0. Using “TF × IDF” value in Logistic Regression algorithm

I found if just using the frequency of words in one sample, the accuracy of Logistic Regression algorithm is not high. Because the frequency of words in the sample cannot reflect the degree of importance or weighting info of the words, for example, sometimes the very common word tends to appear in every sample and it is less important.
The paper **[1]** introduced “TF × IDF”, which is short for “term frequency–inverse document frequency”. It is intended to reflect how important a word is to a document in a collection or corpus. The “TF × IDF” value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. So “TF × IDF” value is much more reasonable to represent the weighting info of word in text.
The paper **[2], [3]** in follow Reference use “TF × IDF” value to preprocess the input text or document. To reduce the impact of document length, the paper **[3]** “cosine-normalize” the feature vectors to have a Euclidean norm of 1.0.

TF: Term Frequency, which measures how frequently a term occurs in a document.
TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, which measures how important a term is.
IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

Reference

1. G. Salton, etc. "Term-weighting approaches in automatic text retrieval."
2. Alexander Genkin, David D. Lewis, David Madigan. " Sparse Logistic Regression for Text Categorization."
3. Alexander Genkin, David D. Lewis. "Large-Scale Bayesian Logistic Regression for Text Categorization."

1. Accuracy before removal of stop words:

(1)The output result of multinomial Naive Bayes algorithm
CorrectCount = 453
ErrorCount = 25
Test Accuracy = 0.9476988

(2) The output result of MCAP Logistic Regression algorithm
Values of λ is 0.001; Learning Rate is 0.1; Iterations is 100.
CorrectCount = 456
ErrorCount = 22
Correct Accuracy = 0.9539748953974896

2. Accuracy after removal of stop words

(1)The output result of multinomial Naive Bayes algorithm
CorrectCount = 454
ErrorCount = 24
Test Accuracy = 0.9497908
(2) The output result of MCAP Logistic Regression algorithm
Values of λ is 0.001; Learning Rate is 0.1; Iterations is 100.
CorrectCount = 457
ErrorCount = 21
Correct Accuracy = 0.9560669456066946

3. Explanation of stop words

After removal of stop words, we can see that:
The Accuracy of Naive Bayes algorithm increases 0.209%.
The Accuracy of Logistic Regression algorithm increases 0.2092%.
Stop words are basically a set of commonly used words in any language, which carry less important meaning than keywords. So if we remove the words that are very commonly used in a given language, we can focus on the important words instead. So we can improve the accuracy of Text Classification.

4. The impact of "λ" in Logistic Regression Algorithm

Before removal of stop words, if we set **LearningRate = 0.01**, we can see the impact of "λ" more clearly under the condition "**Iterations is 100**".
(1) When λ = 0, the LR accuracy is 0.914
(2) When λ = 0.001, the LR accuracy is 0.920
(3) When λ = 0.01, the LR accuracy is 0.926
(4) When λ = 0.1, the LR accuracy is 0.926
(5) When λ = 1, the LR accuracy is 0.939
(6) When λ = 10, the LR accuracy is 0.949
(7) When λ = 20, the LR accuracy is 0.930
Maybe **LearningRate = 0.01** is not a good idea in this example, but we can see the impact of "λ" more clearly

5. The impact of "Iterations" in Logistic Regression Algorithm

Before removal of stop words, if we set **LearningRate = 0.01**, and **λ = 0.001**, we can see the impact of "Iterations" more clearly.
(1) When Iterations = 100, the LR accuracy is 0.920
(2) When Iterations = 500, the LR accuracy is 0.949
(3) When Iterations = 1000, the LR accuracy is 0.953
(4) When Iterations = 5000, the LR accuracy is 0.947
We can see when the times of Iterations reaches a certain degree, the LR accuracy **cannot** be improved.