Categorize Mistakes

After calculated the Mutual Information based on the *Training Set*, we can find the following 10 words that have the highest MI with the categorization:

Call, to, or, FREE, claim, call, &, To, mobile, Txt

And after predicted the *Test Set*, we can have the following confusion matrix:

```
                | Predict True | Predict False
--------------- | ------------ | --------------
 Actually True  |           96 |           106
 Actually False |            3 |          1189
None
Accuracy: 0.9218077474892395
Precision: 0.9696969696969697
Recall: 0.4752475247524752
FPR: 0.0025167785234899327
FNR: 0.5247524752475248
```

From the matrix we can find out that there only have 3 **False Positive** case, so by removing the threshold we can get the 20 sentences that not near the 0 to do the analysis.

Since the **False Positive** case is a little bit complex in our situation, let's discuss about the **False negative** case first.

Worst False Negative

By analyze the 20 sentences, we can identify them with these 4 categories to get better predict value:

- Has URL
- Has digital word
- Has all uppercase word
- Use uncommon punctuation

Here is the program output that show how these categories related with this case:

```
From the 20 sentence, there have 5 sentence has URL
From the 20 sentence, there have 12 sentence has digit word
From the 20 sentence, there have 16 sentence use upper case
From the 20 sentence, there have 16 sentence use uncommon punctuation (! / : @ ;)
```

False Positive

By analyze the 3 **False Positive** sentences, we can identify them with this 1 category:

- Not too long (< 150 characters) and not too short (> 60 characters)

Here is the program output that show how this category related with this case:

```
Only 3 sentence, are in False Positives
From the 3 sentence, there have 3 sentence length is larger than 120
```

By analyze the 3 **False Positive** and other 17 **True Negative** but near the threshold sentences, we can identify them with these 4 categories to get better predict value for those **True Negative** sentences:

- Too Short (<60 characters)
- Too Long (>150 characters)
- Has all uppercase word
- Use uncommon punctuation

Here is the program output that show how these categories related with this case:

```
From the 17 sentence, there have 4 sentence length is less than 60
From the 17 sentence, there have 5 sentence length is larger than 150
From the 17 sentence, there have 12 sentence use upper case
From the 17 sentence, there have 12 sentence use uncommon punctuation (! / : @ ;)
```

New Heuristic Feature

From this process, we can know that the manual check of the data and the out come is very important in our ML process, this can:

- Help us find more features that can reduce the **False Positive** and **False Negative**, also increase the accuracy.
- Find those features and make sure those features will work need a lot of experiments to re-training the model and test it.
- Sometimes we might need to think if those data are actual dirty data, which means the labels are not correct. In this case, it feels like those 3 **False Positive** sentences should be in span message.

For reduce the **False Positive**, we can add the feature that check:

- The length of the sentence if it not too long and not too short. In the manual check we can find that spam messages are usually either too long (> 150 characters) or too short (<60 characters)

For reduce the **False Negative**, we can add the features that check:

- Is the sentence containing a lot of uppercase words?
- Is the sentence containing the un-commonly used punctuation?

Both 2 features have the same high value in our test, which make them a good way to reduce the **False Negative** prediction.