CMPE 442 MACHINE LEARNING
Assignment 2

Arda Andırın
17921934994

Q1) The size of my dictionary is 150138

Q2) The class priors are computed as follows

```
alt.atheism : 0.042425
comp.graphics : 0.051617
comp.os.ms-windows.misc : 0.052236
comp.sys.ibm.pc.hardware : 0.052148
comp.sys.mac.hardware : 0.051087
comp.windows.x : 0.052413
misc.forsale : 0.051706
rec.autos : 0.052501
rec.motorcycles : 0.052855
rec.sport.baseball : 0.052766
rec.sport.hockey : 0.053032
sci.crypt : 0.052590
sci.electronics : 0.052236
sci.med : 0.052501
sci.space : 0.052413
soc.religion.christian : 0.052943
talk.politics.guns : 0.048259
talk.politics.mideast : 0.049850
talk.politics.misc : 0.041100
talk.religion.misc : 0.033322
```

Q3) I have used the most frequent 50000 words. I have observed that doing so accuracy rate increased. Also, the probabilties were going near zero so python couldn't make comparisons, this resulted in a low accuracy (around 40%). To solve this I multipled each new probabilty of a word with 10^4. This way the accuracy rate got up to 80%.

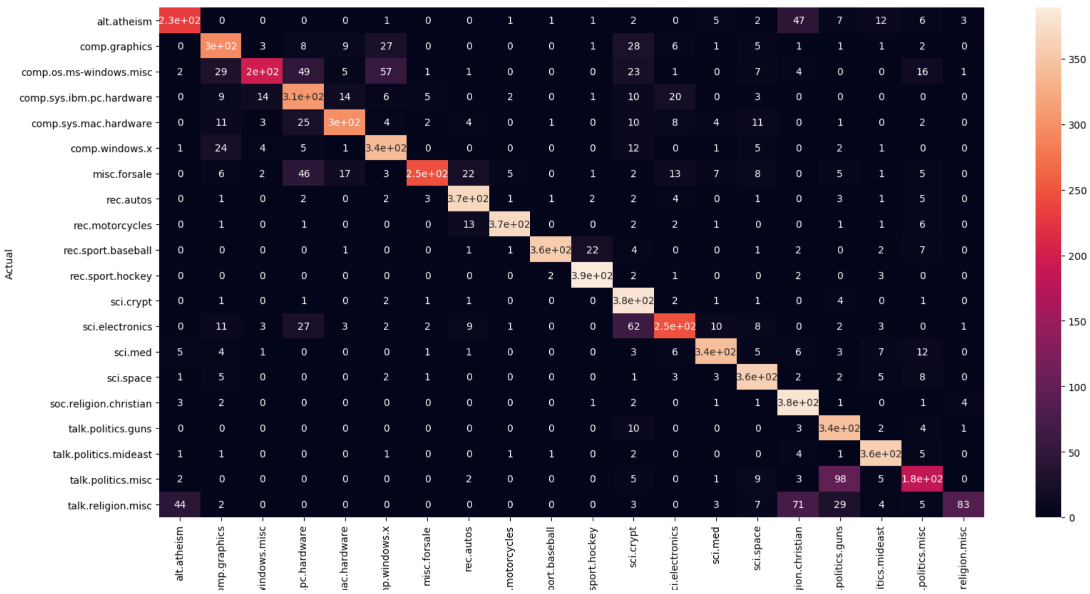
Here is our formula

$$\varphi_{k|y=0} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\left\{X_j{}^{(i)} = k \ \wedge Y^{(i)} = 0\right\} + 1}{\sum_{i=1}^{m} 1\{Y^{\{i\}} = 0\}\, n_i + |V|}$$

Here is the accuracy rate with the confusion matrix.

```
Number of correct predictions = 6087
Number of false predictions = 1445
Total number of tests = 7532

Accuracy rate is =  80.81518852894317 %
```



Q4) Yes for example the algorithm confuses the category talk.religion.misc with religion.christian. This is because the subjects are so related that appearance of a word in one of these categories appear also in the other category. The words god, atheism, jesus etc. are all common in religion related newsgroups, it is normal to see false positives. This can also be observed in other categories that are related to each other. For exapmle, some newsgroups in comp categories are also predicted wrong.

Q5) When we think of the strong indicators, what comes to the mind first is the most common words. Preproccessor already eliminates common words such as stopwords. What would be a really strong indicator is words that appear exclusively in that newsgroup. I couldn't find a way to find those. So here are the top 10 words for each category.

```
-------------------------
alt.atheism
-------------------------
god : 765
one : 722
people : 576
writes : 562
subject : 542
line : 538
would : 501
organization : 472
dont : 458
atheist : 446
-------------------------
comp.graphics
-------------------------
line : 774
image : 758
subject : 627
organization : 581
file : 554
graphic : 460
university : 360
program : 329
would : 295
x : 279
-------------------------
comp.os.ms-windows.misc
-------------------------
maxaxaxaxaxaxaxaxaxaxaxaxaxaxax : 3317
window : 1082
line : 654
file : 641
subject : 618
organization : 579
driver : 375
university : 333
use : 323
problem : 318
```

```
-------------------------
comp.sys.ibm.pc.hardware
-------------------------
drive : 976
scsi : 704
line : 637
subject : 610
organization : 583
card : 478
system : 410
mb : 388
one : 379
disk : 353
-------------------------
comp.sys.mac.hardware
-------------------------
line : 652
subject : 591
organization : 552
mac : 546
apple : 420
problem : 372
drive : 357
one : 350
university : 308
nntppostinghost : 295
-------------------------
comp.windows.x
-------------------------
x : 5152
window : 933
line : 857
subject : 788
file : 778
organization : 599
program : 553
use : 508
widget : 503
server : 474
-------------------------
```

```
-------------------------
misc.forsale
-------------------------
line : 628
subject : 601
organization : 573
sale : 560
new : 329
university : 326
offer : 273
nntppostinghost : 263
distribution : 252
email : 246
-------------------------
rec.autos
-------------------------
car : 1223
line : 642
subject : 625
organization : 589
writes : 484
article : 453
would : 430
one : 362
like : 327
dont : 322
-------------------------
rec.motorcycles
-------------------------
bike : 688
line : 638
organization : 612
subject : 611
writes : 503
article : 473
dod : 455
one : 394
like : 334
nntppostinghost : 303
-------------------------
```

```
rec.sport.baseball
------------------------
line : 648
subject : 618
organization : 601
year : 592
game : 558
writes : 468
team : 432
article : 391
player : 364
run : 343
------------------------
rec.sport.hockey
------------------------
team : 954
game : 917
line : 707
subject : 635
organization : 611
hockey : 594
player : 522
play : 491
year : 438
would : 429
------------------------
sci.crypt
------------------------
key : 1449
encryption : 840
chip : 839
would : 707
clipper : 705
line : 693
system : 668
subject : 665
one : 642
organization : 621
------------------------
```

```
sci.electronics
------------------------
line : 782
subject : 672
organization : 581
one : 481
use : 371
would : 367
writes : 301
university : 291
like : 275
nntppostinghost : 268
------------------------
sci.med
------------------------
subject : 649
line : 619
organization : 610
one : 567
article : 462
writes : 439
would : 419
people : 322
msg : 312
dont : 311
------------------------
sci.space
------------------------
space : 1200
line : 646
subject : 635
organization : 631
would : 553
writes : 452
one : 413
nasa : 407
article : 402
launch : 377
------------------------
```

```
soc.religion.christian
------------------------
god : 1477
one : 817
would : 775
christian : 755
subject : 671
people : 656
line : 636
jesus : 618
organization : 559
say : 520
------------------------
talk.politics.guns
------------------------
gun : 1231
would : 819
people : 657
line : 608
subject : 584
organization : 555
one : 531
writes : 507
article : 492
right : 487
------------------------
talk.politics.mideast
------------------------
armenian : 1268
people : 996
one : 883
israel : 879
israeli : 791
turkish : 712
would : 711
subject : 661
jew : 630
line : 626
------------------------
```

```
------------------------
talk.politics.misc
------------------------
would : 692
people : 684
writes : 611
q : 575
article : 573
line : 525
one : 515
dont : 508
organization : 506
subject : 500
------------------------
talk.religion.misc
------------------------
god : 516
one : 463
subject : 418
line : 418
people : 410
organization : 403
christian : 401
jesus : 390
would : 387
writes : 357
```

Q6) There are some ways to increase the accuracy of the classifier. One of them is with using weights. One of the most popular weight implementation in text classification is "term frequency- inverse tern frequency." It assumes that, how important a word is inversely proportional to how often it occurs across all documents.

Another way to increase the accuracy is removing common words. These words do not have any meaning for us because it appears in all the categories. For example one, would, writes, etc.