BigML Assignment 1: Streaming Naive Bayes

Hao Gao (haog)

Did you receive any help whatsoever from anyone in solving this assignment? No
Did you give any help whatsoever to anyone in solving this assignment? No

Q1
If we discard the some stop words like “a”, “the”, we can reduce the RAM used.
We can also use disk to store the countings, but may influence the running time.

Q2
We can label the instance with labels which have the top k probabilities.
For example, we have probabilities 0.1, 0.2, 0.3, 0.4 and k is 2. Then we label the instance with labels have probabilities 0.3 and 0.4.

Q3
If a word is not in the document, we will get 0 for the probability. To avoid this case, we use the smoothing.

Q4
The expected value of the posterior distribution uses a Dirichlet distribution with smoothing parameter as a prior.