Dallas Delaney - ddelane2

Doug Zhu - dzhu10

Erik Delanois – delanos2

3 Credits

**A.I. MP3 Report**

**Problem 1**

IMPLEMENTATION

In this problem we first read in all of the training data in order to build our P(Fij | class). Once we were able to build this probability for every pixel in every numeric character, we were able to calculate our P(class) based on the frequency of occurrence of each character. Once all of this was stored in a database we were able to calculate P(class) ⋅ P(f1,1 | class) ⋅ P(f1,2 | class) ⋅ ... ⋅ P(f28,28 | class) with the logarithmic adjustment for each image in the test data for each class. The class that returned the highest value would be our estimate. We then saved the most and least prototypical instances of each class by selecting the highest and lowest P(class) ⋅ P(f1,1 | class) ⋅ P(f1,2 | class) ⋅ ... ⋅ P(f28,28 | class) values. After this we were able to calculate the percentage of the test images that were classified properly along with the confusion matrix. We were able to achieve 77.1% correct classification of the images so we are confident in our design. An interesting fact about our approach is that it is very dynamic and flexible. You can make this program run on images with any dimensions you want, with as many characters as you want, and with as many training/test cases as you want and it will calculate all the same output data. This is because we defined everything in terms of constants located at the beginning of the DigitClassification.h file so to change any of the attributes mentioned above you simply need to change the corresponding constant to the appropriate value. This made doing the facial classification much simpler as we only had to edit the constant values.

Smoothing Value = 1

Character Correct Attempts Percentage

0 76 90 84.4444%

1 104 108 96.2963%

2 80 103 77.6699%

3 79 100 79%

4 83 107 77.5701%

5 62 92 67.3913%

6 69 91 75.8242%

7 77 106 72.6415%

8 61 103 59.2233%

9 80 100 80%

Total 771 1000 77.1%

Confusion Matrix:

0 1 2 3 4 5 6 7 8 9

0 84.444 0.000 1.111 0.000 1.111 5.556 3.333 0.000 4.444 0.000

1 0.000 96.296 0.926 0.000 0.000 1.852 0.926 0.000 0.000 0.000

2 0.971 2.913 77.670 3.883 0.971 0.000 5.825 0.971 4.854 1.942

3 0.000 2.000 0.000 79.000 0.000 3.000 2.000 6.000 2.000 6.000

4 0.000 0.935 0.000 0.000 77.570 0.000 2.804 0.935 1.869 15.888

5 2.174 2.174 1.087 13.043 3.261 67.391 1.087 1.087 2.174 6.522

6 1.099 5.495 4.396 0.000 4.396 6.593 75.824 0.000 2.198 0.000

7 0.000 5.660 2.830 0.000 2.830 0.000 0.000 72.642 2.830 13.208

8 0.971 0.971 2.913 13.592 1.942 7.767 0.000 0.971 59.223 11.650

9 1.000 1.000 1.000 3.000 9.000 2.000 0.000 2.000 1.000 80.000

For Character '0' this is the most prototypical image:

+#++

+#####+

+########+

+##########+

+######+ +##+

+####+++ +###+

####+ +##+

+###+ ##+

####+ ##+

###+ +##

###+ ##+

###+ ##+

###+ +##+

###+ +##+

####+ +###+

+###++ +####+

#####++#####+

+##########+

+########+

++++++

For Character '0' this is the least prototypical image:

++++++#+++

+##########+

+##############

+###############+

########++ ++###+

#######+ ####

+###+ +###+

+### +###+

#### ###+

#### ####

###+ ####

#### ####

####+ ####

+####+ ####

+#### ####

+###++ ####

+#####++++#####

+############+

+##########+

+++#######

For Character '1' this is the most prototypical image:

+#+

+##+

+##+

###+

###+

###

+##+

+##+

+##+

###+

###

+##+

+##+

###+

+###

+##+

+###+

+####

+###+

+#+

For Character '1' this is the least prototypical image:

+#

##+

+#++

++#+

+#+#+

+#+++

##+#+

+#++#+

+#+ ++

+#+ +#+

+#+ ++ +

+##+#########++

++##++++++++#+

#+

+#+

+#

+#

#+

#+

#+

For Character '2' this is the most prototypical image:

++++++

######++

+########+

+#++ +###+

++ +##+

+##

+##

+#

+#

+#

+#+

+#+

+++++ ##+

+######+##

+#########+

+##++++#####+

##+ +###+####++

##++####+ +##+

+#####++

+##++

For Character '2' this is the least prototypical image:

+++#+

+#####+

+##+++

##+

+##

+#+

+#+

##+

+#+

+#+

##

##+ +++

+## +#####+

+#+ +###++##+

##+ ### +##

+##+ +##+ +#+

++##+ ##+ +#+

++#####+ +##

+##########

+++++++

For Character '3' this is the most prototypical image:

+#####++

+########+

+++####+

++##

+##+

+##+

++##+

+###+

+###+

+#####+

###+###+

++ +##+

+#+

+##

+##

+##+

++##+

+++++++++###+

+##########+

+++#####++

For Character '3' this is the least prototypical image:

+++++++++#+

++############

+############++

+##++++++++++

##+

+#+

+##+

+##++++++++++++

+##############+

+##############+

+++++++ +++ +##+

+##+

+##+

+##+

++ +##+

#+ +##

##+ +##+

+##+++++++++###+

+###########+

++#######+

For Character '4' this is the most prototypical image:

+ +#

+#+ +#

##+ +#

+##+ +++#

+### +##+

+##+ +##+

### +##+

##+ +###+

#+ ++####+

++####+######

+++#######++

+++++##+

+#+

+#+

##+

+##+

+#+

+#+

+#+

+#+

For Character '4' this is the least prototypical image:

+##++++

++#########+

++############+

++###########+####

+#######++#+++ ++##

+####+ +#+

####+ ##+

###+ ++##+

###+ +++####+

#####++++##########+

+###################

+##########++ +###

++++++++ +###

+###

+##+

+##+

+###

+##+

+##+

+##+

For Character '5' this is the most prototypical image:

++#+

++++++####+

#########+

###+++++

+#++

+#+

+##

+#++++

+######+

+##+++##

+ +#+

+#+

+#+

##

+#+ ##+

+#+ +##

##+ ###

+##+ +##+

##+++###+

###++

For Character '5' this is the least prototypical image:

++############

++++#############+

+##+######++++++

+#####++

+###++

+###++++++

#########++

+++##########++

+++++#######++

+++######+

++ ++####

##+ ###+

###+ +###+

+###++ +####+

+####++ +++####+

++#############+

++#########++

+++###+++

For Character '6' this is the most prototypical image:

+#+

+###+

+###+

+###+

+###+

+###+

+###+

+##+

+##+

+###+

+##+

+##+ ++##+

+###+ +####+

+###++######+

+##+ +######+

+###+#######+

+#########+

+#######+

+#####+

+##+

For Character '6' this is the least prototypical image:

#+

+##+

+##+

+##+

+##+

+###

+###

+###

+###

+##+ +++++++

+##+ +#######+

+#+ +####+###++

+## +##++ +##+

+##+ +###+ +##+

##+ +##+ +##+

+###++##+ +###+

#######+ ++###+

++######+####+

++######++

+++++

For Character '7' this is the most prototypical image:

+#++ ++++++

###########+

++###+++++#+

+++ +#+

+##+

+##+

+##+

+##+

+##+

+##+

+##+

##+

+##+

##+

+##+

+#+

+##

+##

+##

+++

For Character '7' this is the least prototypical image:

####++##+

###########+

############+

+############+

+++#########+

++++++####++

++#####+

++#####

++####

++###+

+####+

+#####

+#####

+####+

+####+

+####+

+###

+###+

+####

+#####

For Character '8' this is the most prototypical image:

+##+

++####+

++######+

+####++##+

++###+ +##+

+##+ +##+

### +#++

### ++###

+##++++###++

+#######++

#####++

+#####

+######+

###++##+

+##+ ###+

+##+ ###+

+##+ +###

+#######+

+######+

++##++

For Character '8' this is the least prototypical image:

+++#####+

+####+++###++++

+####++ ++###+

+###+ ###+

+## ###

+## ###

+### +++++###+

+###+++########+

+######++++++++

+#### +

+#####+

+#+++##++

##+ +###+

##+ +###+

+#++ +###+

###++ ++###+

+####+ +###+

+####+++###+

++#######+

+++###++

For Character '9' this is the most prototypical image:

++##++

+######+

+####++++#+

+###+ ++##

+##+ +####

+##+ +####

+##+ +###+

+##+ ++####

+#########+

+#######+

+++++##+

+##+

+##+

+#+

+##+

+##+

+##

+#+

+#+

+#+

For Character '9' this is the least prototypical image:

++++####++

++###########++

+###++++++++####+

##+ +###+

+##+ +####

+###++ +#####+

+#####++++##++++

++########+

++######++

+#######+

+#+ +###+

##+ +###+

+##+ +###+

##+ ##+

+## ###+

+##+ +##

+##+ +#+

+###++++##+

+#######+

+##+++

This is the odds ratio for 0/5 :

+++++++++ -----

--++++++++++ -----

----++++++++++ ------

+ --- ++++++++++ -------

+--- ++++++++++++ ------

- ++++++++++++++ -----

-++++++++ +++++++++----

-+++++++ - ++++++++--

-+++++ ----- +++++++-

++++++ -------+++++++++

+ ++++++---------++++++++

++++++---------+++++++

-++++++---------++++++

-+++++++------- ++++++

-+++++++-------+++++++-

- ++++++++ ----++++++++

--++++++++ +++++++++

--++++++++++++++++++++-++

+++++++++++++++++++--++

+++++++++++++++ + ---++

++++++++++ --------

-------------------

---------------

-------------

This is the odds ratio for 5/3 :

++++++++++++++++++++++++++++

++++++++++++++++++++++++++++

++++++++++++++++++++++++++++

+++++--------------+++++++++

+++----------------+++++++++

+++--------------- +++++++++

+++--------------- +++++++++

+-+---------- ----- ++++++++

+++------- +++++ --- +++++++

++------+++++++ -----+++++++

++---- +++++++ -------++++++

+++-+++++++++ ---------+++++

++++++++++++ ----------+++++

+++-++++++++ ------- ++++++

++++++++++++ --------+ +++++

++++++++++++ -------++++-+

++--++++++++++++------ +++++

++--- ++++++++++ ------+++++

++--- ++++++++++ -----+++++

++---- +++++++++ ------+++++

++---- +++++++ -------+++++

++----- ---------+++++

++-------- ----------- +++++

++------------------ +++++++

+++----------- +++++++++++

+++------ +++ ++++ ----+++++

++++++++++++++--+++---++++++

++++++++++++++++++++++++++++

This is the odds ratio for 8/3 :

-----------------

-----------------+++++

------------ +++++++++++

---------- ++++++++++

- -------- + ++++++++

------++++++ -- ++++++++

----- ++++++ ----++++++++

-- +++++++++---- ++++++++

-+++++++++ --- ++++++++++

++++++++ --- +++++++++++

-+++++ - ++++++

-+++--- +++++------+++

--- -- +++++++------ + -

----- +++++++++-------+

-----++++++++++-------+

-----++++++++++-------+

----- ++++++++ -------+

------+++++ ---------+

------- ---------++

-------- --- ---+++

-------- ++ ++++

------ ++++++++++++ -

-----------++++++ --

----------- ++

This is the odds ratio for 8/9 :

+++++++++++

+ ++++++++++++++++++

+++++++++++++++++++++

+ ++++++++++++++++++++++

++++++++ --- ++++++++++

+++++ -------- ++++++++

+ + ---------- +++++++

------- +++ - +++++++

------- ++++ ++++++++

--------+++++ --- ++++++

--------+++++ -----++++-

-------- ++++ ----- ++ -

-------- +++ -------++--

------- ++++ ---- ++++--

-- -- ++++++ ---- +++++

--++++++++++--- ++++++

-+++++++++ ---++++++++

-+++++++++ ++++++++

++++++++++++++++++++-

+++++++++++++++++ ++-

-+++++++++++++ -----

-++++++++++ --------

---------------------

--------------------

----- ---

**Problem 2**

IMPLEMENTATION

This part of the MP was done in Python.

In order to implement problem 2.1, we first had make sure we had all of our word likelihoods saved somewhere. In order to do this, we saved the likelihoods of each word from the training data into a dictionary based on whichever label it had. Then we went through each document of the test data and classified it into one label. We had to go through each line and find the sum of the likelihood that the document would fit a certain label. For multinomial naive Bayes we calculated the sum of every word’s likelihood and frequency for each label. The email was then classified as whichever likelihood was greater, and the confusion matrix was updated. To implement the Bernoulli Naive Bayes problem, we commented out the line where the likelihood accounted for the word frequency. The classification rate for both models and both datasheets were the same (email - 97%, movie - 75%). To have our code work for the movie datasheet, we just had to change the way we searched for labels by changing a number in our conditional statement (labels were -1 and 1 instead of 0 and 1). To find the classification rates we divided the number of documents we labeled correctly by the total number of documents.

CONFUSION MATRIX

**Bernoulli:**

|  |  |  |
| --- | --- | --- |
|  | Not Spam | Spam |
| Not Spam | 126 | 4 |
| Spam | 2 | 128 |

Classification Rate: 97.6%

|  |  |  |
| --- | --- | --- |
|  | Good Review | Bad Review |
| Good Review | 382 | 118 |
| Bad Review | 126 | 374 |

Classification Rate: 75.6%

**Multinomial:**

|  |  |  |
| --- | --- | --- |
|  | Not Spam | Spam |
| Not Spam | 126 | 4 |
| Spam | 2 | 128 |

Classification Rate: 97.6%

|  |  |  |
| --- | --- | --- |
|  | Good Review | Bad Review |
| Good Review | 383 | 117 |
| Bad Review | 124 | 376 |

Classification Rate: 75.9%

The top 20 for words for the emails are listed below. They are in the format word\_(# of occurrences)

TOP 20 WORDS FOR EMAILS

TOP SPAM WORDS

email\_1380

s\_1207

order\_1159

report\_1053

our\_965

address\_954

mail\_923

program\_828

send\_800

free\_744

money\_722

list\_713

receive\_662

name\_627

business\_608

one\_553

d\_541

work\_528

com\_524

TOP NORMAL WORDS

language\_1130

university\_906

s\_661

linguistic\_477

de\_445

information\_444

conference\_378

workshop\_360

email\_321

paper\_320

e\_314

english\_312

one\_280

please\_278

include\_277

edu\_271

http\_264

research\_259

abstract\_253

TOP 20 WORDS FOR MOVIES

The top 20 for words for the movies are listed below. They are in the format word\_(# of occurrences)

TOP NEGATIVE REVIEW WORDS

movie\_290

film\_227

like\_163

one\_143

--\_114

bad\_87

story\_85

much\_83

time\_75

even\_70

characters\_64

good\_64

little\_62

would\_58

comedy\_57

never\_53

nothing\_52

makes\_51

plot\_51

TOP POSITIVE REVIEW WORDS

film\_285

movie\_187

--\_136

one\_111

like\_99

story\_94

good\_84

comedy\_83

way\_80

even\_76

time\_73

best\_72

much\_66

performances\_62

funny\_60

make\_60

life\_58

us\_58

makes\_58

EXTRA CREDIT

Problem 1 extra credit: Facial Recognition:

\*see description of problem 1 for details as it has the same implementation.

Smoothing Value = 10

Character Correct Attempts Percentage

0 54 77 70.1299%

1 72 73 98.6301%

Total 126 150 84%

Confusion Matrix:

1. 1

0 70.130 29.870

1 1.370 98.630

For Character '0' this is the most prototypical image:

#### #

### ###########

####

## # ##

# # # # #

## ### #

##

# #

# #####

# # ##

## # #

# # #

# #

## ### ##

##### # # # # ####

# # # # # ## #

### ## ### ## # # #

# # # #

# ######

### ##

#

#### # #

# ## #####

# #

# #

# #

## ###

# # ## ## ##

## # # ## #

# # #

### ## # # ###

#### # ## # # # # #

# # # # # # ## #

# # # # # ####

####### # # #####

## # #

# #

# #

#### ## # #

# # # # #

# # # # #

# ## # # #

# # ## # #

# ### # #

# # # #

### # # ##

# # #

# #

# #

## ##

# #

# #

#

##

# # ##

## # # # #

# # # # #### # #

# # # # ## # # # #

## # # # # # ## # #####

# # # # # # # #

# ## ## # #

### ## # ### ##

# # # # ## #

## ### ## ##

# #####

# #

For Character '0' this is the least prototypical image:

######## ################################################

##

#

# # # # # # #

# # ## # # #### # # #

## # # ## # # # # ######## #

# # ## # # # # # # # #

# # # # # ### # ## # #

# # # ## # # ## # # ## ## #### # #

# # # # # # ## # # # ## # #

# # # # # # # # # ### #

# ### # # # # #### # #

# # # # # ###### # #

# # # # # # #

# # # # # # #

# # # # ### ### # #

# # # ## # ####### # # # # # #

# # # ## # # # # # ## # #

# # # # # # # # # # # #

# # # # # # # ####### # # # #

# ## ##### #### # # ## #### ## #

# ## # # # # #

# ## ### # # # # #

# ## #### # ## ## # ## #

# # # # # # # # # # #

# # # # # ### # ## # # # # #

# # # # ## # ## # # ###

# ## ## # # # # # # #

# # ### # # # # # # ## ### #

# # ######## # # # ## ## #

# # # # # # # # #

# # ### # # # # # ## ### #

# ## # # # # # # #

# # # # # # # ######

# # #### # # # # # # #

# # ## ## # # # # #

# # # # # ### # # # #

# # # # # # #

# ## # # # # # # #

# # # # ## # ### # #

# # # #### # ### # # # ##

# # # # # # # ## #

# # # # # ### # ## # #

# # # # ## ### # ##

# # # # # ## ######## #

# ## #### # #

# # # # #

## # # ## ###

# # # # # ####### #

#### # # # # # #

# # # #

# ## # ######

# # # # #

# ### # ### ### ### #

# # # # # ## ## ## #

# # # # # # # ## ########

# # # # ## # # # ##

#### # ## # # ### # ## # #

# # # # # # # # #

# # # ### # # #

# # # # # ### #

# # # # # #

# ## # #### # # ##

###### ### ####### # # ##

# # # # #

# #

#

### ###### ########################################

For Character '1' this is the most prototypical image:

# #

# #

# #

# ##

# #

# ##

# #

#### #

## ##

# #

# #

## ## #

# # ##

# ## #

# # #

## # ##

## # ##

## ##

# #

# #

## #

# ##

# ##

# #

## #

# #

## ##

# ##

# #

# #

## #

# #

## #

# #

#####

For Character '1' this is the least prototypical image:

#### #### ###

######################### ######### ##

# # #

# # #

# # #

# # ####### #

# ## # ## # # #

# ## ## # # # # # #

# # # # # # # # # #

# # # # ## # # #

# # # # # # # # #

# # # # # # # # # #

# # # # # # # # #

# ## ## # # # # # #

# # # ## ##### # # #

# # #### # # #

# # # # #

# # # # #

# # # # #

# # ### # # #

# # ## # # #

# # # # # # #

# # # ## #

# # # # #

# # ###### # #

# # # # # #

# # # # # # #

# # #### ## # # # # # #

# # # ### ## # # # # # #

# #### # # # # ## #

# #### # # # # #

# ### # # # # # #

# # # # # # # #

# # ####### # # # ### #

# # ## ## # # # # #

# # # # # # # ### #

# # # # ## #

# # # ######## ####### ####### #

# # # ####### ### #### #

##### # # #

# # # #

# #### # #

# # # # #

# # # #

# ### ## ###### # #

# # # # ############## ### # #

# # # # # # #

# # # # # # # #

# # # # # # # #

# # # # # ################### # #

# # # # # # #

# # # # # #

# # # ###### # #

# # # # # #

# # # # # #

# # # ## # #

# # # #### # # #

# # # ######## # # #

# ### # # #

# # # #

# # # # #

############ ####

# # #

# ############################ #

# ############ ###### # #

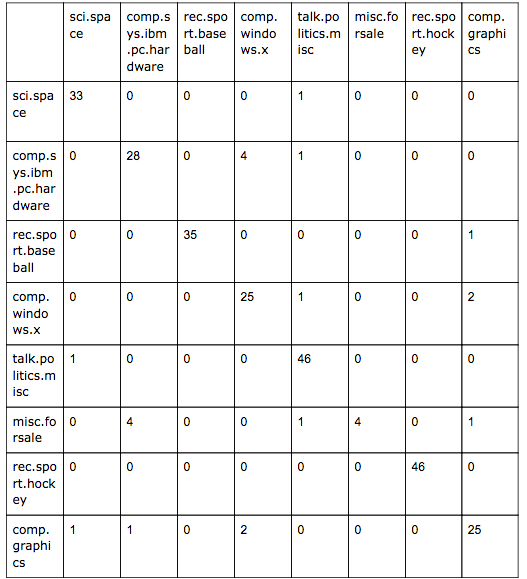
# # ### # # #

# # # # #

Problem 2 extra credit:

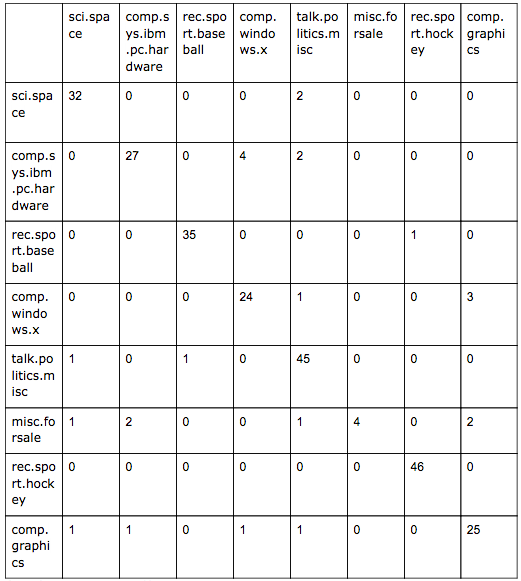
We have implemented problem 2.2. The only thing different about this problem was that there were many more labels. We added more dictionaries to store word frequencies and likelihoods for words in each label, but the way we updated the confusion matrixes and found the classification rates were the same.

**Bernoulli Confusion Matrix**



Classification Rate: 92.015%

**Multinomial Confusion Matrix**



Classification Rate: 90.49%

The top 20 for words for the newsgroups are listed below. They are in the format word\_(# of occurrences)

TOP 20 WORDS FOR EACH CATEGORY (2.2)

TOP sci\_space WORDS

space\_1030

nt\_593

would\_560

one\_384

launch\_352

nasa\_345

earth\_332

subject\_328

like\_304

us\_280

system\_278

also\_277

writes\_271

could\_263

first\_253

data\_253

time\_253

orbit\_251

edu\_251

TOP comp\_sys\_ibm\_pc\_hardware WORDS

drive\_496

scsi\_416

nt\_392

ide\_306

one\_262

card\_253

drives\_232

controller\_229

system\_216

disk\_216

subject\_205

use\_204

would\_203

edu\_198

hard\_191

bus\_189

get\_177

m\_176

data\_164

TOP rec\_sport\_baseball WORDS

nt\_936

would\_454

year\_427

edu\_416

writes\_355

one\_316

game\_316

good\_299

team\_294

subject\_293

last\_288

article\_287

think\_287

players\_275

like\_267

baseball\_255

games\_242

better\_240

well\_222

TOP comp\_windows\_x WORDS

x\_3598

window\_522

use\_455

nt\_433

subject\_426

file\_396

server\_363

also\_323

get\_312

available\_312

edu\_286

motif\_284

version\_277

system\_270

sun\_256

program\_256

c\_254

one\_252

m\_248

TOP talk\_politics\_misc WORDS

nt\_1400

would\_951

people\_831

q\_692

one\_601

mr\_599

think\_571

writes\_552

president\_552

article\_509

government\_497

stephanopoulos\_452

know\_448

us\_417

edu\_413

like\_404

subject\_378

going\_372

get\_331

TOP misc\_forsale WORDS

new\_218

edu\_189

dos\_160

sale\_145

appears\_144

art\_135

subject\_132

wolverine\_128

shipping\_117

cover\_115

price\_115

one\_112

list\_112

comics\_109

drive\_108

nt\_105

hulk\_104

good\_101

vs\_98

TOP rec\_sport\_hockey WORDS

nt\_838

game\_653

team\_635

hockey\_564

would\_402

play\_374

subject\_339

period\_338

season\_333

nhl\_327

games\_322

one\_312

first\_290

year\_283

think\_276

players\_271

get\_262

la\_262

edu\_259

TOP comp\_graphics WORDS

image\_889

jpeg\_468

edu\_438

nt\_419

file\_406

images\_389

data\_369

also\_365

graphics\_346

software\_315

available\_301

use\_289

one\_259

program\_252

files\_242

format\_240

get\_235

version\_221

system\_219

**Contributions**

Erik Delanois - Problem 1

Doug Zhu, Dallas Delaney - Problem 2