Q1. *Report the evaluation results of your model using 10-fold cross-validation.*

```
Time taken to build model: 0.68 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7973               98.3107 %
Incorrectly Classified Instances       137                1.6893 %
Kappa statistic                          0.9647
Mean absolute error                      0.0177
Root mean squared error                  0.122
Relative absolute error                  3.6872 %
Root relative squared error             24.8979 %
Total Number of Instances             8110

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.995    0.034      0.978     0.995     0.986      0.997    ham
                 0.966    0.005      0.992     0.966     0.979      0.998    spam
Weighted Avg.    0.983    0.023      0.983     0.983     0.983      0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 4838   26 |    a = ham
  111 3135 |    b = spam
```

There are 4838/ (4838+111) = 97.7% of all hams labeled accurately by the model and 3135/(3135+26) = 99.18% of all spams labeled accurately by the model.

The overall accuracy for this model is:

(4838+3135)/(4838+3135+26+111)=12811/12948=98.9%

Recall for this model is:

4838/ (4838+111) =97.7%

Precision for this model is:

4838/(4838+26)=99.5%

Conclusion: This model has shown great classification ability with a high accuracy rate, especially when to identify spams.

Q2. ***Report the 10-fold cross-validation results and compare with the occurrence-based results in the previous question.***

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7959               98.1381 %
Incorrectly Classified Instances       151                1.8619 %
Kappa statistic                          0.961
Mean absolute error                      0.02
Root mean squared error                  0.1223
Relative absolute error                  4.1695 %
Root relative squared error             24.9558 %
Total Number of Instances             8110

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.996     0.04       0.974     0.996     0.985      0.997    ham
                 0.96      0.004      0.993     0.96      0.976      0.998    spam
Weighted Avg.    0.981     0.026      0.982     0.981     0.981      0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 4843   21 |    a = ham
  130 3116 |    b = spam
```

Naïve Bayes Multinomial is much more effective than Naïve Bayes with much shorter time to build the model.

The frequency-based model performs similar to the occurrence based model, the former with a little worse overall performance. However, this model is relatively performs a little better on predicting spam rather than ham, which means it is more likely to classify ham into spam while the occurrence-based model is more likely to do vice versa.

Q3. Calculate the total cost and expected cost (per email) based on the confusion matrix you obtained in question 1. ***Copy the confusion matrix and present the formulas you used to get the results.*** [Be careful with the dimensions of the confusion matrix: which are the "actuals" and which are the "predictions"?]

For occurrence-based model:

```
Time taken to build model: 0.68 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7973               98.3107 %
Incorrectly Classified Instances       137                1.6893 %
Kappa statistic                          0.9647
Mean absolute error                      0.0177
Root mean squared error                  0.122
Relative absolute error                  3.6872 %
Root relative squared error             24.8979 %
Total Number of Instances             8110

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.995     0.034      0.978     0.995     0.986      0.997    ham
                 0.966     0.005      0.992     0.966     0.979      0.998    spam
Weighted Avg.    0.983     0.023      0.983     0.983     0.983      0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 4838   26 |    a = ham
  111 3135 |    b = spam
```

The cost would be: 111 * 5/100 + 26 * 5= $135.55

For frequency-based model:

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7959               98.1381 %
Incorrectly Classified Instances       151                1.8619 %
Kappa statistic                          0.961
Mean absolute error                      0.02
Root mean squared error                  0.1223
Relative absolute error                  4.1695 %
Root relative squared error             24.9558 %
Total Number of Instances             8110

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.996     0.04       0.974     0.996     0.985      0.997    ham
                 0.96      0.004      0.993     0.96      0.976      0.998    spam
Weighted Avg.    0.981     0.026      0.982     0.981     0.981      0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 4843   21 |    a = ham
  130 3116 |    b = spam
```

The cost would be 130 * 5/100 + 21 * 5 = $111.5

Q4. *Calculate the total cost and expected cost. Compare the accuracy and the costs with those of the cost "insensitive" model you built earlier.*

```
Time taken to build model: 0.58 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7817               96.3872 %
Incorrectly Classified Instances       293                3.6128 %
Kappa statistic                          0.9236
Mean absolute error                      0.0361
Root mean squared error                  0.1901
Relative absolute error                  7.5251 %
Root relative squared error             38.7948 %
Total Number of Instances             8110

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0.089      0.944     0.999     0.971      0.955    ham
                 0.911     0.001      0.999     0.911     0.953      0.955    spam
Weighted Avg.    0.964     0.054      0.966     0.964     0.964      0.955

=== Confusion Matrix ===

    a    b   <-- classified as
 4860    4 |    a = ham
  289 2957 |    b = spam
```

In this case, the cost would be 289*5/100 + 4*5 = $34.45

The total cost, compared to that of models built before, has been decreased dramatically. But the model accuracy, at the same time, is impaired. Only the precision for spam class and recall for ham class are improved.

Q5. *Compare the results of evaluating this new model with the one you generated in question 1 using the full set of features. What is your observation?*

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7564               93.2676 %
Incorrectly Classified Instances       546                6.7324 %
Kappa statistic                          0.8624
Mean absolute error                      0.1138
Root mean squared error                  0.2391
Relative absolute error                 23.6952 %
Root relative squared error             48.7934 %
Total Number of Instances             8110

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.905     0.026      0.981     0.905     0.942      0.961    ham
                 0.974     0.095      0.872     0.974     0.921      0.961    spam
Weighted Avg.    0.933     0.053      0.938     0.933     0.933      0.961

=== Confusion Matrix ===

    a    b   <-- classified as
 4401  463 |    a = ham
   83 3163 |    b = spam
```

Conclusion: 30 features are obviously too few for the model to learn. We can see from the decreasing overall accuracy that the model is under-fitting because we've left out too many important features. It would be better if we increase the input feature number. However, the model still performs well on predicting hams, with only 83 out of (83+4401) hams missed. This model is also very likely to classify spam into ham, which makes this model yield large cost.

Q6

Applying TFIDF and using the algorithm of NaïveBayes:

```
Time taken to build model: 4 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7268               89.6178 %
Incorrectly Classified Instances       842               10.3822 %
Kappa statistic                          0.7741
Mean absolute error                      0.1023
Root mean squared error                  0.3097
Relative absolute error                 21.3135 %
Root relative squared error             63.2123 %
Total Number of Instances             8110

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0.258      0.853     0.999     0.92       0.994    ham
                 0.742     0.001      0.998     0.742     0.851      0.996    spam
Weighted Avg.    0.896     0.155      0.911     0.896     0.893      0.995

=== Confusion Matrix ===

    a    b   <-- classified as
 4860    4 |    a = ham
  838 2408 |    b = spam
```

Conclusion: The model is less efficient when applying TFIDF, with the running time of 4 seconds. It is very good at identifying spam though, with only 4 out of 2412 spam missed. However, it misses 838 hams, classifying them into spam. The cost of this model would be:

838 * 5/100 + 4 * 5 = $ 61.9