Assignment 3: Naive Bayes Classification

**Team**

Roychowdhury, Saikat <rychwdh2@illinois.edu>; 3 CREDITS

Abhinav Sharma <abhinavsharma3105@gmail.com> : 3 CREDITS

Shyam Rajendran <srajend2@illinois.edu> : 3 CREDITS

INDEX

Part 1: Digit classification 3

Part 1.1 Single pixels as features 3

Estimated priors P(class) 3

Maximum a Posteriori (MAP) classification 3

MAP Vs ML observation 4

Experimenting with different K values 5

PART 1 EXTRA CREDIT 6

Apply your Naive Bayes classifier with various features to face data. 6

Part 2 (for everybody): Text Document Classification 7

Spam detection 7

Classification Rate per class 7

Experimenting with different K values 7

Top 20 Words in SPAM CLASS with highest likelihood [ WORD : Likelihood ] 8

Top 20 Words in NON SPAM CLASS with highest likelihood [ word : likelihood ] 9

Eight newsgroups 10

Results with K=1 as below 10

Classification rate per class: 10

Experimenting with different K values 10

Top 20 Words in NEWS CATEGORIES with highest likelihood 12

Highest-confusion pairs 16

Top 20 words with highest log-odds ration for each pair is as below 16

Extra Credit for Part 2 18

bag-of-words representations of the documents using word cloud maps 18

Statement of individual contribution 20

# Part 1: Digit classification

## Part 1.1 Single pixels as features

### Estimated priors P(class)

We estimated the priors for each class from the

**“class” -> “P(class)**

"0" -> "0.0958"

"1" -> "0.1126"

"2" -> "0.0976"

"3" -> "0.0986"

"4" -> "0.107"

"5" -> "0.0868"

"6" -> "0.1002"

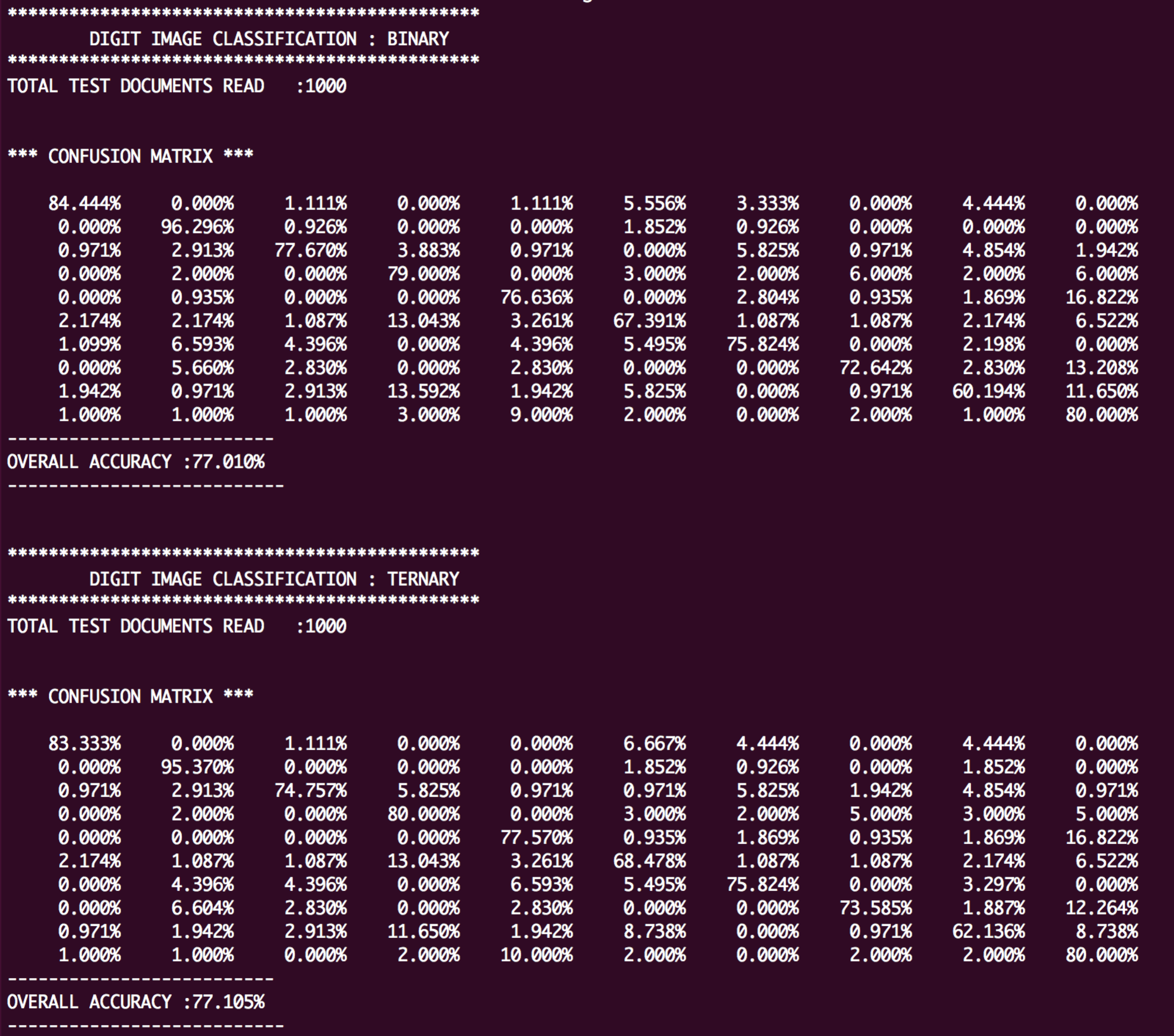
"7" -> "0.11"

"8" -> "0.0924"

"9" -> "0.099"

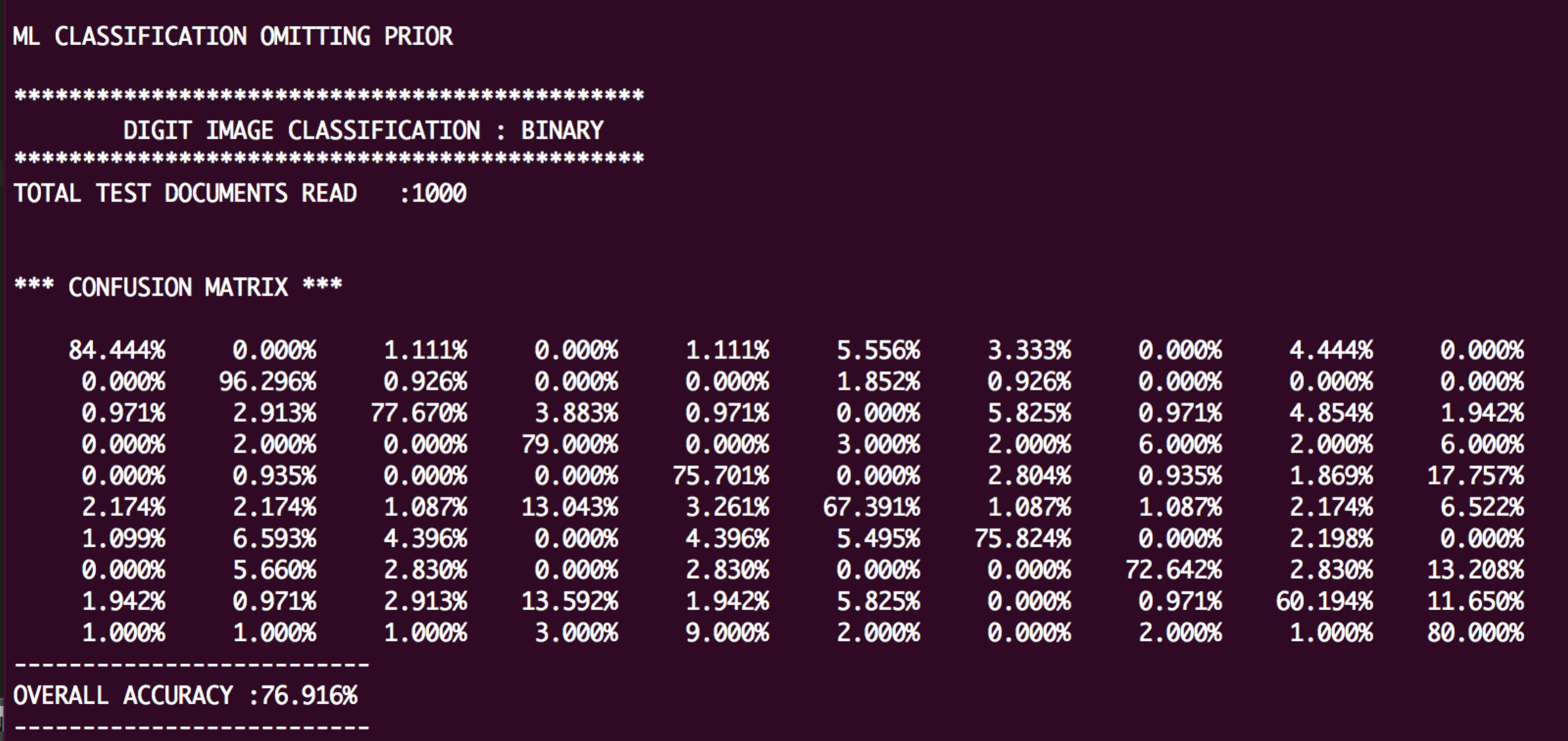
### Maximum a Posteriori (MAP) classification

Below is the screenshot of our MAP confusion matrix run with K = 1. The diagonal indicates the classification success for each digit class [0-9]



### MAP Vs ML observation

We compared the results with and without omitting the prior terms during classification and below are our results for K=1 ( binary classification )



As can be seen, the difference between ML and MAP is not very significant. **~ 1%**

### Experimenting with different K values

We also ran the classifier with different values of Laplace Smoothing factors in the range 1-50. We observed that maximum accuracy was achieved when K was 1.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**DIGIT IMAGE CLASSIFICATION : BINARY CLASSIFIER**

**K , Accuracy ( in % )**

**1.0,77.00973220352992**

2.0,76.51786976979153

3.0,76.20100217137053

4.0,76.10391479272975

5.0,75.67663337849181

6.0,75.6679377263179

7.0,75.45924207414399

8.0,75.36334915321942

9.0,75.36626177457865

10.0,75.37906450582798

11.0,75.14612271423155

12.0,75.14162792818337

13.0,75.03173781829325

14.0,74.93584489736867

15.0,74.83875751872789

16.0,74.5230843780231

17.0,74.4143887258492

18.0,74.3991510176005

19.0,74.49349064024202

20.0,74.3847949880681

21.0,74.49590609917921

22.0,74.38152120324091

23.0,74.48152120324092

24.0,74.16704252025231

25.0,73.96125948943762

26.0,73.74112042909053

27.0,73.64403305044975

28.0,73.54694567180897

29.0,73.1279722027722

30.0,72.93379744549064

31.0,72.93379744549064

32.0,72.73076217067522

33.0,72.73076217067522

34.0,72.52087206078511

35.0,72.52087206078511

36.0,72.5220665185013

37.0,72.32789176121975

38.0,72.21919610904584

39.0,72.31265405297107

40.0,72.10687102215638

41.0,72.09817536998247

42.0,72.09817536998247

43.0,72.09817536998247

44.0,72.00108799134169

45.0,72.00108799134169

46.0,71.99834023534243

47.0,71.99834023534243

48.0,71.88964458316852

49.0,71.78964458316851

50.0,71.69255720452773

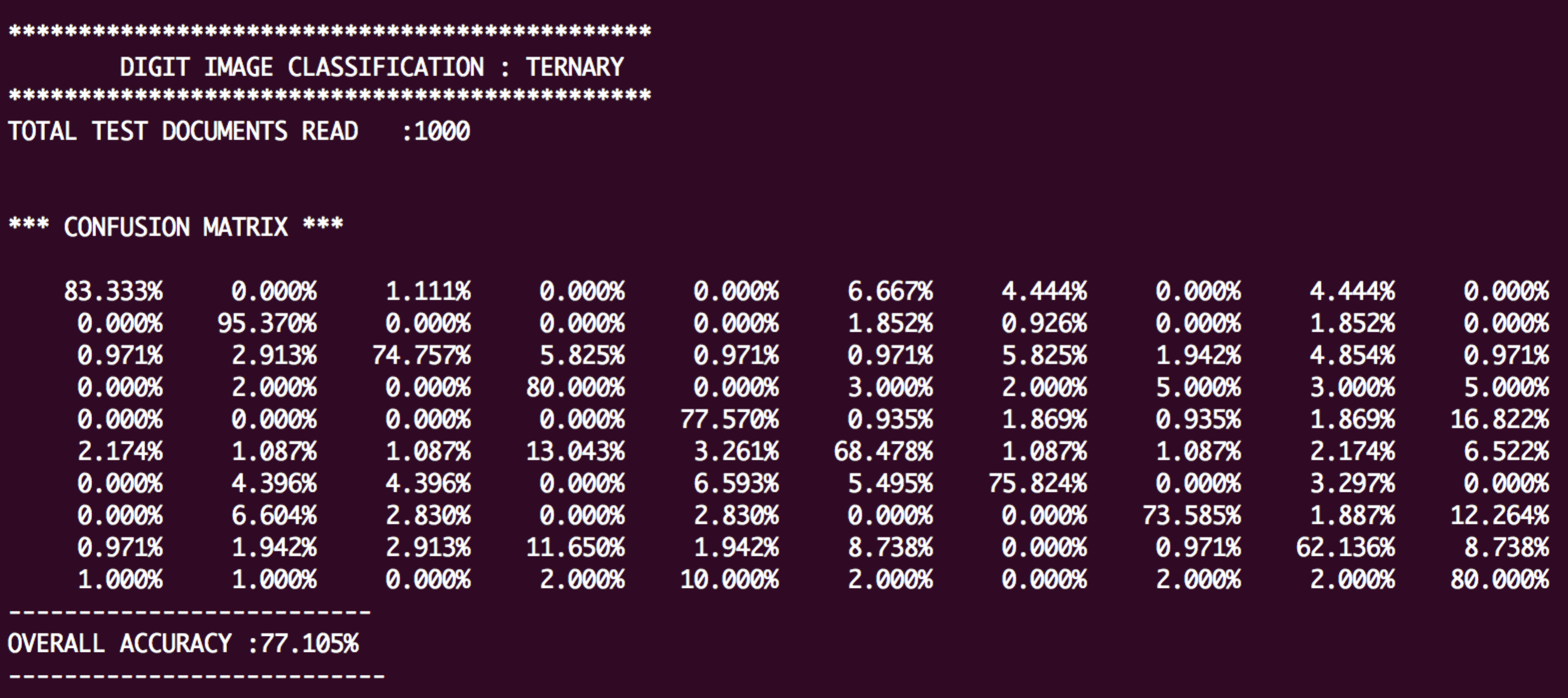
## PART 1 EXTRA CREDIT

Implement Ternary Features

Instead of considering the pixel values as either “1” or “0” depending on whether it is “background” : space or “foreground” : “+” or “#” , we gave unique values for each type of pixel to have ternary classification.

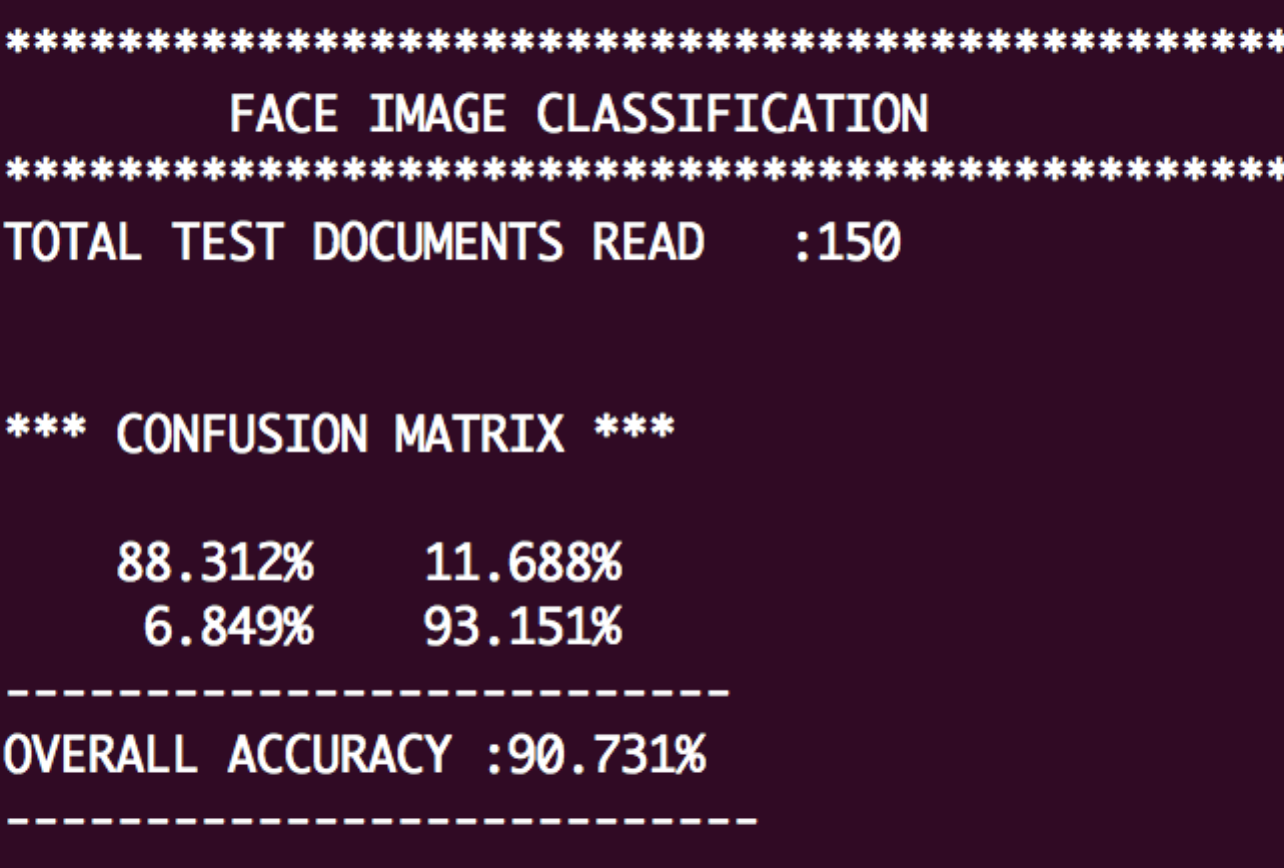
Below is the screen shot of the classifier with K=1.

We can see a slight improvement over binary classification.



### Apply your Naive Bayes classifier with various features to [face data](http://web.engr.illinois.edu/~slazebni/spring15/assignment3/facedata.zip).

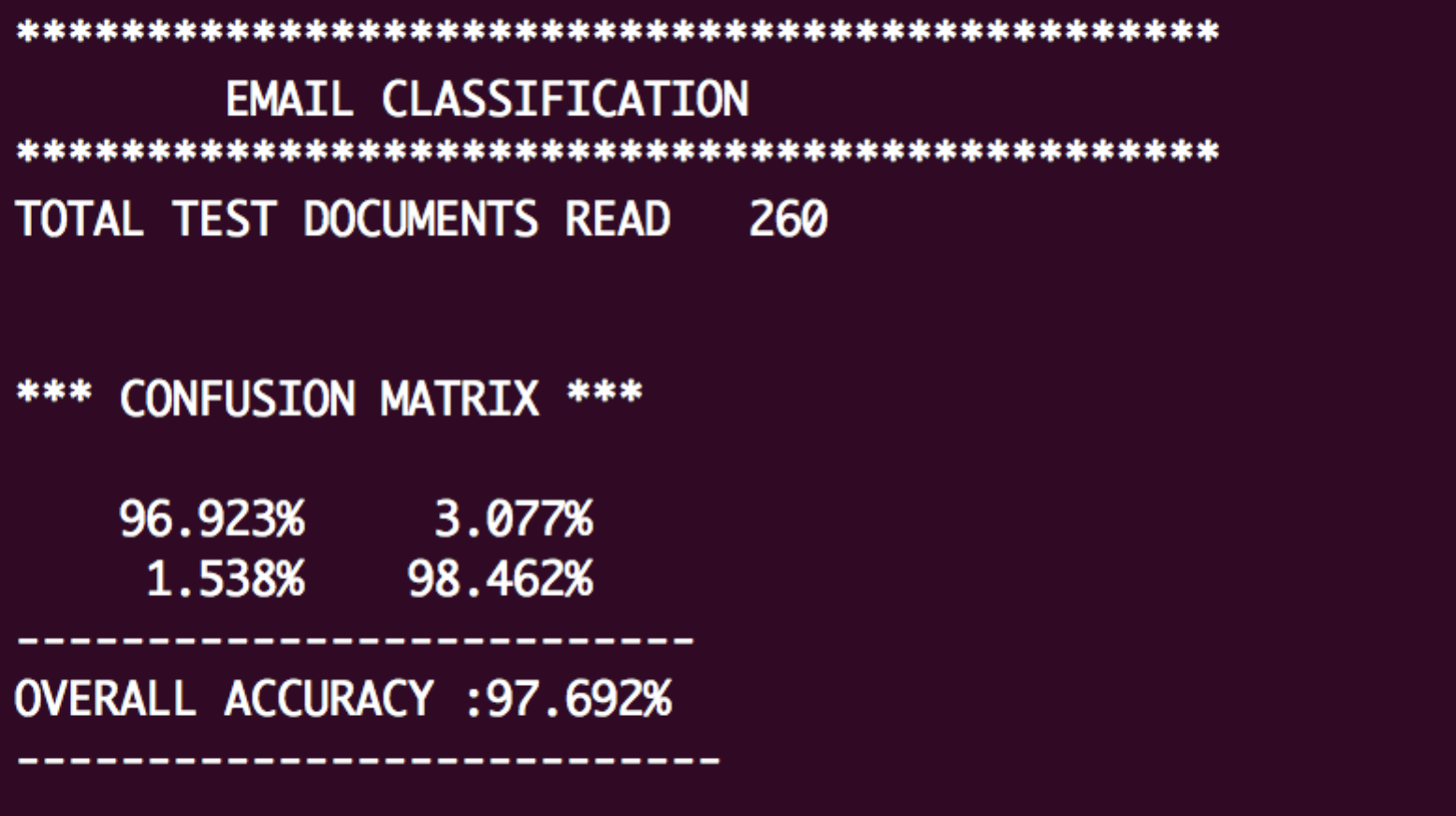
We applied the Naïve Bayes classifier to the face data. Below is the confusion matrix with K=1.



# Part 2 (for everybody): Text Document Classification

## Spam detection

Spam email classification : Naive Bayes classifier with K=1



## Classification Rate per class

**NonSpam Class classification Rate : 96.923%**

**Spam Class classification Rate : 98.423%**

### Experimenting with different K values

We also ran the classifier for K=1-50 and below are the results

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**SPAM CLASSIFICATION**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**K, Accuracy (in %)**

1 ,97.692

2 ,97.308

3 ,97.308

4 ,97.308

5 ,96.538

6 ,96.538

7 ,96.538

8 ,96.538

9 ,96.538

10 ,96.538

11 ,96.538

12 ,96.538

13 ,96.538

14 ,95.769

15 ,95.769

16 ,95.769

17 ,95.769

18 ,95.769

19 ,95.769

20 ,95.769

21 ,95.769

22 ,95.769

23 ,95.769

24 ,95.769

25 ,95.769

26 ,95.769

27 ,95.769

28 ,95.769

29 ,95.769

30 ,95.769

31 ,95.769

32 ,95.769

33 ,95.769

34 ,95.769

35 ,95.769

36 ,95.769

37 ,95.769

38 ,95.769

39 ,95.769

40 ,95.769

41 ,95.769

42 ,95.769

43 ,95.769

44 ,95.769

45 ,95.769

46 ,95.769

47 ,95.769

48 ,95.769

49 ,95.769

50 ,95.769

### Top 20 Words in SPAM CLASS with highest likelihood [ WORD : Likelihood ]

|  |  |
| --- | --- |
| email | 0.010771475 |
| s | 0.009422116 |
| order | 0.009047727 |
| report | 0.008220952 |
| our | 0.007534573 |
| address | 0.007448775 |
| mail | 0.007206982 |
| program | 0.006466005 |
| send | 0.006247611 |
| free | 0.005810825 |
| money | 0.00563923 |
| list | 0.005569032 |
| receive | 0.005171244 |
| name | 0.004898252 |
| business | 0.004750057 |
| one | 0.00432107 |
| d | 0.004227472 |
| work | 0.004126075 |
| com | 0.004094876 |
| nt | 0.004063677 |

### Top 20 Words in NON SPAM CLASS with highest likelihood [ word : likelihood ]

|  |  |
| --- | --- |
| language | 0.01041331 |
| university | 0.008350904 |
| s | 0.006095147 |
| linguistic | 0.004401028 |
| de | 0.004106398 |
| information | 0.004097191 |
| conference | 0.003489518 |
| workshop | 0.003323789 |
| email | 0.002964709 |
| paper | 0.002955502 |
| e | 0.002900259 |
| english | 0.002881844 |
| one | 0.002587215 |
| please | 0.002568801 |
| include | 0.002559593 |
| edu | 0.00250435 |
| http | 0.0024399 |
| abstract | 0.002338621 |
| address | 0.002329414 |
| papers | 0.002274171 |

## Eight newsgroups

### Results with K=1 as below

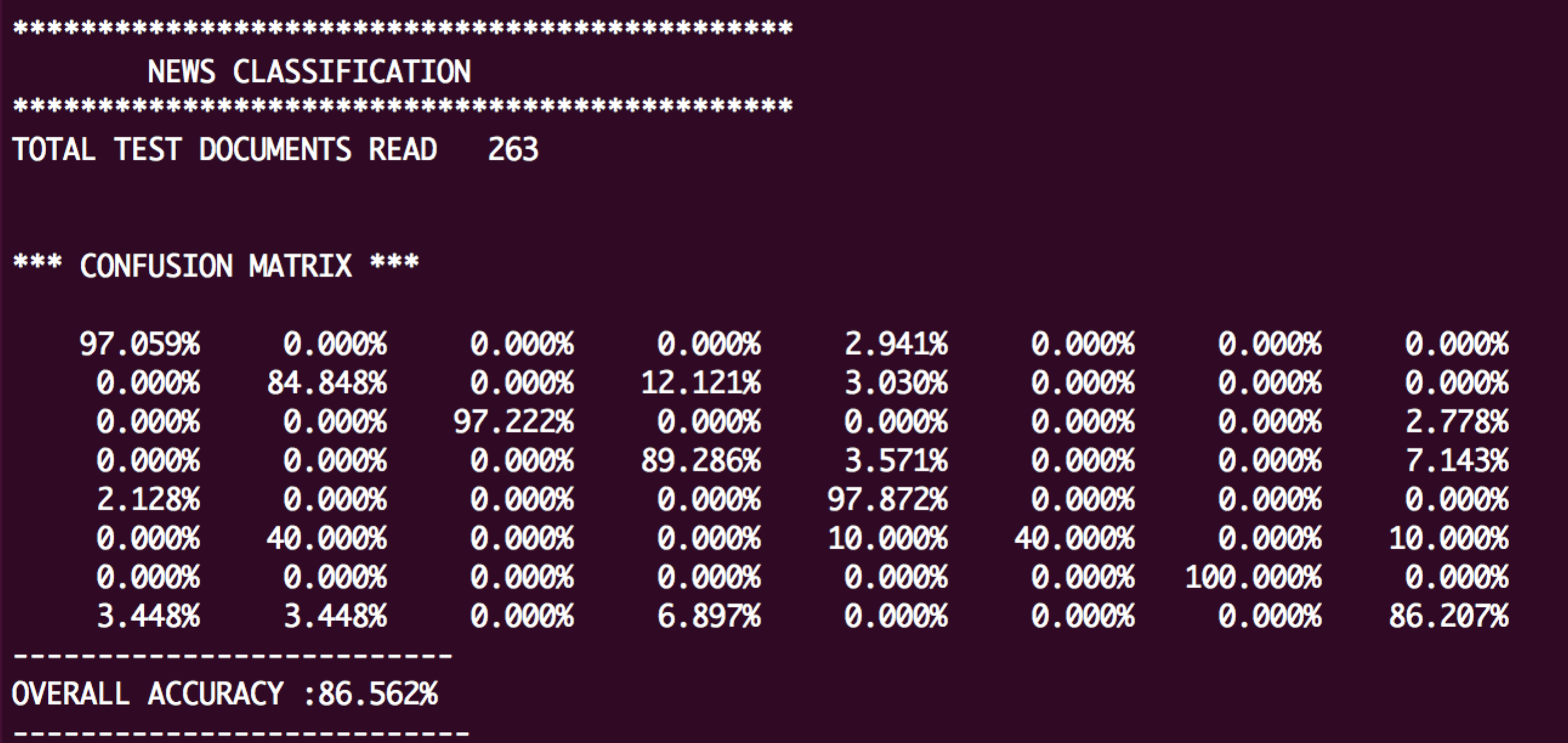


fig 6

### Classification rate per class:

|  |  |
| --- | --- |
| **CLASS** | **CLASSIFICATION RATE** |
| sci.space | **97.059%** |
| comp.sys.ibm.pc.hardware | **84.848%** |
| rec.sport.baseball | **97.22%** |
| comp.windows.x | **89.286%** |
| talk.politics.misc | **97.872%** |
| misc.forsale | **40%** |
| rec.sport.hockey | **100%** |
| comp.graphics | **86.207%** |

### Experimenting with different K values

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

NEWS CLASSIFICATION

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**K, Accuracy ( in %)**

1 ,86.562

2 ,87.189

3 ,89.249

4 ,89.249

5 ,88.983

6 ,88.711

7 ,88.439

8 ,88.439

9 ,88.396

10 ,88.130

11 ,87.540

12 ,86.310

13 ,84.664

14 ,84.664

15 ,82.839

16 ,81.774

17 ,80.529

18 ,79.465

19 ,78.235

20 ,77.436

21 ,75.466

22 ,75.487

23 ,74.157

24 ,72.901

25 ,71.096

26 ,69.658

27 ,68.204

28 ,66.767

29 ,65.946

30 ,64.600

31 ,63.350

32 ,62.184

33 ,61.753

34 ,60.495

35 ,58.733

36 ,57.744

37 ,55.963

38 ,55.277

39 ,54.930

40 ,53.500

41 ,52.341

42 ,51.962

43 ,51.057

44 ,48.546

45 ,47.913

46 ,47.218

47 ,47.484

48 ,45.988

49 ,45.269

50 ,44.650

### Top 20 Words in NEWS CATEGORIES with highest likelihood

|  |  |
| --- | --- |
| **comp.graphics** | |
| TRUE | 0.0002725403 |
| FALSE | 0.0000165176 |
| zyxel | 0.0000412940 |
| zvi | 0.0000247764 |
| zurich | 0.0000247764 |
| zur | 0.0000165176 |
| zugcsmil | 0.0000247764 |
| zphigs | 0.0000412940 |
| zorn | 0.0000165176 |
| zorg | 0.0000495528 |
| zopfi | 0.0000247764 |
| zoomingin | 0.0000247764 |
| zooming | 0.0000495528 |
| zoom | 0.0000908468 |
| zippy | 0.0000412940 |
| zipped | 0.0000165176 |
| zip | 0.0000330352 |
| zhenghao | 0.0000165176 |
| zhao | 0.0000247764 |
| zero | 0.0000330352 |
|  |  |
|  |  |
| **rec.sport.hockey** | |
| TRUE | 0.0003454367 |
| FALSE | 0.0000157017 |
| zurich | 0.0000157017 |
| zupancic | 0.0000157017 |
| zubov | 0.0000628067 |
| zoomed | 0.0000157017 |
| zones | 0.0000235525 |
| zone | 0.0003689892 |
| zombo | 0.0001099117 |
| zmolek | 0.0000471050 |
| zippety | 0.0000235525 |
| zipper | 0.0000314033 |
| zholtok | 0.0000157017 |
| zhivov | 0.0000471050 |
| zhitnik | 0.0000706575 |
| zhamnov | 0.0001177625 |
| zezel | 0.0000392542 |
| zettler | 0.0000314033 |
| zeta | 0.0000157017 |
| zero | 0.0000157017 |
|  |  |
| **misc.forsale** | |
| zuiko | 0.0000273999 |
| zorro | 0.0000273999 |
| zooromancer | 0.0000273999 |
| zoom | 0.0001232995 |
| zones | 0.0000273999 |
| zone | 0.0000821997 |
| zombies | 0.0000821997 |
| zimmerman | 0.0000410998 |
| zildjian | 0.0000410998 |
| zhen | 0.0000410998 |
| zeus | 0.0000273999 |
| zero | 0.0000821997 |
| zenith | 0.0000410998 |
| zell | 0.0000273999 |
| zack | 0.0000273999 |
| z | 0.0000273999 |
| yugoslavia | 0.0000273999 |
| yuanchieh | 0.0000273999 |
| ysaron | 0.0000273999 |
| youself | 0.0000273999 |
|  |  |
|  |  |
| **talk.politics.misc** | |
| FALSE | 0.0001224440 |
| zumwalt | 0.0000122444 |
| zoology | 0.0000122444 |
| zoologists | 0.0000122444 |
| zoologist | 0.0000122444 |
| zooid | 0.0000122444 |
| zoo | 0.0000122444 |
| zoning | 0.0000122444 |
| zones | 0.0000122444 |
| zone | 0.0000306110 |
| zionists | 0.0000122444 |
| zimmers | 0.0000122444 |
| zimmerman | 0.0000122444 |
| zimmer | 0.0000122444 |
| zeus | 0.0000122444 |
| zeroes | 0.0000122444 |
| zero | 0.0000244888 |
| zealand | 0.0001346884 |
| zb | 0.0000122444 |
| zabrecky | 0.0000122444 |
|  |  |
|  |  |
| **comp.windows.x** | |
| TRUE | 0.0004352452 |
| FALSE | 0.0002498630 |
| zzgbavk | 0.0000161202 |
| zymos | 0.0000241803 |
| zylorqgv | 0.0000161202 |
| zx | 0.0000161202 |
| zwdtklxiwwsa | 0.0000161202 |
| zt | 0.0000241803 |
| zsh | 0.0000241803 |
| zs | 0.0000161202 |
| zrmdfl | 0.0000161202 |
| zooms | 0.0000161202 |
| zooming | 0.0000564207 |
| zoom | 0.0000403005 |
| zone | 0.0000161202 |
| zombies | 0.0000161202 |
| zombie | 0.0000161202 |
| zok | 0.0000322404 |
|  |  |
|  |  |
| **rec.sport.baseball** | |
| TRUE | 0.0003740450 |
| FALSE | 0.0000187023 |
| zzzzzz | 0.0000467556 |
| zz | 0.0000467556 |
| zupcic | 0.0000467556 |
| zot | 0.0000187023 |
| zones | 0.0000187023 |
| zone | 0.0002244270 |
| zip | 0.0000280534 |
| zimmers | 0.0000654579 |
| zero | 0.0001028624 |
| zeile | 0.0000748090 |
| zavatson | 0.0000187023 |
| zaphod | 0.0000187023 |
| zap | 0.0000187023 |
| zane | 0.0000561068 |
| z | 0.0000280534 |
| yup | 0.0000561068 |
| youth | 0.0000187023 |
| youre | 0.0000187023 |
|  |  |
| comp.sys.ibm.pc.hardware | |
| TRUE | 0.0003640207 |
| zurichswitzerland | 0.0000234852 |
| zurich | 0.0000234852 |
| zou | 0.0000234852 |
| zorro | 0.0000234852 |
| zoom | 0.0000234852 |
| zone | 0.0000587130 |
| zip | 0.0000469704 |
| zhao | 0.0000234852 |
| zeus | 0.0000234852 |
| zero | 0.0000821982 |
| zeos | 0.0000469704 |
| zenon | 0.0000234852 |
| zeeff | 0.0000234852 |
| zawodny | 0.0000234852 |
| zappa | 0.0000234852 |
| zabbal | 0.0000234852 |
| yuri | 0.0000234852 |
| yung | 0.0000352278 |
| yulaev | 0.0000234852 |
|  |  |
|  |  |
| **sci.space** | |
| TRUE | 0.0003411186 |
| FALSE | 0.0000290314 |
| zwarte | 0.0000217735 |
| zware | 0.0000145157 |
| zwakke | 0.0000145157 |
| zwak | 0.0000145157 |
| zwaartepunten | 0.0000145157 |
| zulu | 0.0000145157 |
| zullen | 0.0000217735 |
| zoology | 0.0000725784 |
| zonker | 0.0000145157 |
| zoning | 0.0000145157 |
| zone | 0.0000580627 |
| zond | 0.0000362892 |
| zogeheten | 0.0000217735 |
| zodiacal | 0.0000435471 |
| zo | 0.0000362892 |
| zipping | 0.0000145157 |
| zijn | 0.0000362892 |
| zien | 0.0000145157 |
| zo | 0.0000161202 |
| znvpjw | 0.0000161202 |

### Highest-confusion pairs

From the confusion matrix in fig 6 Four highest-confusion pairs are as below

1. misc.forsale - comp.sys.ibm.pc.hardware : 40%
2. comp.windows.x - comp.sys.ibm.pc.hardware : 12.12%
3. comp.graphics - omp.windows.x : 7.43%
4. comp.windows.x - comp.graphics : 6.89%

### Top 20 words with highest log-odds ration for each pair is as below

**misc.forsale - comp.sys.ibm.pc.hardware : 40%**

|  |  |
| --- | --- |
| sale | 4.039162627 |
| shipping | 3.315415006 |
| copies | 3.300473426 |
| cover | 3.298320573 |
| appearance | 3.28966251 |
| art | 3.27506371 |
| items | 3.245210747 |
| books | 3.17459318 |
| games | 3.098607273 |
| miller | 3.044540052 |
| nm | 3.016369175 |
| appears | 2.933677459 |
| silver | 2.862218495 |
| mpc | 2.828316943 |
| car | 2.793225624 |
| book | 2.719117651 |
| stereo | 2.679896938 |
| black | 2.639074944 |
| star | 2.596515329 |
| fair | 2.596515329 |

**comp.windows.x - comp.sys.ibm.pc.hardware : 12.12%**

|  |  |
| --- | --- |
| scsi | 5.310771896 |
| drives | 4.728724127 |
| bus | 4.524709746 |
| chip | 4.137498078 |
| pin | 3.887843401 |
| isa | 3.820980456 |
| jumper | 3.581750767 |
| gateway | 3.511792178 |
| meg | 3.489813272 |
| heads | 3.489813272 |
| drive | 3.366012164 |
| led | 3.179658343 |
| controllers | 3.179658343 |
| motherboard | 3.110665472 |
| uwoca | 3.084348164 |
| jumpers | 3.061875308 |
| nec | 3.038885789 |
| mhz | 2.94124732 |
| modem | 2.907724628 |
| floppy | 2.861204612 |

**comp.graphics - omp.windows.x : 7.43%**

|  |  |
| --- | --- |
| xterm | 4.175351536 |
| xlib | 3.6263047 |
| shared | 3.47215402 |
| xt | 3.387893676 |
| openwindows | 3.337178584 |
| x | 3.288860443 |
| null | 3.181099263 |
| sco | 3.089161768 |
| hostname | 3.089161768 |
| consortium | 3.020168896 |
| gcc | 2.946060924 |
| printf | 2.893417191 |
| shell | 2.852031974 |
| makefile | 2.837847339 |
| cc | 2.748235181 |
| guidelines | 2.716486482 |
| FALSE | 2.716486482 |
| tcpip | 2.68369666 |
| startup | 2.578336144 |
| biz | 2.578336144 |

**comp.windows.x - comp.graphics : 6.89%**

|  |  |
| --- | --- |
| jpeg | 5.076344021 |
| gopher | 3.593886238 |
| surface | 3.374257629 |
| visualization | 3.229806346 |
| gif | 3.155005133 |
| quality | 3.1334145 |
| viewers | 3.044778428 |
| gamma | 2.986184263 |
| search | 2.968792521 |
| movie | 2.968792521 |
| image | 2.944373993 |
| siggraph | 2.933074438 |
| analysis | 2.857566886 |
| texture | 2.827713922 |
| rt | 2.827713922 |
| mirror | 2.817561551 |
| images | 2.81244645 |
| tracing | 2.765193566 |
| algorithms | 2.765193566 |
| fullcolor | 2.732403743 |

COMPLETE

dataset, and over 80% on the newsgroup dataset. Additionally, for each class, report the top 20 words with the highest likelihood. Finally, as in Part 1.1, take the pair of classes from the email dataset and four highest-confusion pairs from the newsgroup dataset, and display the top 20 words with the highest log-odds ratio for that pair of classes.

## Extra Credit for Part 2

### bag-of-words representations of the documents using word cloud maps

We created a visualization of bag of words and have hosted it online @ <URL PASTE>

Screenshot of the visualization as below < PASTE SCREEN SHOT BELOW>

# Statement of individual contribution

|  |  |
| --- | --- |
| **SHYAM RAJENDRAN** | **Implemented**  Part1  + Single pixels as features classification  + Ternary feature based classification  + Face data classification  Part2  + Spam Detection  + Newsgroup classification |
| **ABHINAV SHARMA** |  |
| **SAIKAT ROYCHOWDHURY** |  |