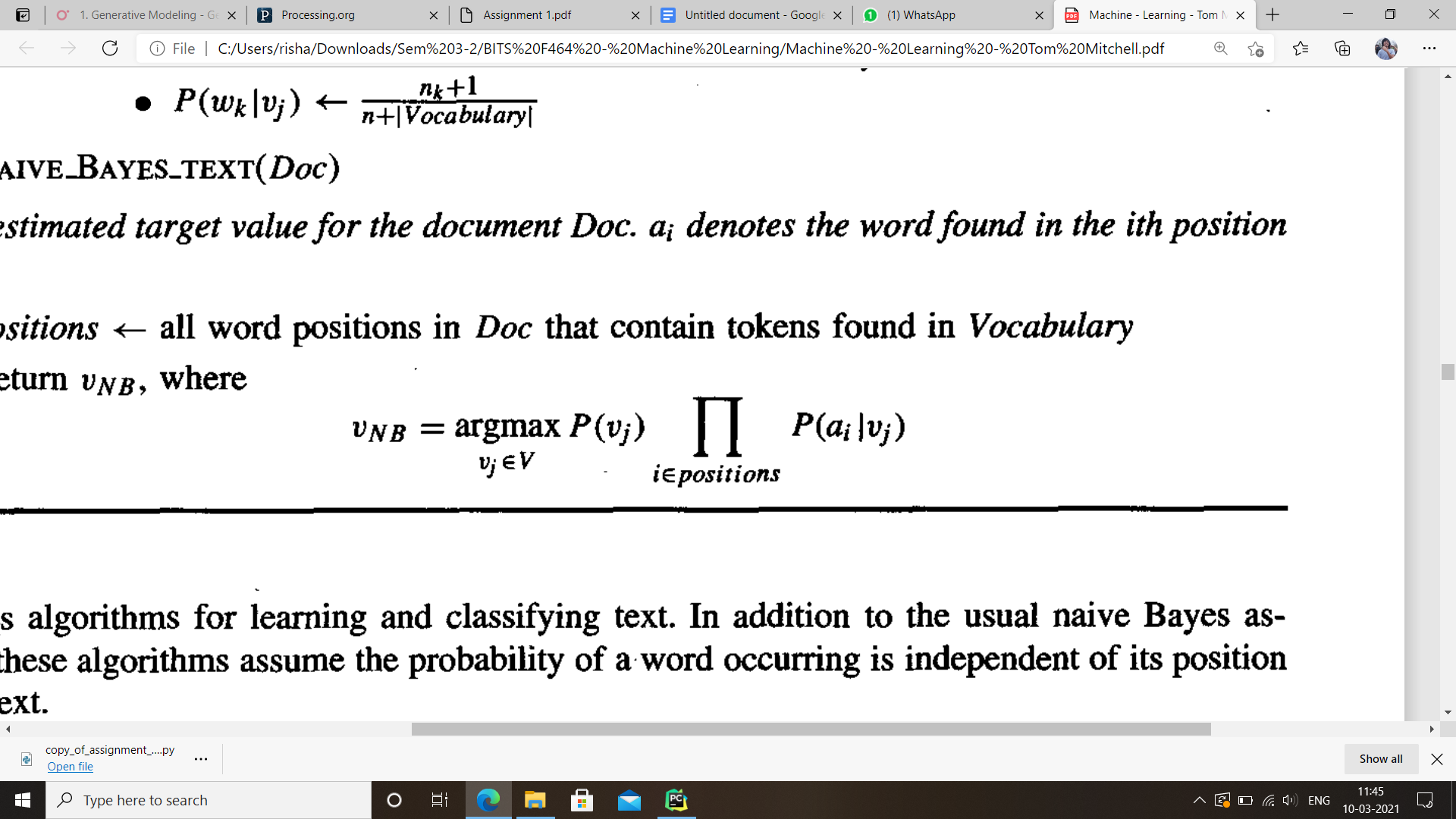
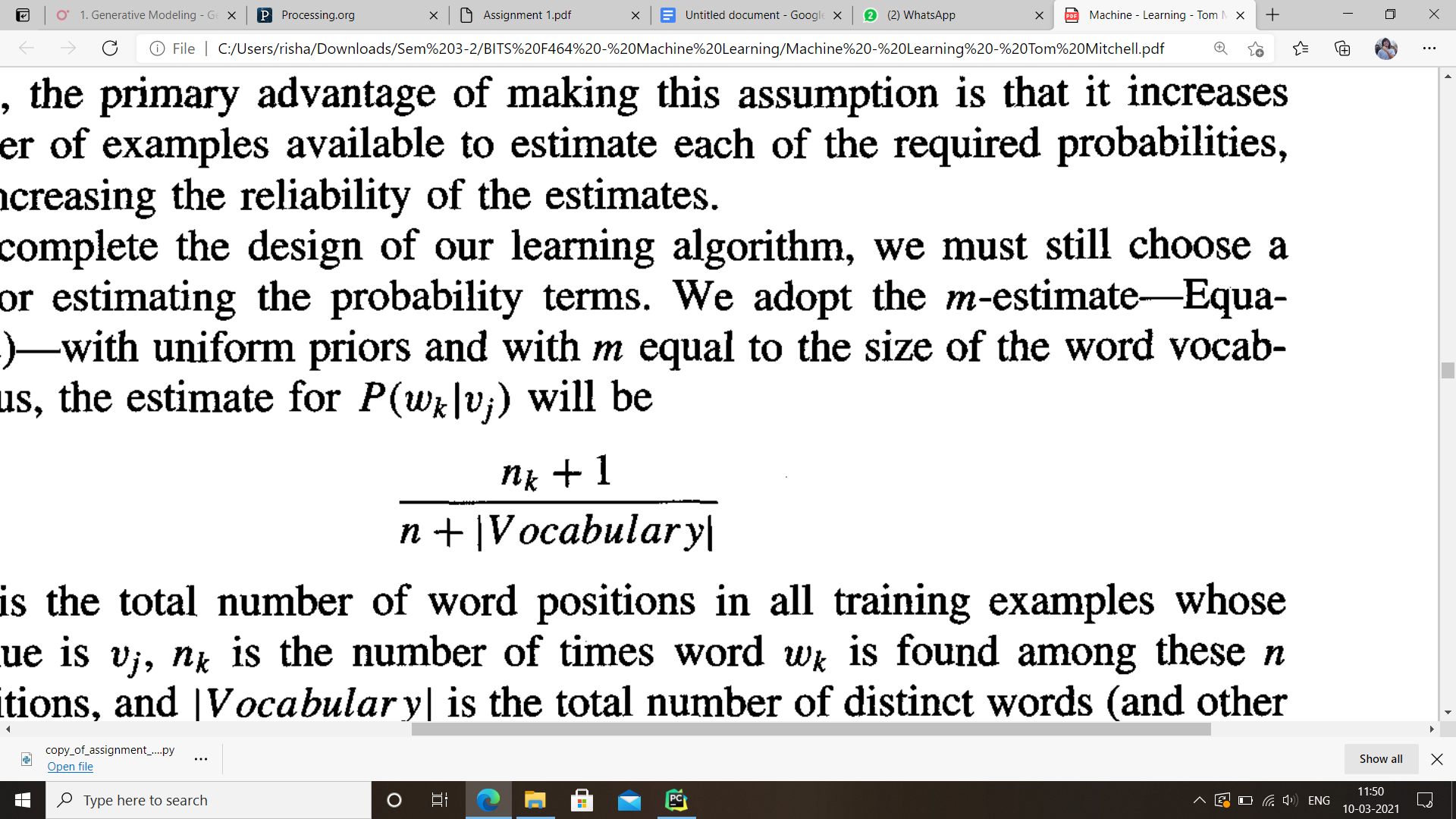
Q1. A very brief description of your model and its implementation.

The model is a direct implementation of naive bayes classifier which has been mentioned in the book “Machine Learning by Tom Mitchell”. We have taken an assumption that each word that will occur is not dependent on the words before it. This assumption allows us to greatly simplify the model. For a sentence w1w2w3w4 , we calculate two probabilities, the first one is p(spam|w1w2w3w4) and second one p(not spam|w1w2w3w4).

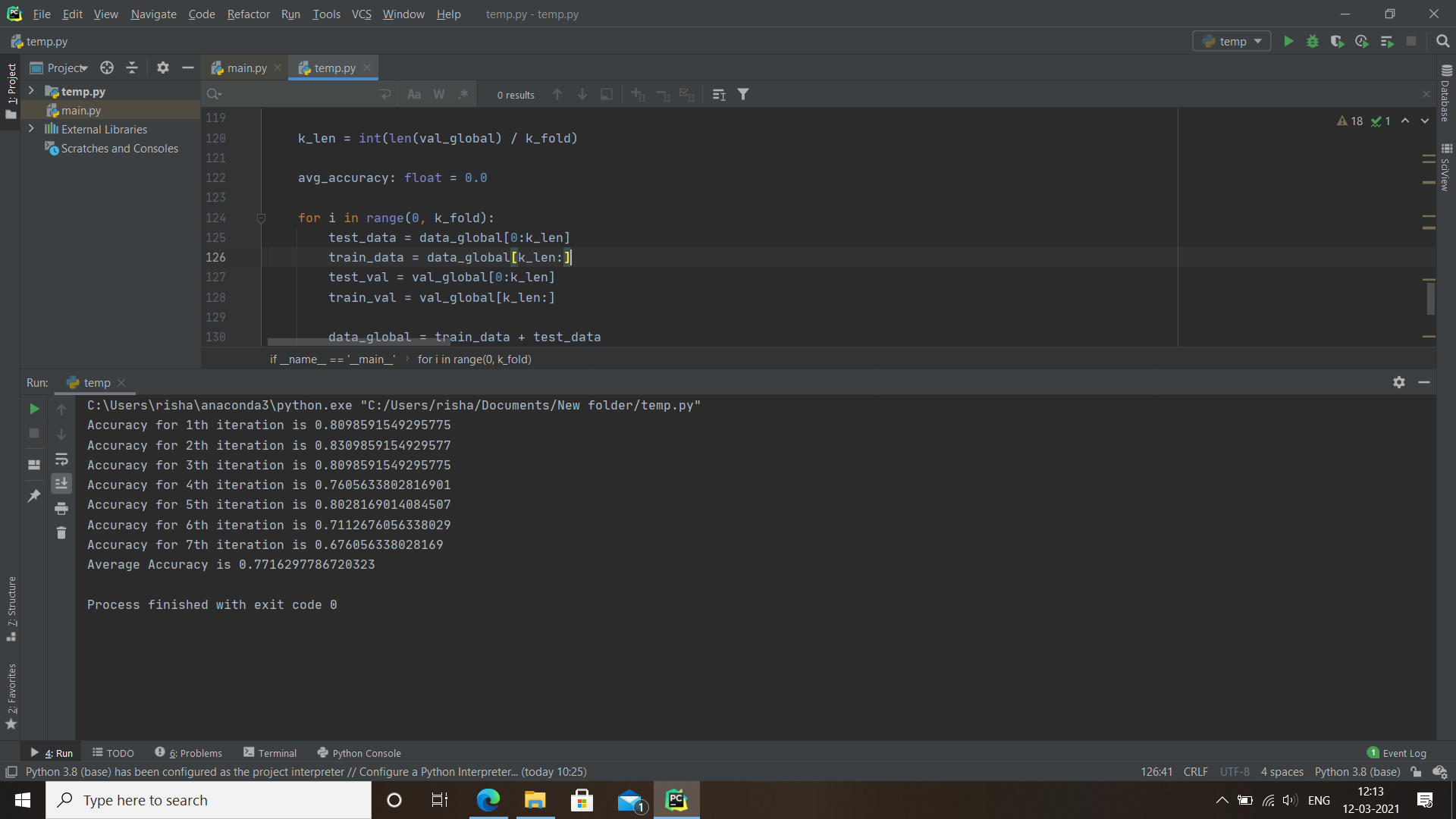
We use the above mentioned formula for getting the probability of being spam or not and then whichever is the higher decides the nature of mail.

P(vj) is calculated as (Total training example of vj)/(Total training examples)

Here ai represents the word found in the test sentence. (similar to w1,w2,etc)

Moreover to avoid the problem of probability becoming zero P(ai|vj) getting zero we use laplacian smoothing which in the book is given as. Here we actually use the value of alpha as 1 thus instead of (nk + alpha)/(n+alpha\*|vocabulary|) we get this

Q2. Accuracy of your model over each fold and the overall average accuracy.



Q3. Major limitations of the Naive Bayes classifier.

The very first limitation is due to its simplistic nature. We take an assumption of independence, although this makes the model very easy to compute, the assumptions itself as logically not sound. Seldom do we see any sort of independence of features as taken in this classifier. Thus due to this independence we see very poor average accuracy of 77.16%.