': 1, 'apr': 1, 'dont': 1, 'organ
ization': 1, 'date': 1, 'think': 1, 'newsgroups': 1, 'path': 1, 'messageid': 1, 'president': 1, 'reference': 1, 'gmt': 1, 'go
vernment': 1, 'know': 1, 'mr': 1, 'one thousand, nine hundred and ninety-three': 1, 'right': 1, 'xref': 1, 'cantaloupesrvcscm
uedu': 1, 'u': 1, 'state': 1, 'like': 1, 'make': 1, 'well': 1, 'time': 1, 'get': 1, 'say': 1, 'new': 1, 'going': 1, 'said':
1, 'two': 1, 'want': 1, 'way': 1, 'sender': 1, 'university': 1, 'law': 1, 'go': 1, 'even': 1, 'american': 1, 'child': 1, 'thi
ng': 1, 'ninety-three': 1, 'year': 1, 'nntppostinghost': 1, 'im': 1, 'also': 1, 'could': 1, 'much': 1, 'believe': 1, 'questio
n': 1, 'good': 1, 'many': 1, 'system': 1, 'made': 1, 'mean': 1, 'see': 1, 'first': 1, 'may': 1, 'case': 1, 'take': 1, 'news':
1, 'job': 1, 'drug': 1, 'number': 1, 'thats': 1, 'day': 1, 'something': 1, 'fact': 1, 'country': 1, 'world': 1, 'work': 1, 'n
eed': 1, 'point': 1, 'part': 1, 'care': 1, 'private': 1, 'come': 1, 'back': 1, 'really': 1, 'opinion': 1, 'money': 1, 'men':
1, 'group': 1, 'program': 1, 'last': 1, 'sure': 1, 'let': 1, 'general': 1, 'support': 1, 'house': 1, 'life': 1, 'use': 1, 'si
nce': 1, 'public': 1, 'war': 1, 'still': 1, 'pay': 1, 'service': 1, 'problem': 1, 'issue': 1, 'usa': 1, 'might': 1, 'ten': 1,
'anything': 1, 'white': 1, 'twenty': 1, 'study': 1, 'decision': 1, 'cant': 1, 'free': 1, 'national': 1, 'political': 1, 'kin
d': 1, 'lot': 1, 'reason': 1, 'look': 1, 'fire': 1, 'gun': 1, 'force': 1, 'three': 1, 'yes': 1, 'someone': 1, 'person': 1, 'b
ill': 1, 'power': 1, 'today': 1, 'give': 1, 'action': 1, 'tell': 1, 'place': 1, 'without': 1, 'party': 1, 'five': 1, 'fiftee
n': 1, 'never': 1, 'got': 1, 'united': 1, 'try': 1, 'put': 1, 'four': 1, 'situation': 1, 'doesnt': 1, 'package': 1, 'every':
1, 'didnt': 1, 'used': 1, 'show': 1, 'human': 1, 'plan': 1, 'next': 1, 'sixteen': 1, 'nothing': 1, 'done': 1, 'c': 1, 'anothe
r': 1, 'anyone': 1, 'six': 1, 'idea': 1, 'federal': 1, 'whether': 1, 'report': 1, 'evidence': 1, 'distribution': 1, 'workin
g': 1, 'find': 1, 'enough': 1, 'press': 1, 'administration': 1, 'youre': 1, 'thought': 1, 'argument': 1, 'statement': 1, 'h
```

10. Top 40 % tokens for class **rec.sport.hockey** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for rec.sport.hockey are: \n ",tf_idf_final['rec.sport.hockey'])
```

```
Top 10 % terms using tf_idf for feature selection for rec.sport.hockey are:
  {'one': 1, 'zero': 1, 'two': 1, 'game': 1, 'three': 1, 'team': 1, 'line': 1, 'subject': 1, 'apr': 1, 'organization': 1, 'da
te': 1, 'newsgroups': 1, 'messageid': 1, 'path': 1, 'four': 1, 'gmt': 1, 'one thousand, nine hundred and ninety-three': 1, 'f
ive': 1, 'six': 1, 'player': 1, 'reference': 1, 'writes': 1, 'university': 1, 'would': 1, 'go': 1, 'year': 1, 'play': 1, 'sen
der': 1, 'get': 1, 'article': 1, 'nntppostinghost': 1, 'think': 1, 'goal': 1, 'fan': 1, 'seven': 1, 'win': 1, 'like': 1, 'tim
e': 1, 'dont': 1, 'ninety-three': 1, 'first': 1, 'period': 1, 'know': 1, 'good': 1, 'twenty-five': 1, 'five hundred and fift
y': 1, 'ten': 1, 'last': 1, 'see': 1, 'im': 1, 'la': 1, 'twenty': 1, 'point': 1, 'new': 1, 'twenty-one': 1, 'well': 1, 'v':
1, 'shot': 1, 'blue': 1, 'pittsburgh': 1, 'leaf': 1, 'wing': 1, 'eleven': 1, 'eight': 1, 'even': 1, 'fifteen': 1, 'sixteen':
1, 'back': 1, 'news': 1, 'twenty-three': 1, 'really': 1, 'toronto': 1, 'second': 1, 'going': 1, 'nine': 1, 'boston': 1, 'divi
sion': 1, 'way': 1, 'best': 1, 'make': 1, 'let': 1, 'john': 1, 'also': 1, 'people': 1, 'got': 1, 'series': 1, 'played': 1, 'l
eague': 1, 'next': 1, 'many': 1, 'night': 1, 'mike': 1, 'say': 1, 'l': 1, 'could': 1, 'canada': 1, 'right': 1, 'twenty-six':
1, 'take': 1, 'st': 1, 'didnt': 1, 'much': 1, 'great': 1, 'better': 1, 'power': 1, 'eighteen': 1, 'april': 1, 'twelve': 1, 'n
ineteen': 1, 'king': 1, 'twenty-two': 1, 'guy': 1, 'ice': 1, 'since': 1, 'twenty-four': 1, 'name': 1, 'lead': 1, 'thirteen':
1, 'final': 1, 'mon': 1, 'thirty': 1, 'cant': 1, 'bos': 1, 'made': 1, 'chicago': 1, 'usa': 1, 'show': 1, 'playing': 1, 'jet':
1, 'never': 1, 'come': 1, 'vancouver': 1, 'fourteen': 1, 'world': 1, 'state': 1, 'van': 1, 'calgary': 1, 'winnipeg': 1, 'stil
l': 1, 'give': 1, 'said': 1, 'pick': 1, 'he': 1, 'flame': 1, 'seventeen': 1, 'may': 1, 'post': 1, 'roger': 1, 'usenet': 1, 't
hing': 1, 'put': 1, 'look': 1, 'fri': 1, 'net': 1, 'louis': 1, 'want': 1, 'ny': 1, 'mark': 1, 'anyone': 1, 'u': 1, 'sure': 1,
'nj': 1, 'lot': 1, 'joseph': 1, 'round': 1, 'little': 1, 'g': 1, 'coverage': 1, 'third': 1, 'bad': 1, 'another': 1, 'san': 1,
'red': 1, 'distribution': 1, 'thirty-three': 1, 'mean': 1, 'thirty-one': 1, 'quebec': 1, 'european': 1, 'end': 1, 'canadian':
```

11. Top 40 % tokens for class **sci.space** using TF-IDF as feature selection.

```
print("Top",k,"% terms using tf_idf for feature selection for sci.space are: \n ",tf_idf_final['sci.space'])
```

```
Top 10 % terms using tf_idf for feature selection for sci.space are:
  {'space': 1, 'one': 1, 'line': 1, 'date': 1, 'subject': 1, 'organization': 1, 'path': 1, 'newsgroups': 1, 'messageid': 1,
'would': 1, 'apr': 1, 'one thousand, nine hundred and ninety-three': 1, 'gmt': 1, 'reference': 1, 'writes': 1, 'article': 1,
'system': 1, 'like': 1, 'u': 1, 'sender': 1, 'may': 1, 'time': 1, 'two': 1, 'nasa': 1, 'nntppostinghost': 1, 'get': 1, 'missi
on': 1, 'university': 1, 'also': 1, 'year': 1, 'earth': 1, 'dont': 1, 'could': 1, 'moon': 1, 'think': 1, 'first': 1, 'ninety-
three': 1, 'xref': 1, 'cantaloupesrvcscmuedu': 1, 'program': 1, 'new': 1, 'know': 1, 'three': 1, 'much': 1, 'science': 1, 'th
ing': 1, 'make': 1, 'people': 1, 'see': 1, 'data': 1, 'distribution': 1, 'pat': 1, 'news': 1, 'well': 1, 'technology': 1, 'us
e': 1, 'cost': 1, 'world': 1, 'work': 1, 'day': 1, 'way': 1, 'good': 1, 'project': 1, 'six': 1, 'need': 1, 'go': 1, 'even':
1, 'im': 1, 'research': 1, 'idea': 1, 'flight': 1, 'right': 1, 'question': 1, 'part': 1, 'center': 1, 'problem': 1, 'five':
1, 'information': 1, 'high': 1, 'station': 1, 'many': 1, 'long': 1, 'four': 1, 'since': 1, 'power': 1, 'back': 1, 'might': 1,
'something': 1, 'want': 1, 'design': 1, 'around': 1, 'sun': 1, 'give': 1, 'light': 1, 'find': 1, 'still': 1, 'planet': 1, 'lo
ok': 1, 'available': 1, 'used': 1, 'twenty-three': 1, 'large': 1, 'enough': 1, 'astronomy': 1, 'put': 1, 'twenty': 1, 'say':
1, 'going': 1, 'better': 1, 'launch': 1, 'star': 1, 'anyone': 1, 'software': 1, 'national': 1, 'international': 1, 'using':
1, 'said': 1, 'group': 1, 'mass': 1, 'ten': 1, 'computer': 1, 'commercial': 1, 'twenty-two': 1, 'point': 1, 'really': 1, 'us
a': 1, 'april': 1, 'usenet': 1, 'support': 1, 'small': 1, 'lot': 1, 'development': 1, 'life': 1, 'etc': 1, 'thats': 1, 'mone
y': 1, 'thirty': 1, 'book': 1, 'tue': 1, 'service': 1, 'eight': 1, 'satellite': 1, 'fifteen': 1, 'j': 1, 'communication': 1,
'source': 1, 'number': 1, 'big': 1, 'command': 1, 'object': 1, 'energy': 1, 'model': 1, 'field': 1, 'test': 1, 'sixteen': 1,
'real': 1, 'email': 1, 'probably': 1, 'little': 1, 'institute': 1, 'eleven': 1, 'cant': 1, 'made': 1, 'last': 1, 'different':
1, 'twenty-seven': 1, 'w': 1, 'set': 1, 'please': 1, 'mean': 1, 'come': 1, 'next': 1, 'post': 1, 'general': 1, 'access': 1,
```

For K=1

1. Actual classes for testing documents.

```
Actual classes for test documents are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'comp.graphics', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'comp.gr
aphics', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'sci.med', '537
52': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '178934': 'talk.
politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphics', '3906
3': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graphics', '592
96': 'sci.med', '60905': 'sci.space', '179096': 'talk.politics.misc', '54776': 'rec.sport.hockey', '53876': 'rec.sport.hocke
y', '58045': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'sci.med',
'38259': 'comp.graphics', '59562': 'sci.med', '59515': 'sci.med', '58941': 'sci.med', '58110': 'sci.med', '37950': 'comp.grap
hics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space', '58770': 'sc
i.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'comp.graphics', '58930': 'sci.me
d', '61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '58112': 'sci.med',
'61487': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 'rec.sport.hockey',
'59262': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey', '60184': 'sci.spac
e', '59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics', '58863': 'sci.me
d', '38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'talk.politics.misc',
```

2. Predicted classes using **mutual information** as feature selection.

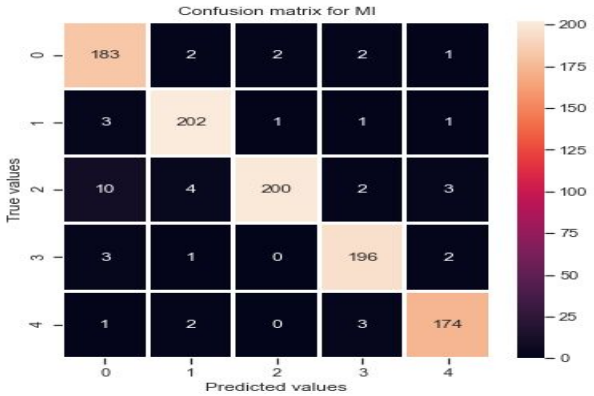
```
Predicted classes for test files using mutual information as feature selection are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'rec.sport.hockey', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'sci.
space', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'comp.graphics',
'53752': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '178934': 't
alk.politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphics',
'39063': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graphics',
'59296': 'sci.med', '60905': 'sci.space', '179096': 'talk.politics.misc', '54776': 'rec.sport.hockey', '53876': 'rec.sport.ho
ckey', '58045': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'sci.me
d', '38259': 'comp.graphics', '59562': 'sci.med', '59515': 'sci.med', '58941': 'sci.med', '58110': 'sci.med', '37950': 'comp.
graphics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space', '58770':
'sci.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'sci.med', '58930': 'sci.med',
'61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '58112': 'sci.med', '6148
7': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 'rec.sport.hockey', '5926
2': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey', '60184': 'sci.space',
'59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics', '58863': 'sci.med',
'38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'talk.politics.misc', '60
```

3. Predicted classes using **TF-IDF** as feature selection.

```
Predicted classes for test files using tf_idf as feature selection are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'rec.sport.hockey', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'sci.
space', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'comp.graphics',
'53752': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '178934': 't
alk.politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphics',
'39063': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graphics',
'59296': 'sci.med', '60905': 'sci.space', '179096': 'talk.politics.misc', '54776': 'rec.sport.hockey', '53876': 'rec.sport.ho
ckey', '58045': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'sci.me
d', '38259': 'comp.graphics', '59562': 'sci.med', '59515': 'sci.med', '58941': 'sci.med', '58110': 'sci.med', '37950': 'comp.
graphics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space', '58770':
'sci.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'sci.med', '58930': 'sci.med',
'61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '58112': 'sci.med', '6148
7': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 'rec.sport.hockey', '5926
2': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey', '60184': 'sci.space',
'59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics', '58863': 'sci.med',
'38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'talk.politics.misc', '60
```

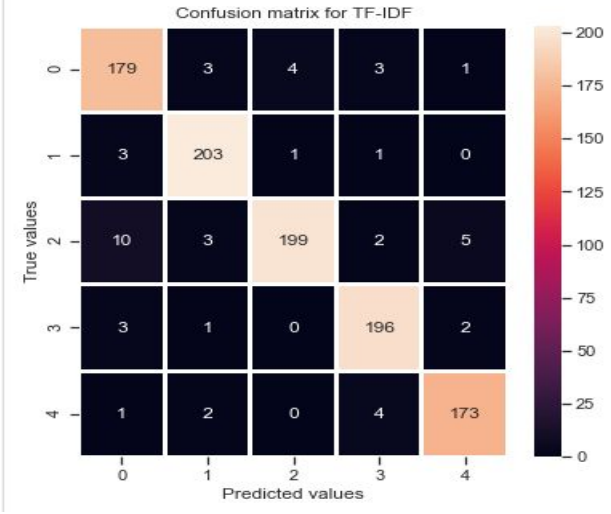
4. Accuracy and confusion matrix for **mutual information**.

```
accuracy after choosing mutual information as feature selection in KNN is:  0.955955955955956
Confusion matrix for MI is:
 [[183   2   2   2   1]
 [  3 202   1   1   1]
 [ 10   4 200   2   3]
 [  3   1   0 196   2]
 [  1   2   0   3 174]]
```



5. Accuracy and confusion matrix for **TF-IDF.**

```
accuracy after choosing tf_idf as feature selection in KNN is:  0.950950950950951
Confusion matrix for MI is:
 [[179   3   4   3   1]
 [  3 203   1   1   0]
 [ 10   3 199   2   5]
 [  3   1   0 196   2]
 [  1   2   0   4 173]]
```



For K=3

1. Actual classes for testing documents.

```
Actual classes for test documents are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'comp.graphics', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'comp.gr
aphics', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'sci.med', '537
52': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '178934': 'talk.
politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphics', '3906
3': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graphics', '592
96': 'sci.med', '60905': 'sci.space', '179096': 'talk.politics.misc', '54776': 'rec.sport.hockey', '53876': 'rec.sport.hocke
y', '58045': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'sci.med',
'38259': 'comp.graphics', '59562': 'sci.med', '59515': 'sci.med', '58941': 'sci.med', '58110': 'sci.med', '37950': 'comp.grap
hics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space', '58770': 'sc
i.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'comp.graphics', '58930': 'sci.me
d', '61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '58112': 'sci.med',
'61487': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 'rec.sport.hockey',
'59262': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey', '60184': 'sci.spac
e', '59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics', '58863': 'sci.me
d', '38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'talk.politics.misc',
```

2. Predicted classes using **mutual information** as feature selection.

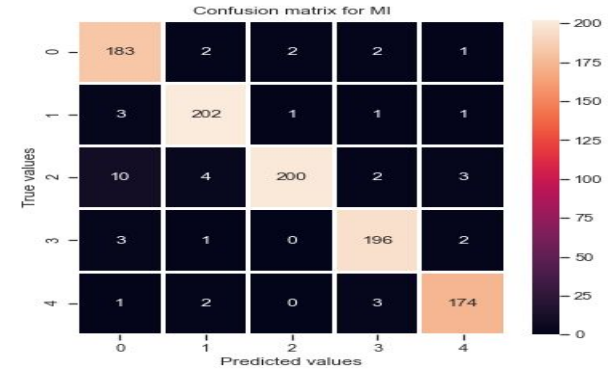
```
Predicted classes for test files using mutual information as feature selection are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'rec.sport.hockey', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'sci.
space', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'comp.graphics',
'53752': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '178934': 't
alk.politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphics',
'39063': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graphics',
'59296': 'sci.med', '60905': 'sci.space', '179096': 'talk.politics.misc', '54776': 'rec.sport.hockey', '53876': 'rec.sport.ho
ckey', '58045': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'sci.me
d', '38259': 'comp.graphics', '59562': 'sci.med', '59515': 'rec.sport.hockey', '58941': 'sci.med', '58110': 'sci.med', '3795
0': 'comp.graphics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space',
'58770': 'sci.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'comp.graphics', '5893
0': 'comp.graphics', '61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '581
12': 'sci.med', '61487': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 're
c.sport.hockey', '59262': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey',
'60184': 'sci.space', '59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics',
'58863': 'sci.med', '38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'tal
```

3. Predicted classes using **TF-IDF** as feature selection.

```
Predicted classes for test files using tf_idf as feature selection are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'rec.sport.hockey', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'sci.
space', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'comp.graphics',
'53752': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '178934': 't
alk.politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphics',
'39063': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graphics',
'59296': 'sci.med', '60905': 'sci.space', '179096': 'sci.med', '54776': 'rec.sport.hockey', '53876': 'rec.sport.hockey', '580
45': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'sci.med', '38259':
'comp.graphics', '59562': 'sci.med', '59515': 'rec.sport.hockey', '58941': 'sci.med', '58110': 'sci.med', '37950': 'comp.grap
hics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space', '58770': 'sc
i.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'comp.graphics', '58930': 'comp.gr
aphics', '61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '58112': 'sci.me
d', '61487': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 'rec.sport.hocke
y', '59262': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey', '60184': 'sci.
space', '59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics', '58863': 'sc
i.med', '38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'talk.politics.mi
```

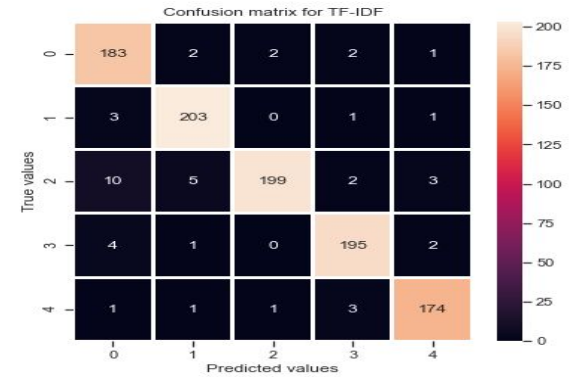
4. Accuracy and confusion matrix for **mutual information**.

```
accuracy after choosing mutual information as feature selection in KNN is:  0.955955955955956
Confusion matrix for MI is:
 [[183   2   2   2   1]
 [  3 202   1   1   1]
 [ 10   4 200   2   3]
 [  3   1   0 196   2]
 [  1   2   0   3 174]]
```



5. Accuracy and confusion matrix for **TF-IDF.**

```
accuracy after choosing tf_idf as feature selection in KNN is:  0.954954954954955
Confusion matrix for MI is:
 [[183   2   2   2   1]
 [  3 203   0   1   1]
 [ 10   5 199   2   3]
 [  4   1   0 195   2]
 [  1   1   1   3 174]]
```

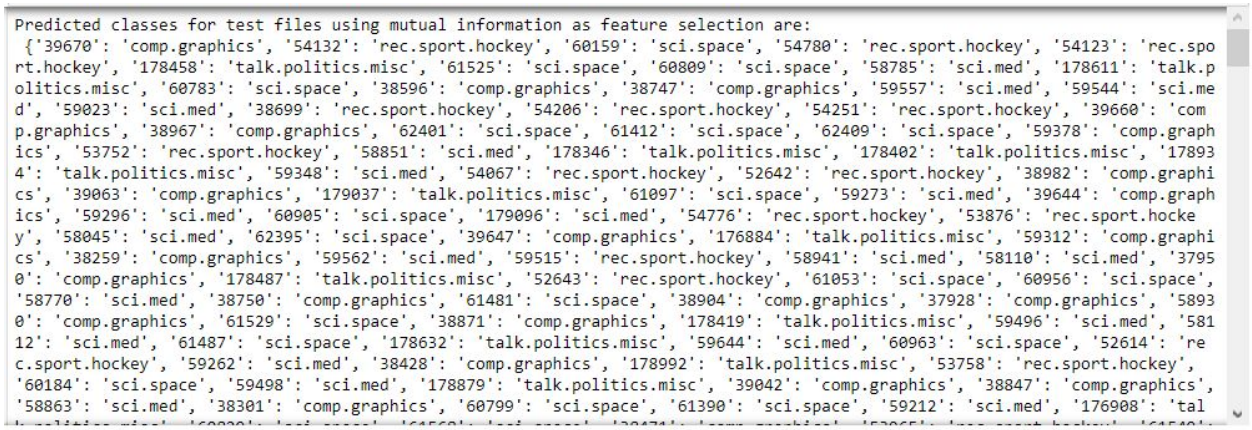


For K=5

1. Actual classes for testing documents.

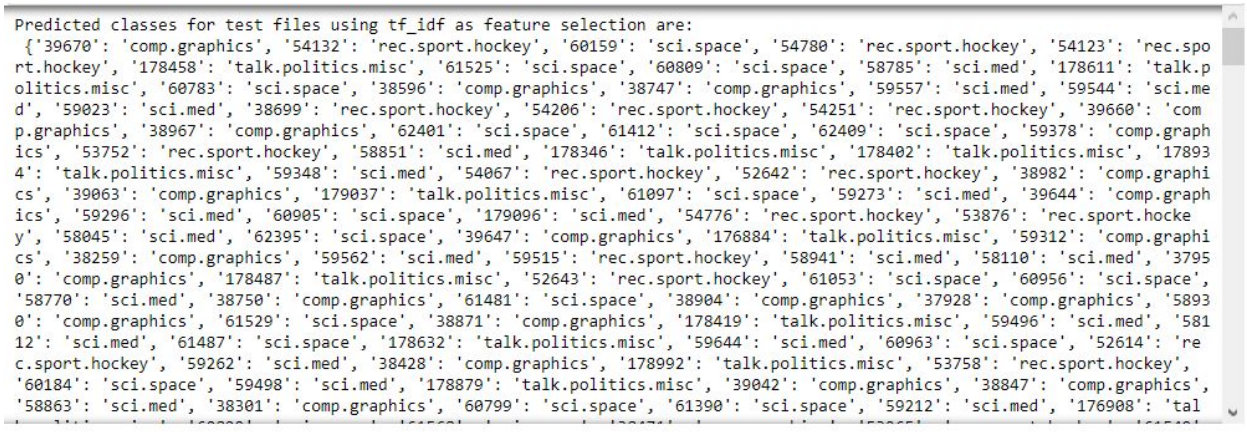
```
Actual classes for test documents are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'comp.graphics', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'comp.gr
aphics', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'sci.med', '537
52': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '178934': 'talk.
politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphics', '3906
3': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graphics', '592
96': 'sci.med', '60905': 'sci.space', '179096': 'talk.politics.misc', '54776': 'rec.sport.hockey', '53876': 'rec.sport.hocke
y', '58045': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'sci.med',
'38259': 'comp.graphics', '59562': 'sci.med', '59515': 'sci.med', '58941': 'sci.med', '58110': 'sci.med', '37950': 'comp.grap
hics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space', '58770': 'sc
i.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'comp.graphics', '58930': 'sci.me
d', '61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '58112': 'sci.med',
'61487': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 'rec.sport.hockey',
'59262': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey', '60184': 'sci.spac
e', '59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics', '58863': 'sci.me
d', '38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'talk.politics.misc',
```

2. Predicted classes using **mutual information** as feature selection.



```
Predicted classes for test files using mutual information as feature selection are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'rec.sport.hockey', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'com
p.graphics', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'comp.graph
ics', '53752': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '17893
4': 'talk.politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphi
cs', '39063': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graph
ics', '59296': 'sci.med', '60905': 'sci.space', '179096': 'sci.med', '54776': 'rec.sport.hockey', '53876': 'rec.sport.hocke
y', '58045': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'comp.graphi
cs', '38259': 'comp.graphics', '59562': 'sci.med', '59515': 'rec.sport.hockey', '58941': 'sci.med', '58110': 'sci.med', '3795
0': 'comp.graphics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space',
'58770': 'sci.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'comp.graphics', '5893
0': 'comp.graphics', '61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '581
12': 'sci.med', '61487': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 're
c.sport.hockey', '59262': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey',
'60184': 'sci.space', '59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics',
'58863': 'sci.med', '38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'tal
```

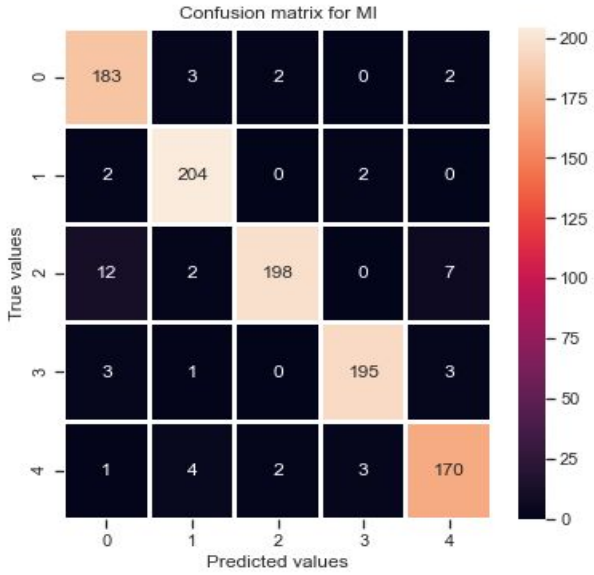
3. Predicted classes using **TF-IDF** as feature selection.



```
Predicted classes for test files using tf_idf as feature selection are:
 {'39670': 'comp.graphics', '54132': 'rec.sport.hockey', '60159': 'sci.space', '54780': 'rec.sport.hockey', '54123': 'rec.spo
rt.hockey', '178458': 'talk.politics.misc', '61525': 'sci.space', '60809': 'sci.space', '58785': 'sci.med', '178611': 'talk.p
olitics.misc', '60783': 'sci.space', '38596': 'comp.graphics', '38747': 'comp.graphics', '59557': 'sci.med', '59544': 'sci.me
d', '59023': 'sci.med', '38699': 'rec.sport.hockey', '54206': 'rec.sport.hockey', '54251': 'rec.sport.hockey', '39660': 'com
p.graphics', '38967': 'comp.graphics', '62401': 'sci.space', '61412': 'sci.space', '62409': 'sci.space', '59378': 'comp.graph
ics', '53752': 'rec.sport.hockey', '58851': 'sci.med', '178346': 'talk.politics.misc', '178402': 'talk.politics.misc', '17893
4': 'talk.politics.misc', '59348': 'sci.med', '54067': 'rec.sport.hockey', '52642': 'rec.sport.hockey', '38982': 'comp.graphi
cs', '39063': 'comp.graphics', '179037': 'talk.politics.misc', '61097': 'sci.space', '59273': 'sci.med', '39644': 'comp.graph
ics', '59296': 'sci.med', '60905': 'sci.space', '179096': 'sci.med', '54776': 'rec.sport.hockey', '53876': 'rec.sport.hocke
y', '58045': 'sci.med', '62395': 'sci.space', '39647': 'comp.graphics', '176884': 'talk.politics.misc', '59312': 'comp.graphi
cs', '38259': 'comp.graphics', '59562': 'sci.med', '59515': 'rec.sport.hockey', '58941': 'sci.med', '58110': 'sci.med', '3795
0': 'comp.graphics', '178487': 'talk.politics.misc', '52643': 'rec.sport.hockey', '61053': 'sci.space', '60956': 'sci.space',
'58770': 'sci.med', '38750': 'comp.graphics', '61481': 'sci.space', '38904': 'comp.graphics', '37928': 'comp.graphics', '5893
0': 'comp.graphics', '61529': 'sci.space', '38871': 'comp.graphics', '178419': 'talk.politics.misc', '59496': 'sci.med', '581
12': 'sci.med', '61487': 'sci.space', '178632': 'talk.politics.misc', '59644': 'sci.med', '60963': 'sci.space', '52614': 're
c.sport.hockey', '59262': 'sci.med', '38428': 'comp.graphics', '178992': 'talk.politics.misc', '53758': 'rec.sport.hockey',
'60184': 'sci.space', '59498': 'sci.med', '178879': 'talk.politics.misc', '39042': 'comp.graphics', '38847': 'comp.graphics',
'58863': 'sci.med', '38301': 'comp.graphics', '60799': 'sci.space', '61390': 'sci.space', '59212': 'sci.med', '176908': 'tal
```

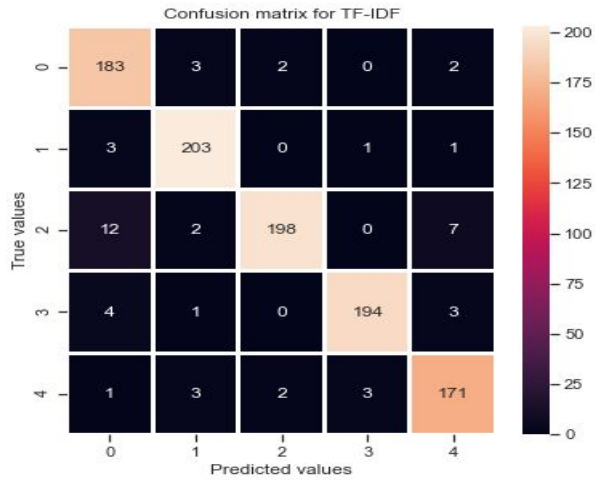
4. Accuracy and confusion matrix for **mutual information**.

```
accuracy after choosing mutual information as feature selection in KNN is:  0.950950950950951
Confusion matrix for MI is:
 [[183   3   2   0   2]
 [  2 204   0   2   0]
 [ 12   2 198   0   7]
 [  3   1   0 195   3]
 [  1   4   2   3 170]]
```



5. Accuracy and confusion matrix for **TF-IDF.**

```
accuracy after choosing tf_idf as feature selection in KNN is:  0.94994994994995
Confusion matrix for MI is:
 [[183   3   2   0   2]
 [  3 203   0   1   1]
 [ 12   2 198   0   7]
 [  4   1   0 194   3]
 [  1   3   2   3 171]]
```



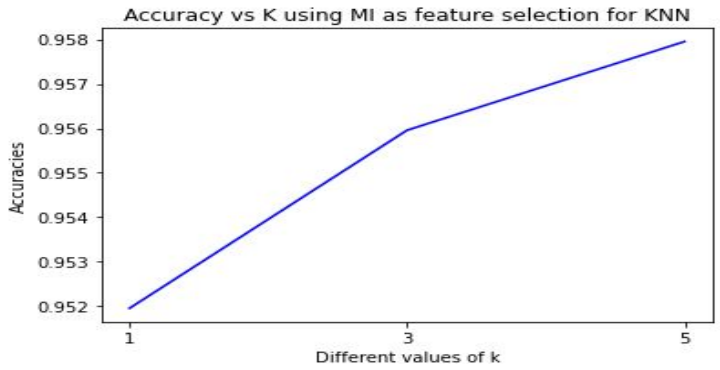
Graph between different values of k and accuracies using **mutual information** as feature selection technique.

```
: import matplotlib.pyplot as plt
  a=[0.9519519519519519,0.955955955955956,0.957950950950951]

  p=["1", "3", "5"]

  plt.plot(p, a,color="blue")
  plt.xlabel("Different values of k")
  plt.ylabel("Accuracies")
  plt.title("Accuracy vs K using MI as feature selection for KNN")
```

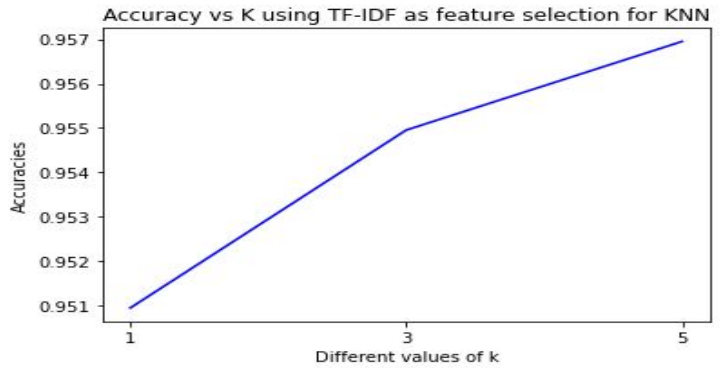
```
: Text(0.5, 1.0, 'Accuracy vs K using MI as feature selection for KNN')
```



Graph between different values of k and accuracies using **TF-IDF** as feature selection technique.

```
: a1=[0.950950950950951, 0.954954954954955,0.95694994994995]
  p=["1", "3", "5"]
  plt.plot(p, a1,color="blue")
  plt.xlabel("Different values of k")
  plt.ylabel("Accuracies")
  plt.title("Accuracy vs K using TF-IDF as feature selection for KNN")
```

```
: Text(0.5, 1.0, 'Accuracy vs K using TF-IDF as feature selection for KNN')
```



Analysis:

1. As each document is represented in a high dimensional space where dimension corresponds to a term, there can be many rare terms or noises which may lead the classifier in different directions. In other words, the presence of such terms can reduce the accuracy of our classifiers. Hence, there is a need to pick features which can increase the efficiency and effectiveness of our text classification algorithms.
2. Feature selection also reduces overfitting of the classifier.
3. Mutual information is being used for feature selection because it tells how much information a term contains about the class.
4. TF-IDF is being used for feature selection because it takes into account the frequency of term that occurs in a class and that particular term is coming in how many numbers of documents of a particular class. If a term occurs frequently in every document, then it can be a stop word. If a term is occurring frequently in less numbers of documents then it may be an important term. So, it is necessary to identify such terms.
5. Feature selection also reduces the complexity because we work on a subset of the terms.
6. As both the feature selection techniques are different we cannot say which one is better than the other.
7. Naïve Bayes classifier takes probabilistic **estimation** route and generates probabilities for each class. It assumes **conditional independence** between the features and uses a maximum likelihood hypothesis.
8. KNN classifier calculates the **similarity** between classes and chooses K nearest neighbors to classify documents.
9. Naïve Bayes is a quick learning algorithm hence it is faster than KNN.
10. As our dataset is very large and as in KNN test time proportional to the size of the training set, the larger the training set, the longer it takes time to classify documents. While in Naïve Bayes, we already have stored conditional probability for each term at the time of training. Hence, during test time we just have to take probabilities of each term of the test document and multiply them. So, Naïve Bayes is faster than KNN and performs efficiently for large datasets.
11. In Spite of being a lazy classifier, KNN classifies the documents more accurately than Naïve Bayes. Both the classifiers have their own advantages and disadvantages.
12. When graph is plotted between different top k% features and accuracy values for both tf-idf and mutual information based feature selection, we can see that with the increasing value of k, the accuracies are also increasing because we are selecting more number of important features each time. But if we will select 100% terms for feature selection then definitely it will not perform well because it may contain less informative terms or noises which will decrease the efficiency of our algorithms. Also, there will be no point of feature selection in this case.

13. For Naïve Bayes, graphs are plotted between different split ratios of testing and training documents. We can see that as the percentage of training documents is decreasing, the accuracies are also decreasing. This happened because we are getting less documents for training and hence testing documents are not classified accurately due to smaller training sets.
14. For KNN classifier, a graph is plotted between different values of K, where K is k nearest neighbors and accuracies. We can see that for k=1 the accuracies are coming low both for mutual information and TF-IDF feature selection techniques. This is because if we choose only one nearest neighbor, then any outlier can also come into account which may degrade the performance of the classifier. In other words we can say that noise may have higher influence on the result.
15. For higher values of K, accuracies are increasing because we are taking into account more number of similar documents for classification.