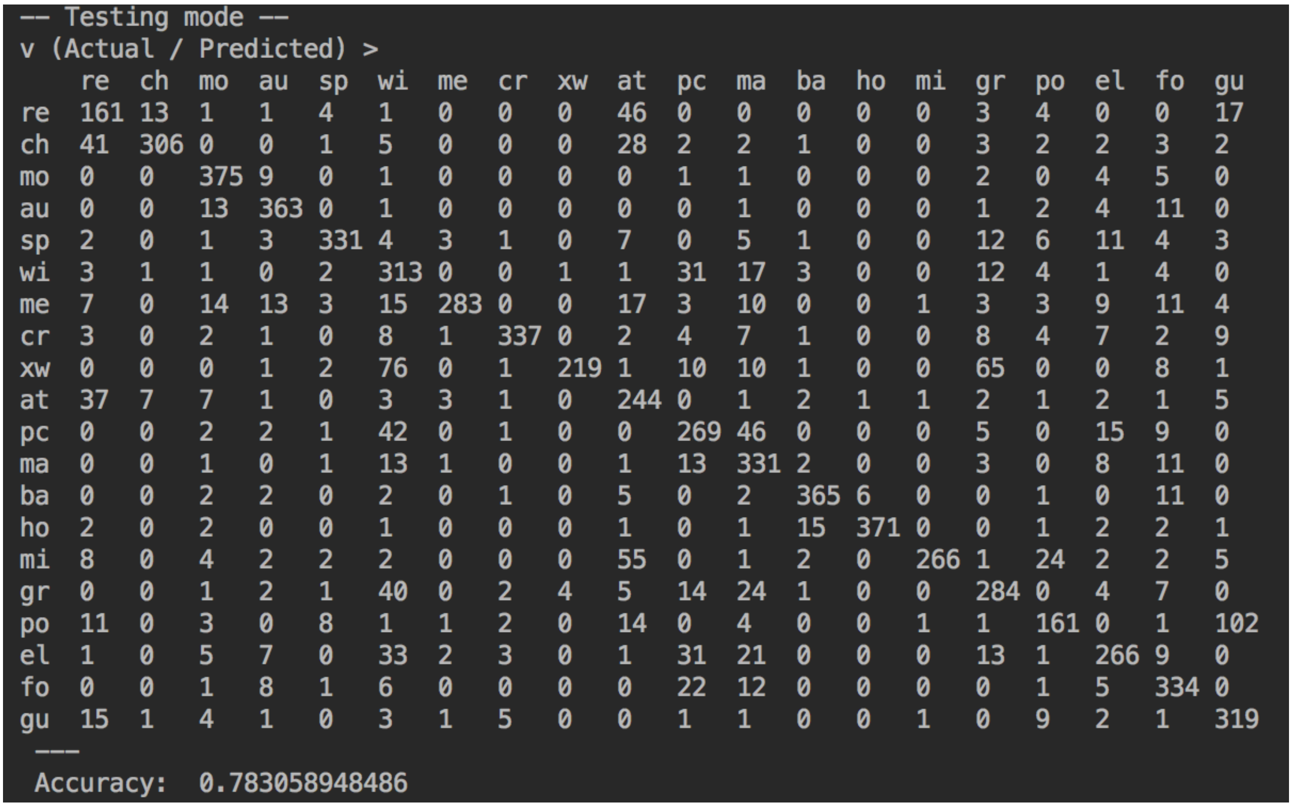
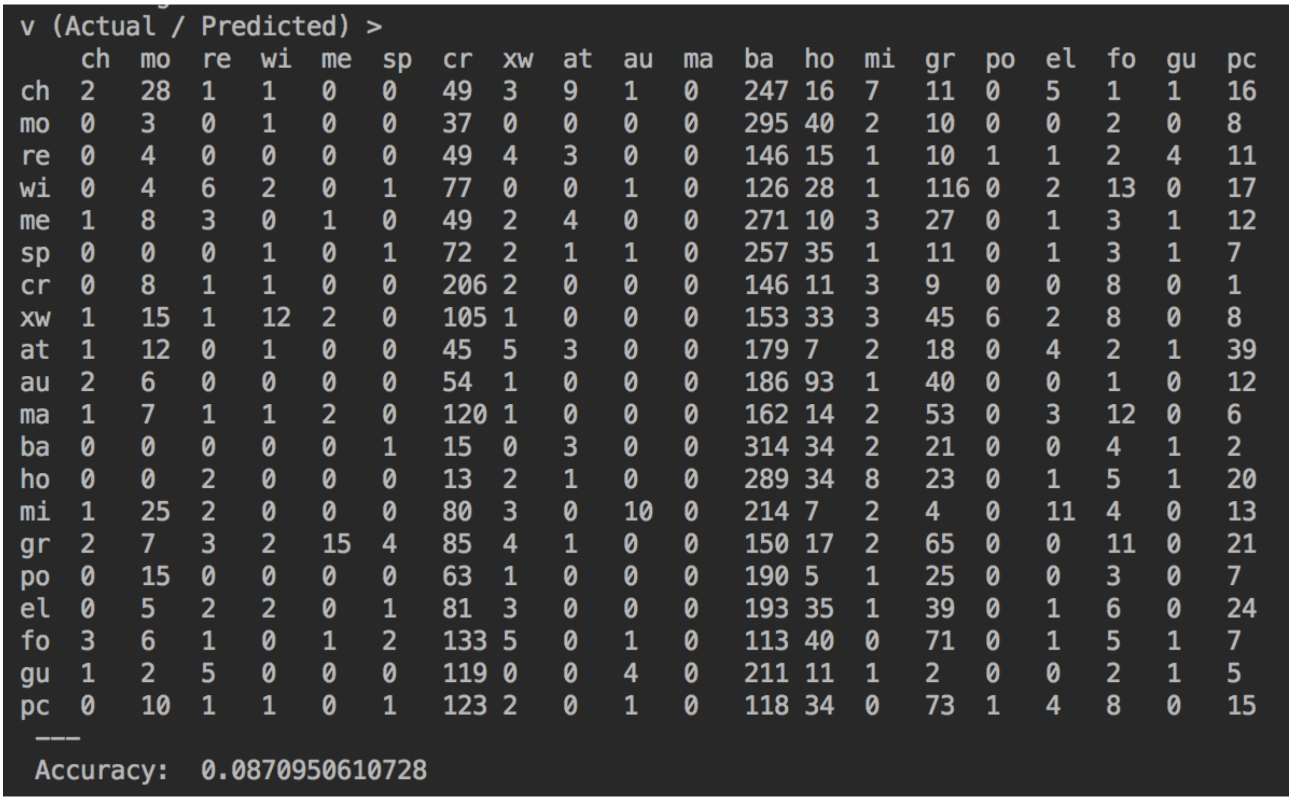
AI - Assignment 4 Part-2 Report

1. Training:
    1. Training data will hold doc_dictionary which holds details like:
        1. Total Words
        2. Words list
        3. Topics and their probability
        4. If it's labelled or not
    2. Topic Word Table is represented in terms of a dictionary:
        1. This contains list of words & frequencies in each topic
2. Testing:
    1. Picks up from a model file of the above details and applies bayes formula.
3. Accuracy:
    1. Supervised Learning (Fraction: 1.0)



```
-- Testing mode --
v (Actual / Predicted) >
    re  ch  mo  au  sp  wi  me  cr  xw  at  pc  ma  ba  ho  mi  gr  po  el  fo  gu
re  161 13  1   1   4   1   0   0   0   46  0   0   0   0   0   3   4   0   0   17
ch  41  306 0   0   1   5   0   0   0   28  2   2   1   0   0   3   2   2   3   2
mo  0   0   375 9   0   1   0   0   0   0   1   1   0   0   0   2   0   4   5   0
au  0   0   13  363 0   1   0   0   0   0   0   1   0   0   0   1   2   4   11  0
sp  2   0   1   3   331 4   3   1   0   7   0   5   1   0   0   12  6   11  4   3
wi  3   1   1   0   2   313 0   0   1   1   31  17  3   0   0   12  4   1   4   0
me  7   0   14  13  3   15  283 0   0   17  3   10  0   0   1   3   3   9   11  4
cr  3   0   2   1   0   8   1   337 0   2   4   7   1   0   0   8   4   7   2   9
xw  0   0   0   1   2   76  0   1   219 1   10  10  1   0   0   65  0   0   8   1
at  37  7   7   1   0   3   3   1   0   244 0   1   2   1   1   2   1   2   1   5
pc  0   0   2   2   1   42  0   1   0   0   269 46  0   0   0   5   0   15  9   0
ma  0   0   1   0   1   13  1   0   0   1   13  331 2   0   0   3   0   8   11  0
ba  0   0   2   2   0   2   0   1   0   5   0   2   365 6   0   0   1   0   11  0
ho  2   0   2   0   0   1   0   0   0   1   0   1   15  371 0   0   1   2   2   1
mi  8   0   4   2   2   2   0   0   0   55  0   1   2   0   266 1   24  2   2   5
gr  0   0   1   2   1   40  0   2   4   5   14  24  1   0   0   284 0   4   7   0
po  11  0   3   0   8   1   1   2   0   14  0   4   0   0   1   1   161 0   1   102
el  1   0   5   7   0   33  2   3   0   1   31  21  0   0   0   13  1   266 9   0
fo  0   0   1   8   1   6   0   0   0   0   22  12  0   0   0   0   1   5   334 0
gu  15  1   4   1   0   3   1   5   0   0   1   1   0   0   1   0   9   2   1   319
 ---
 Accuracy:  0.783058948486
```

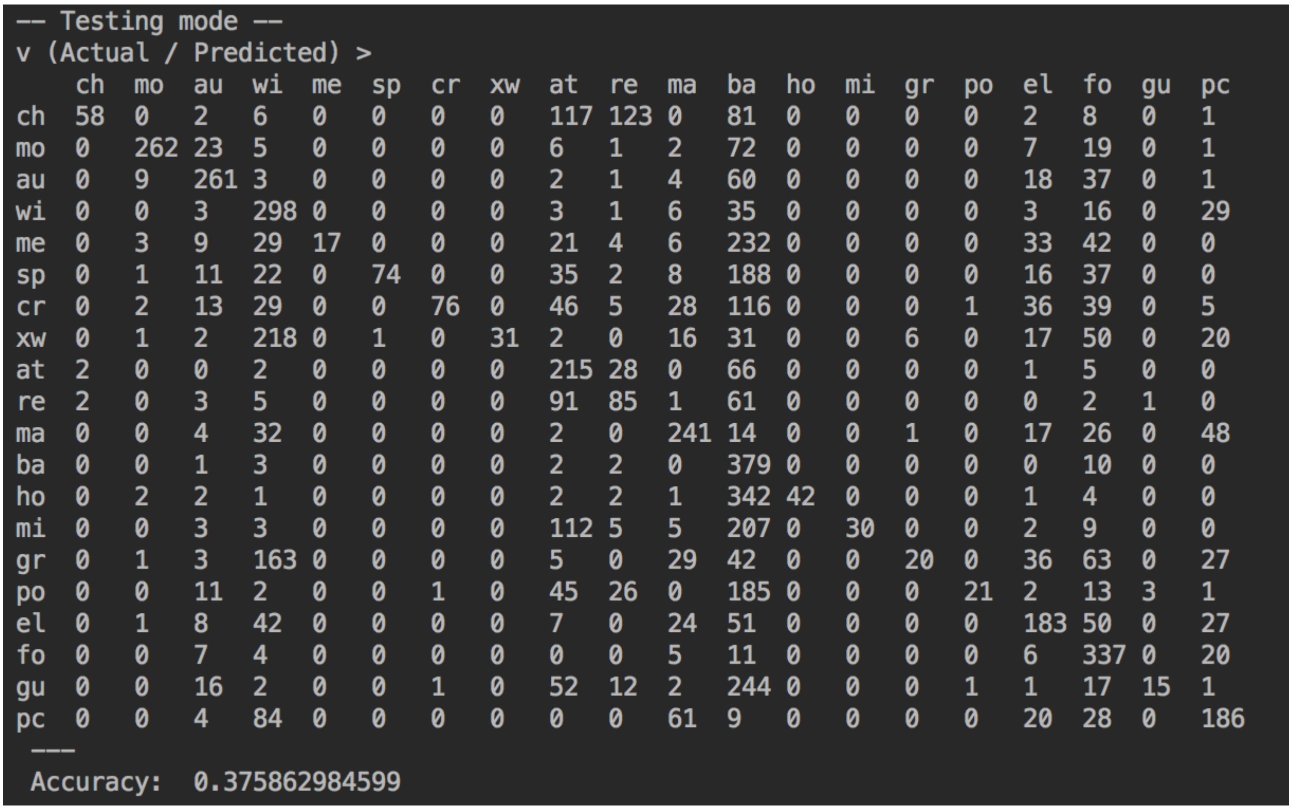
2. Unsupervised Learning (Fraction: 0.0)



```
v (Actual / Predicted) >
    ch  mo  re  wi  me  sp  cr  xw  at  au  ma  ba  ho  mi  gr  po  el  fo  gu  pc
ch  2   28  1   1   0   0   49  3   9   1   0   247 16  7   11  0   5   1   1   16
mo  0   3   0   1   0   0   37  0   0   0   0   295 40  2   10  0   0   2   0   8
re  0   4   0   0   0   0   49  4   3   0   0   146 15  1   10  1   1   2   4   11
wi  0   4   6   2   0   1   77  0   0   1   0   126 28  1   116 0   2   13  0   17
me  1   8   3   0   1   0   49  2   4   0   0   271 10  3   27  0   1   3   1   12
sp  0   0   0   1   0   1   72  2   1   1   0   257 35  1   11  0   1   3   1   7
cr  0   8   1   1   0   0   206 2   0   0   0   146 11  3   9   0   0   8   0   1
xw  1   15  1   12  2   0   105 1   0   0   0   153 33  3   45  6   2   8   0   8
at  1   12  0   1   0   0   45  5   3   0   0   179 7   2   18  0   4   2   1   39
au  2   6   0   0   0   0   54  1   0   0   0   186 93  1   40  0   0   1   0   12
ma  1   7   1   1   2   0   120 1   0   0   0   162 14  2   53  0   3   12  0   6
ba  0   0   0   0   0   1   15  0   3   0   0   314 34  2   21  0   0   4   1   2
ho  0   0   2   0   0   0   13  2   1   0   0   289 34  8   23  0   1   5   1   20
mi  1   25  2   0   0   0   80  3   0   10  0   214 7   2   4   0   11  4   0   13
gr  2   7   3   2   15  4   85  4   1   0   0   150 17  2   65  0   0   11  0   21
po  0   15  0   0   0   0   63  1   0   0   0   190 5   1   25  0   0   3   0   7
el  0   5   2   2   0   1   81  3   0   0   0   193 35  1   39  0   1   6   0   24
fo  3   6   1   0   1   2   133 5   0   1   0   113 40  0   71  0   1   5   1   7
gu  1   2   5   0   0   0   119 0   0   4   0   211 11  1   2   0   0   2   1   5
pc  0   10  1   1   0   1   123 2   0   1   0   118 34  0   73  1   4   8   0   15
 ---
 Accuracy:  0.0870950610728
```

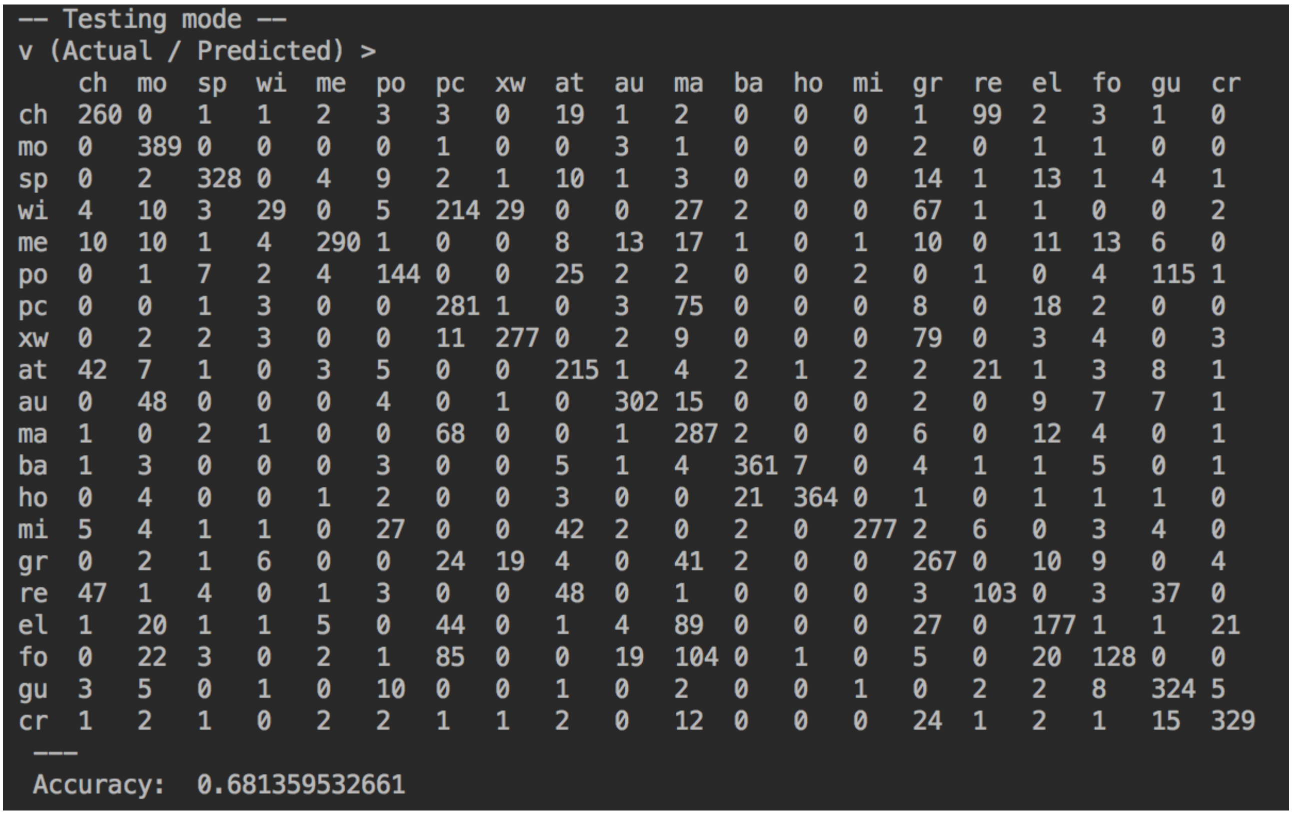
3. Semi-Supervised Learning:

Fraction 0.1 — Retrained 0 times:



```
-- Testing mode --
v (Actual / Predicted) >
    ch  mo  au  wi  me  sp  cr  xw  at  re  ma  ba  ho  mi  gr  po  el  fo  gu  pc
ch  58  0   2   6   0   0   0   0   117 123 0   81  0   0   0   0   2   8   0   1
mo  0   262 23  5   0   0   0   0   6   1   2   72  0   0   0   0   7   19  0   1
au  0   9   261 3   0   0   0   0   2   1   4   60  0   0   0   0   18  37  0   1
wi  0   0   3   298 0   0   0   0   3   1   6   35  0   0   0   0   3   16  0   29
me  0   3   9   29  17  0   0   0   21  4   6   232 0   0   0   0   33  42  0   0
sp  0   1   11  22  0   74  0   0   35  2   8   188 0   0   0   0   16  37  0   0
cr  0   2   13  29  0   0   76  0   46  5   28  116 0   0   0   1   36  39  0   5
xw  0   1   2   218 0   1   0   31  2   0   16  31  0   0   6   0   17  50  0   20
at  2   0   0   2   0   0   0   0   215 28  0   66  0   0   0   0   1   5   0   0
re  2   0   3   5   0   0   0   0   91  85  1   61  0   0   0   0   0   2   1   0
ma  0   0   4   32  0   0   0   0   2   0   241 14  0   0   1   0   17  26  0   48
ba  0   0   1   3   0   0   0   0   2   2   0   379 0   0   0   0   0   10  0   0
ho  0   2   2   1   0   0   0   0   2   2   1   342 42  0   0   0   1   4   0   0
mi  0   0   3   3   0   0   0   0   112 5   5   207 0   30  0   0   2   9   0   0
gr  0   1   3   163 0   0   0   0   5   0   29  42  0   0   20  0   36  63  0   27
po  0   0   11  2   0   0   1   0   45  26  0   185 0   0   0   21  2   13  3   1
el  0   1   8   42  0   0   0   0   7   0   24  51  0   0   0   0   183 50  0   27
fo  0   0   7   4   0   0   0   0   0   0   5   11  0   0   0   0   6   337 0   20
gu  0   0   16  2   0   0   1   0   52  12  2   244 0   0   0   1   1   17  15  1
pc  0   0   4   84  0   0   0   0   0   0   61  9   0   0   0   0   20  28  0   186
 ---
 Accuracy:  0.375862984599
```

Fraction 0.1 — Retrained 5 times:



```
-- Testing mode --
v (Actual / Predicted) >
    ch  mo  sp  wi  me  po  pc  xw  at  au  ma  ba  ho  mi  gr  re  el  fo  gu  cr
ch  260 0   1   1   2   3   3   0   19  1   2   0   0   0   1   99  2   3   1   0
mo  0   389 0   0   0   0   1   0   0   3   1   0   0   0   2   0   1   1   0   0
sp  0   2   328 0   4   9   2   1   10  1   3   0   0   0   14  1   13  1   4   1
wi  4   10  3   29  0   5   214 29  0   0   27  2   0   0   67  1   1   0   0   2
me  10  10  1   4   290 1   0   0   8   13  17  1   0   1   10  0   11  13  6   0
po  0   1   7   2   4   144 0   0   25  2   2   0   0   2   0   1   0   4   115 1
pc  0   0   1   3   0   0   281 1   0   3   75  0   0   0   8   0   18  2   0   0
xw  0   2   2   3   0   0   11  277 0   2   9   0   0   0   79  0   3   4   0   3
at  42  7   1   0   3   5   0   0   215 1   4   2   1   2   2   21  1   3   8   1
au  0   48  0   0   0   4   0   1   0   302 15  0   0   0   2   0   9   7   7   1
ma  1   0   2   1   0   0   68  0   0   1   287 2   0   0   6   0   12  4   0   1
ba  1   3   0   0   0   3   0   0   5   1   4   361 7   0   4   1   1   5   0   1
ho  0   4   0   0   1   2   0   0   3   0   0   21  364 0   1   0   1   1   1   0
mi  5   4   1   1   0   27  0   0   42  2   0   2   0   277 2   6   0   3   4   0
gr  0   2   1   6   0   0   24  19  4   0   41  2   0   0   267 0   10  9   0   4
re  47  1   4   0   1   3   0   0   48  0   1   0   0   0   3   103 0   3   37  0
el  1   20  1   1   5   0   44  0   1   4   89  0   0   0   27  0   177 1   1   21
fo  0   22  3   0   2   1   85  0   0   19  104 0   1   0   5   0   20  128 0   0
gu  3   5   0   1   0   10  0   0   1   0   2   0   0   1   0   2   2   8   324 5
cr  1   2   1   0   2   2   1   1   2   0   12  0   0   0   24  1   2   1   15  329
 ---
 Accuracy:  0.681359532661
```

Formula goes something like:
P (Topic I Doc) = P (Topic) . P (Doc I Topic) / P (Doc)

Here, P (Doc) is a constant where document for each will be the same. P (Topic) will be occurrence of topicA in documents by the total number of documents. With P (Doc I Topic) it will be words in that document. Hence,

P (Doc I Topic1) = P (Word-1 I Topic1) . P (Word-2 I Topic1)………….P (Word-n I Topic1)