Info284 Fev004

Second obligatory assignment

The cleaning function and the function to create the vocabulary has been greatly improved and now the vocabulary and the dataset have the same words in them. I notice in my last attempt that I was removing way to many things.

In the bayer() function I have simplified the formula so that “a” represents what is before divison, which is the same as what is before the “+” sign in bayers rule. “b” is then what is after the “+” sign. Before the result of bayers rule the prior-probabilites are multiplied to a and b.

The “VocabularyWithProb.txt” file has this order:
word
P(0)
P(1)

I omitted more explanation in the vocabulary itself so it would be easier to read in to the program later.

I have a problem in that each probability value for each word is quite low, like 0.000345… If I put in an article into my bayes function and multiply lets say 80 of those numbers, I will in the end get 0. Now if all the parameter-numbers to bayes is 0 then obviously it wont work. There is something amiss in my naive bayes understanding. A small fix to atleast get the bayes function running is to change the product function to just add up all the numbers but not multiply them(also changed createVocabularyWithProb function to default to 0.0 if word is only in one class, which would not be possible if product function was in fact multiplying).

How to try out the classifier with your own text:
1. Open terminal/powershell in same directory as main.py is in.
2. Run the command: python main.py “Your text comes here”.

As of now I have not calculated an error rate since I suspect it will be quite high. The classifier is finished, but not really working. Im pretty sure I need to redeliver this project