Report for Naïve Bayes and Logistic Regression

Yidan Sheng yxs173130

|  |  |  |
| --- | --- | --- |
| Accuacy | Before removing stop words | After removing stop words |
|  | 0.9476987447698745 | 0.9414225941422594 |

Naïve Bayes

Logistic Regression

|  |  |  |
| --- | --- | --- |
| Accuracy  λ | Before removing stop words | After removing stop words |
| λ = 0.001 | 0.9581589958158996 | 0.9560669456066946 |
| λ = 0.01 | 0.9581589958158996 | 0.9560669456066946 |
| λ = 0.1 | 0.9497907949790795 | 0.9539748953974896 |
| λ = 0.2 | 0.9435146443514645 | 0.9581589958158996 |
| λ = 0.3 | 0.9518828451882845 | 0.9539748953974896 |

The hard limit on numbers of iteration is 50.

After removing stop words, the accuracy of Naïve Bayes decreases, and accuracy of Logistic Regression sometimes increases or decreases, but the accuracy flows between 0.94 and 0.96.

Although stop words don’t have any sense for classifying, it is still possible that certain words are associated with certain topic. These stop words are general such as daily, new or literature, may not for email subject topic.

For example in Naïve Bayes, conditional probability of ‘me’ in ham is 0.004(higher than average) which is a good word to classify ham email. The similar example in Logistic Regression, the weight of the feature which is stop word is big, means using heavily by the model. When we remove this stop word, the accuracy might decrease. However, it might increase the accuracy due to different lambda, because the lambda penalize the lambda.