Exercise 4: Programming Exercise: Naive Bayes Classifier

We provide you with a data set that contains spam and non-spam emails ("hw5_nb.zip"). Please use the Naive Bayes Classifier to detect the spam emails. Finish the following exercises by programming. You can use your favorite programming language.

1. Remove all the tokens that contain non-alphabetic characters.
2. Train the Naive Bayes Classifier on the training set according to Algorithm 2.
3. Test the Naive Bayes Classifier on the test set according to Algorithm 3. You may encounter a problem that the likelihood probabilities you calculate approach 0. How do you deal with this problem?
4. Compute the confusion matrix, accuracy, precision, recall, and F-score.
5. Without the Laplace smoothing technique, complete the steps again.

Algorithm 2 Training Naive Bayes Classifier

```
Input: The training set with the labels $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$.
$\mathcal{V} \leftarrow$ the set of distinct words and other tokens found in $\mathcal{D}$
for each target value $c$ in the labels set $\mathcal{C}$ do
    $\mathcal{D}_c \leftarrow$ the training samples whose labels are $c$
    $P(c) \leftarrow \frac{|\mathcal{D}_c|}{|\mathcal{D}|}$
    $T_c \leftarrow$ a single document by concatenating all training samples in $\mathcal{D}_c$
    $n_c \leftarrow |T_c|$
    for each word $w_k$ in the vocabulary $\mathcal{V}$ do
        $n_{c,k} \leftarrow$ the number of times the word $w_k$ occurs in $T_c$
        $P(w_k|c) = \frac{n_{c,k}+1}{n_c+|\mathcal{V}|}$
    end for
end for
```

Algorithm 3 Testing Naive Bayes Classifier

```
Input: An email $\mathbf{x}$. Let $x_i$ be the $i^{th}$ token in $\mathbf{x}$ . $\mathcal{I} = \emptyset$.
for $i = 1, \ldots, |\mathbf{x}|$ do
    if $\exists\, w_{k_i} \in \mathcal{V}$ such that $w_{k_i} = x_i$ then
        $\mathcal{I} \leftarrow \mathcal{I} \cup i$
    end if
end for
predict the label of $\mathbf{x}$ by
    $$\hat{y} = \arg\max_{c \in \mathcal{C}} P(c) \prod_{i \in \mathcal{I}} P(w_{k_i}|c)$$
```

Solution: ■