ASSIGNMENT – 2

1. TEXT CLASSIFICATION

(a) (i) Accuracy over Train Data: **94.152 %**
Accuracy over Test Data: **82.586 %**
(ii) Word Clouds

Figure 1: Negative Word Cloud

.

Figure 2: Positive Word Cloud

(b) (i) Accuracy over Test Data for Random Prediction: **49.68 %**
(ii) Accuracy over Test Data if all classes are predicted Positive: **57.386 %**
(iii) The model created by me is 66.236 % better than Random Prediction Baseline and 43.913 % better than Positive Baseline.

(c) (i) Confusion Matrix for Test Data for My Model: **[[7998, 2002], [610, 4390]]**
Confusion Matrix for Test Data for Random Model: **[[4346, 3178], [4262, 3214]]**
Confusion Matrix for Test Data if all classes are predicted positive: **[[8608, 6392], [0, 0]]**

(ii) For Prediction using My Model, the max value in Confusion Matrix is for True Positive Predictions. It should be because our test set has 10000 positive comments out of which (according to accuracy) lot of them are classified correctly.

For Random Prediction, the max value in Confusion Matrix is for True Positive Predictions. The total number of positively classified comments appears similar to negatively classified comments.

For Positive Baseline, the max value in Confusion Matrix is for True Positive Predictions. The True Negative and False Negative Predictions are 0 as expected.

(iii) The only pattern that seems to be visible is that for every model, the value of True Positive Predictions is highest.

(d) (i) Performed in the code.

(ii) Word Clouds

Figure 3: Negative WordCloud - After Stemming and Removing StopWords

Figure 4: Positive WordCloud: After performing Stemming and removing StopWords

(iii) Accuracy over Test Data: **84.933 %**

(iv) Accuracy increased around **2.35 %** when stemming was performed.

(e) (i) Accuracy over Test Data: **87.393 %**

(ii) I added tri-grams as well as tetra-grams to my model to further increase the efficiency. Accuracy over Test Data: **89.273 %**

(iii) The accuracy in part (a) was 82.586 % and accuracy in part (d) was 84.933 % which significantly improved when bi-grams were introduced to 87.393 % and further improved to 89.273 % when tri-grams and tetra-grams were improved.

(f) (i) The required metrics for best performing model:

POSITIVE:

Precision = **0.8915**
Recall = **0.94448**
F1 Score = **0.91722**

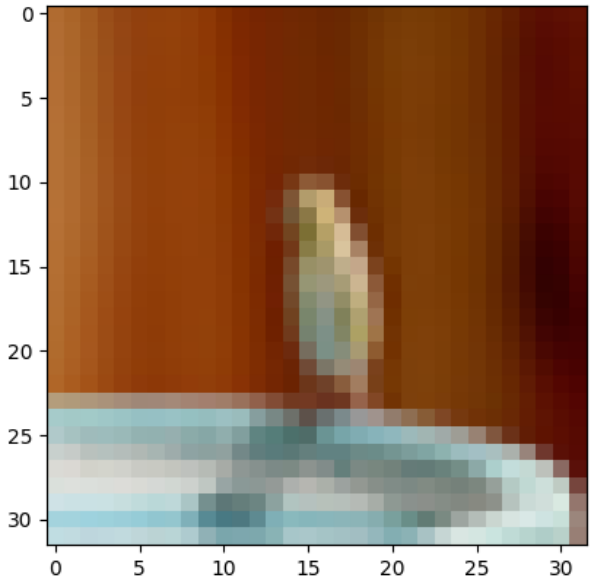
NEGATIVE:

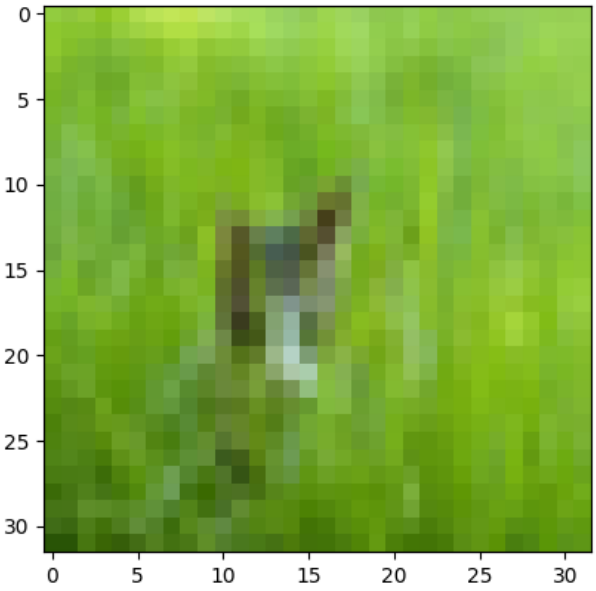
Precision = **0.8952**
Recall = **0.80489**
F1 Score = **0.84764**

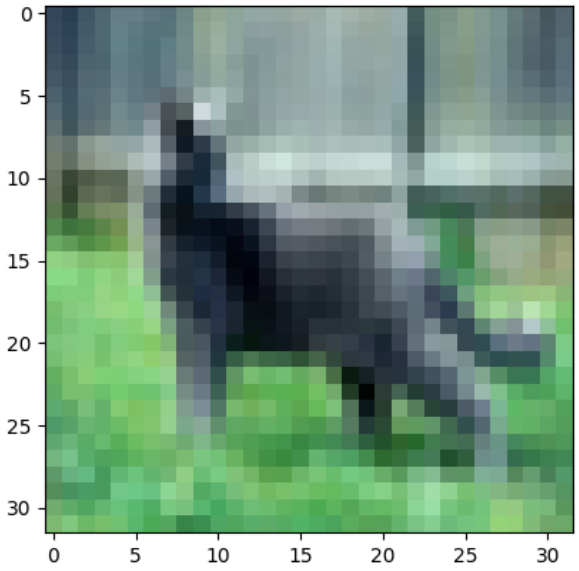
(ii)

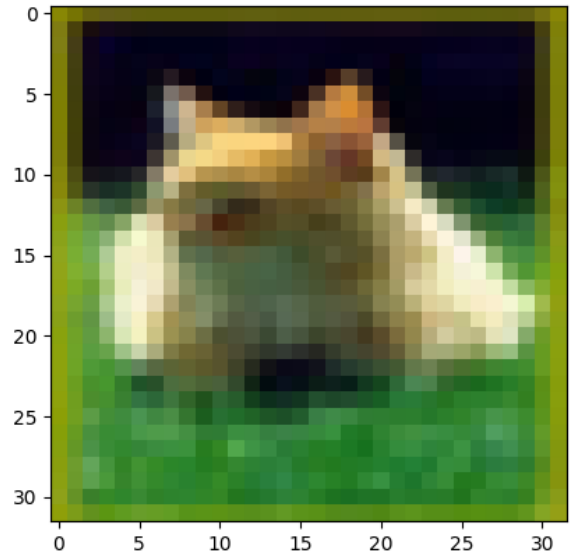
2. BINARY IMAGE CLASSIFICATION

(a) (i) No. of Support Vectors: **4000**, Percentage of Training Samples = **100 %**
(ii) Test Set Accuracy = **63.95%**
(iii) Images corresponding to top 5 co-efficients:






0
5
10
15
20
25
30
0
5
10
15
20
25
30


0
5
10
15
20
25
30
0
5
10
15
20
25
30

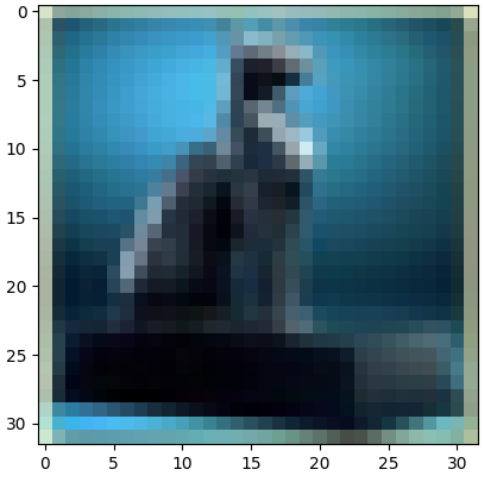
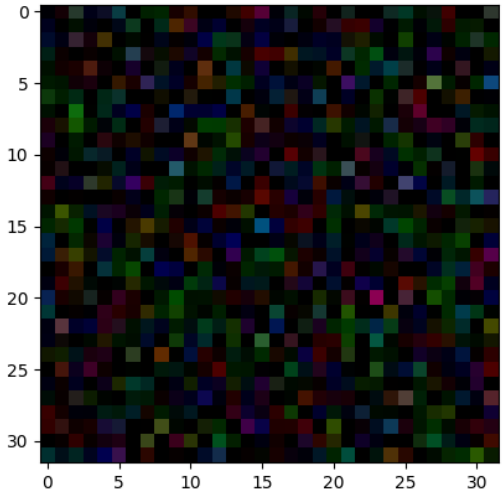
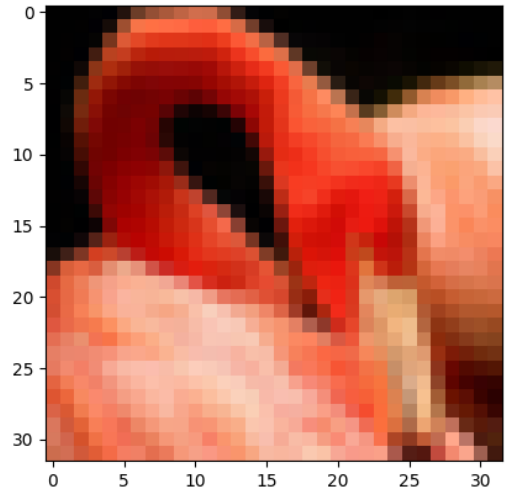
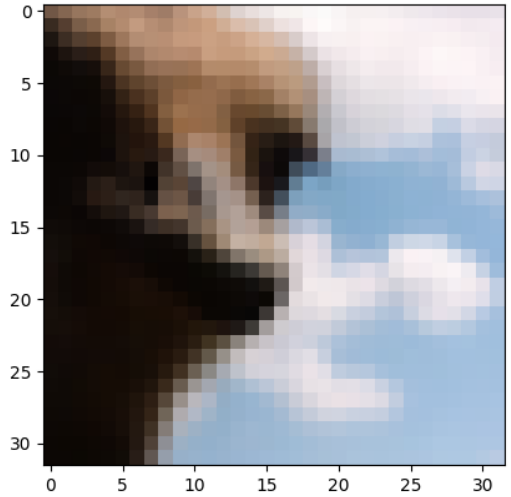


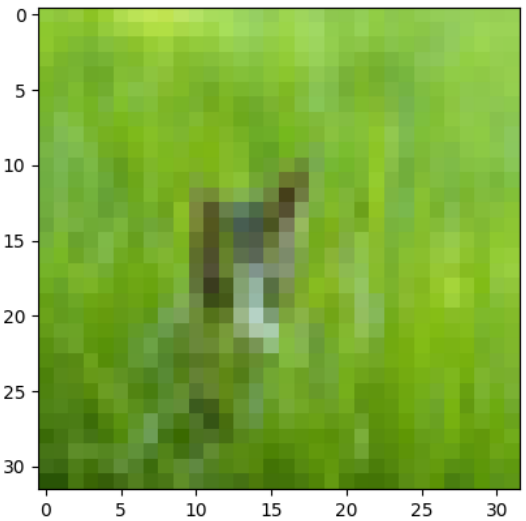
Image corresponding to reshaping W:

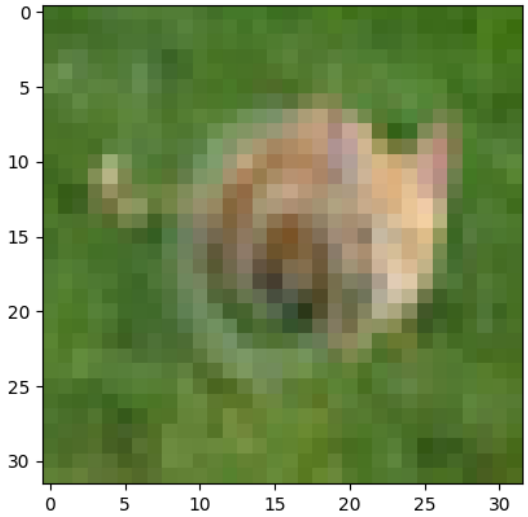


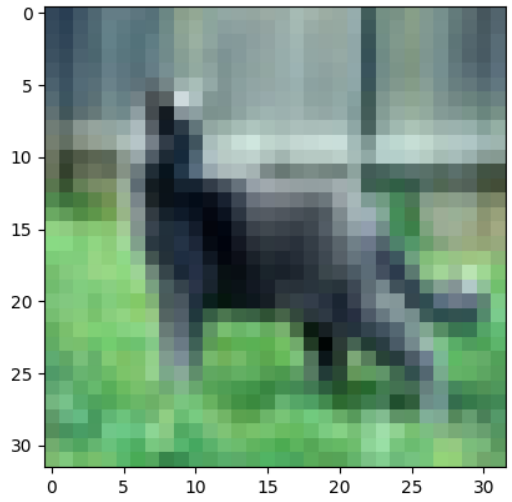
(b) (i) No. of Support Vectors: **4000**, compared to 4000 for part (a), this number is same.
(ii) Test Set Accuracy = **76.1%**
(iii) Images corresponding to top 5 co-efficients:


0
5
10
15
20
25
30
0
5
10
15
20
25
30


0
5
10
15
20
25
30
0
5
10
15
20
25
30


0
5
10
15
20
25
30
0
5
10
15
20
25
30





(iv) Test Set Accuracy decreased when Gaussian Kernel Model is used compared to Linear Kernel Model.

(c) (i) nSV for Linear Model = **2332**
nSV for Gaussian Model = **2756**
(ii) RMS Error in W = **0.014446**
Absolute Error in B = **0.18755**
(iii) LINEAR MODEL: Test Data Accuracy: **64.05 %**
GAUSSIAN MODEL: Test Data Accuracy: **76.15 %**
(iv) CVXOPT Models:
Linear Model: Time taken: **38.51448**
Gaussian Model: Time taken: **167.35019**
SK-LEARN Models:
Linear Model: Time taken: **50.68936**
Gaussian Model: Time taken: **15.03073**

3. MULTI-CLASS IMAGE CLASSIFICATION

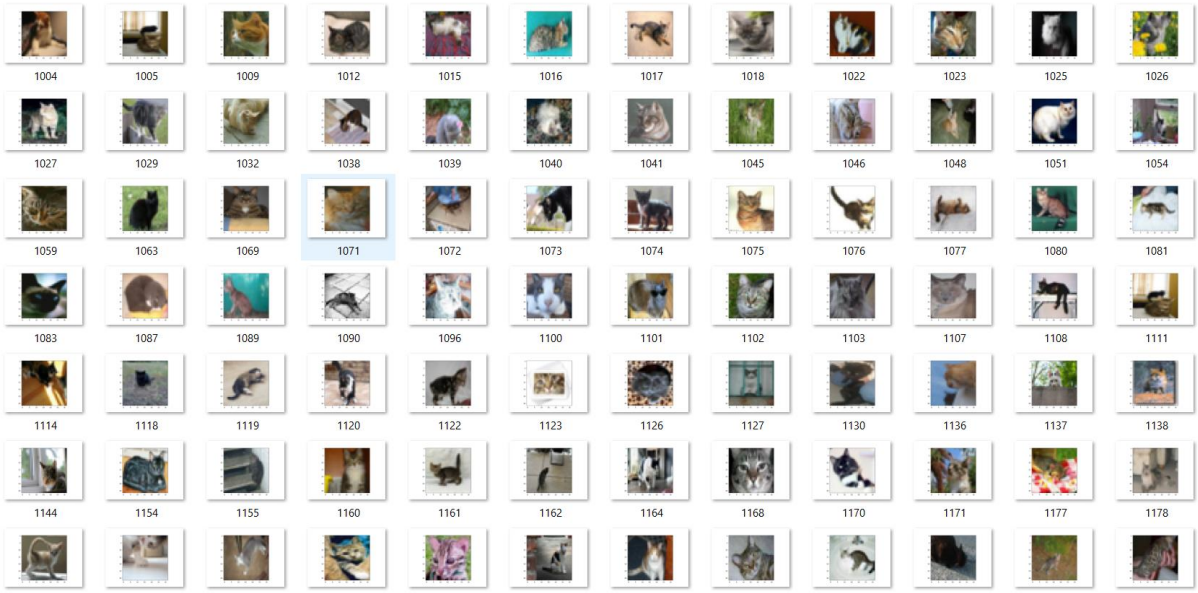
(a) Correct: 2952
Incorrect: 2048
Accuracy = **59.04 %**

(b) (i) Test Data Accuracy = **59.3%**
Time Taken by SK-LEARN: **301.14669 sec**
Time Taken by CVXOPT: **1656.95358 sec**

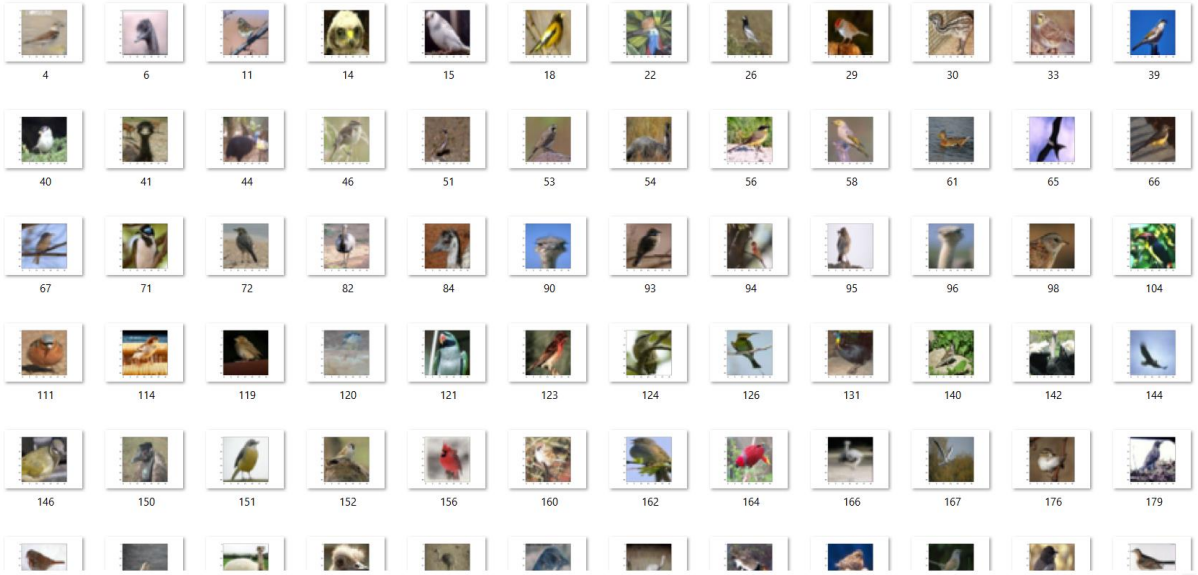
(c) Confusion Matrix for Linear Model:
[683 404]
[317 596]
Confusion Matrix for Gaussian Model:
[753 231]
[247 769]

Observation(s):
According to the Confusion Matrices above, the classification into True Positive and True Negative classes is similar. Also, the classification into False Positive and False Negative is automatically coming out to be the same.
The mis-classification is almost same for both the classes.
The True Positive class for my case consists of photos of birds whereas the True Positive class for my case consists of photos of cats.
One distinctive feature that seems to be optimal for classification is that all the cats that are facing towards the camera and face view is visible whereas in case of birds, it seems that most of the birds classified as birds are facing sideways or whole of their bodies are visible from side view.

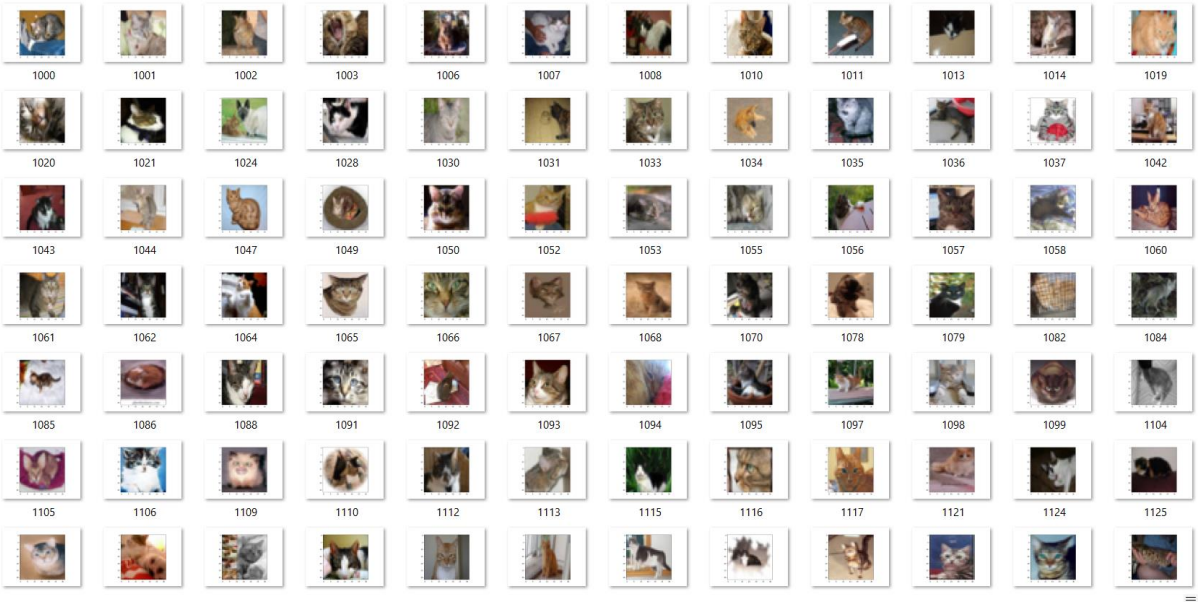
FP:



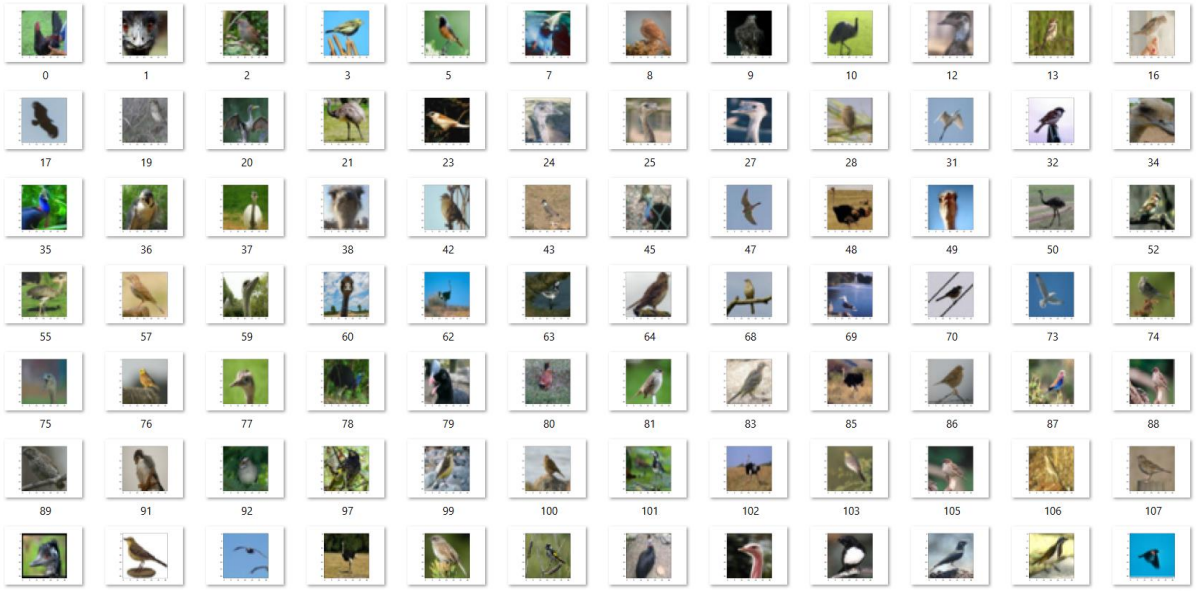
FN:


4 6 11 14 15 18 22 26 29 30 33 39
40 41 44 46 51 53 54 56 58 61 65 66
67 71 72 82 84 90 93 94 95 96 98 104
111 114 119 120 121 123 124 126 131 140 142 144
146 150 151 152 156 160 162 164 166 167 176 179

TN:


1000 1001 1002 1003 1006 1007 1008 1010 1011 1013 1014 1019
1020 1021 1024 1028 1030 1031 1033 1034 1035 1036 1037 1042
1043 1044 1047 1049 1050 1052 1053 1055 1056 1057 1058 1060
1061 1062 1064 1065 1066 1067 1068 1070 1078 1079 1082 1084
1085 1086 1088 1091 1092 1093 1094 1095 1097 1098 1099 1104
1105 1106 1109 1110 1112 1113 1115 1116 1117 1121 1124 1125

TP:


0
1
2
3
5
7
8
9
10
12
13
16
17
19
20
21
23
24
25
27
28
31
32
34
35
36
37
38
42
43
45
47
48
49
50
52
55
57
59
60
62
63
64
68
69
70
73
74
75
76
77
78
79
80
81
83
85
86
87
88
89
91
92
97
99
100
101
102
103
105
106
107