TEXT CLASSIFICATION

Naive Bayes and Logistic Regression - Analysis Report

Evaluate Naive Bayes and Logistic Regression for text classification for the Spam or Ham Dataset

Naive Bayes

Category	Accuracy
Without Stop Words	94.56066945606695
With Stop Words	94.76987447698745

We can notice that the Accuracy without stop words slightly decreases. Smoothing made the model more accurate

Logistic Regression

For an initial run with $\lambda = 0.1$, *Iterations=100,* $\eta = 0.1$ the result is as follows

Category	Accuracy
Without Stop Words	95.60669456066945
With Stop Words	94.35146443514645

Below is the result for varied values of Lambda (λ) and Iterations Count. Learning Rate $(\eta) = 0.1$ was used for all the runs.

λ	Iterations	Accuracy; W/o Stop Words	Accuracy- With Stop Words
0.1	100	95.60669456066945	94.35146443514645
0.01	100	95.60669456066945	94.76987447698745
0.001	100	95.39748953974896	94.97907949790795
0.1	300	94.56066945606695	94.56066945606695

0.01	300	95.60669456066945	94.76987447698745
0.001	300	95.81589958158996	94.56066945606695

Accuracy of Logistic Regression decreases when the stop words are removed but we can notice that the value is high among the rows with high values of λ. The accuracy with the presence of stop words increases but we notice that the value is low among the rows with high λ.