MACHINE LEARNING ASSIGNMENT 3

Name: Akshay Kumar B Kale NetId: Abk190001

Question 1, 3: Naïve Bayes Algorithm with and without using stopwords.

General Command:

python <python file> <Path for spam training files> <Path for ham training files> <Path for test spam files> <Path for ham test files> <stopWords file path> <yes/no to remove stop-words>

Results:

Commands for **with** StopWords:

python NB.py train\spam train\ham test\spam test\ham stopWords.txt yes

Accuracy: 94.5607

Commands for **without** StopWords:

python NB.py train\spam train\ham test\spam test\ham stopWords.txt No

Accuracy: 94.7699

Accuracy decreases after removing the stop words.

Question 2, 3: Logistic Regression Algorithm with and without using stopwords

General Command:

python <python file> <Path for spam training files> <Path for ham training files> <Path for test spam files> <Path for ham test files> <stopWords file path> <yes/no to remove stop-words> <lamda value>

Commands for **with** StopWords:

python LR.py train\spam train\ham test\spam test\ham stopWords.txt yes 0.1

Commands for **without** StopWords:

python LR.py train\spam train\ham test\spam test\ham stopWords.txt No 0.1

Hard limit on Number of iteration=100

Results:

λ	η	Number of iterations	Accuracy without stop words	Accuracy with stop words
0.1	0.1	100	95.6067	94.3515

0.01	0.1	100	95.6067	94.7699
0.001	0.1	100	95.3975	94.9790
0.005	0.1	100	95.6067	94.7699
0.005	0.01	100	95.1882	96.0251

Observations for Both:

- In Naïve Bayes, Accuracy is increasing without stopwords is higher than the with stop words.
- In Logistic Regression, from the above table if **η** is very small then changes in the value of **λ** results in negligible/no change in accuracy
- Accuracy of Logistic Regression sometimes decreases or increases, its in the range between 94 -96 %.
- With stop words, the accuracy of LR increases, but decreases for high values of λ because some stop words aren't useful for classification of a mail as spam or ham sometime produces change in the overall accuracy
- Without stop words, the accuracy of Logistic Regression decreases in some of the cases and increases for high values of λ because λ is the penalty on higher values to avoid overfitting.