Aren Deza
2019-50022

MACHINE LEARNING - PROBLEM SET 3

Results and Discussion

Stopwords	Accuracy	Recall	Precision
removed	0.518333	0.304348	0.875000
kept	0.515000	0.299233	0.873134

1. As indicated by the table presented above, it is necessary to remove stop words to improve the accuracy, recall, and precision of the model.

   As stop words are not useful in classification and reduce the dimensionality of a given dictionary, making sure to remove them beforehand produces a slight positive effect on the accuracy and precision of the model, and a noticeable improvement in the model's recall.

```
FP: 17
FN: 272
TP: 119
TN: 192
-----
Acc: 0.5183333333333333
r:   0.30434782608695654
P:   0.875
```

```
FP: 194
FN: 25
TP: 366
TN: 15
-----
Acc: 0.635
r:   0.9360613810741688
P:   0.6535714285714286
```

2. Shown above are the performance metrics of the model when the dictionary size is untouched (left) and limited (right). Here, the dictionary has been filtered to only include words that appear more than 100000 times. Having a smaller dictionary size has significantly increased the rate at which emails are classified as spam, which is risky in a practical setting as high false positive detections can cause more inconveniences to email users than a large amount of false negatives.

   While it has produced dramatically improved effects on the model's accuracy and recall, its precision has been negatively affected.

```
alpha: 2.0

FP: 5
FN: 280
TP: 111
TN: 204
-----
Acc: 0.525
r:   0.28388746803069054
P:   0.9568965517241379
```

```
alpha: 0.5

FP: 66
FN: 225
TP: 166
TN: 143
-----
Acc: 0.515
r:   0.42455242966751916
P:   0.7155172413793104
```

```
alpha: 0.1

FP: 157
FN: 145
TP: 246
TN: 52
-----
Acc: 0.49666666666666665
r:   0.629156010230179
P:   0.6104218362282878
```

```
alpha: 0.005

FP: 165
FN: 120
TP: 271
TN: 44
-----
Acc: 0.525
r:   0.6930946291560103
P:   0.6215596330275229
```

3. The figures presented above illustrate the effects that adjusting the alpha has on the model. From merely observing each result and comparing them to the default results shown in the previous item, the following connections can be made:

	Larger Alpha	**Smaller Alpha**
FP	decreases	increases
FN	increases	decreases
TP	decreases	increases
TN	increases	decreases
Acc	increases	decreases, then increases past a certain threshold.
r	decreases	increases
P	increases	decreases, then increases past a certain threshold.

4. To make potential improvements to the model, the following recommendations are made:
    a. Based on the observations made from the other parts of the discussion, it is necessary to make sure that stop words are removed from the dictionary. If possible, the code must include implementations to make this operation much faster so as to cover more amounts of data.
    b. Similarly, the minimum number of occurrences for a word to be included in the dictionary must not be restricted so as to limit the occurrence of false positives in the model's classifications.
    c. Furthermore, the alpha should be kept at 1 in order to provide balanced and optimal results.
    d. Outside of the performance metrics, I have observed that although we are capable of removing stop words, the dictionary is still cluttered by other words that cannot be accounted for, such as people's names and email addresses.
    If possible, these words should be removed from the dictionary in order to maintain the dimensionality of the model. Ideally, this could be done by limiting words that enter the dictionary to only those that are part of a known language (such as with the PyEnchant module).
    e. Singular and plural nouns should be counted as the same word, and the same applies for verbs that possess multiple forms.
    In 000/000, there are several instances of the words 'Catholic' and 'Catholics', which could be easier to parse through if singular and plural nouns were treated interchangeably.
    f. A spell checker should also be introduced so as to prevent spam emails from bypassing the filters by using deliberate mispronunciation, such as with the names of the drugs on 071/002.
    g. The code should be optimized to work faster to allow it to process more emails for more accurate classifications. As it is currently, it takes about 2 minutes to classify 1000 emails.
    h. HTML tags should also be filtered out. In the case of emails like 000/023, there are more HTML tags than text, and these tags can potentially skew the results if they are not handled appropriately.
    i. Finally, there should also be a mechanism for detecting and evaluating links. If not to act as an additional measure for spam detection, they should at least be removed to avoid cluttering the dictionary.