DS 222 : Machine Learning with Large Datasets
Assignment 1

Shreyas R

1. Local Naive Bayes

In this part of assignment, a Naive Bayes document classifier was implemented in-memory.
Please refer to the code **nb_local.py**

The accuracy was fount to be and the time taken was fount to be seconds.

2. Hadoop MAP-REDUCE implementation of Naive Bayes

In this part of assignment, a Naive Bayes document classifier was implemented in Hadoop using the map-reduce framework.
Please refer to the codes **mapper_().py** and **reducer_().py** . The final code to run the entire setup is run_nb.sh .

The accuracy was fount to be

The graph of time taken with different number of reducers is obtained as follows. There is not a linear decrease in the time taken. The time taken reduces somewhat exponentially with the number of reducers. This is due to the minimum amount of overheads that can not be avoided.

3. Acknowlegements

1. I thank my classmates Bharath P and Sonali Singh for helping me with the implementation issues.
2. I have used beautiful soup package to remove unwanted stopwords.