Email Classification

Aim

Email users spend a large amount of time regularly deleting spam emails, which needlessly occupy storage space and consumes bandwidth. Hence the development of classifiers that are able to separate legitimate from spam emails is required. The aim of this study is to examine spam detection using the Naive Bayesian technique.

With email dominating as the main form of communication, spam detection to remove spam is important as it is a global issue that negatively affects productivity and uses up valuable resources.

Data Preprocessing and Feature Selection

Subject vs Body: Results and Discussion

Corpus: Subject	
	Accuracy [%]
ZeroR	
OneR	
1-NN	
3-NN	
NB	
DT	
MLP	
MyNB	

Corpus: Body	
	Accuracy [%]
ZeroR	
OneR	
1-NN	
3-NN	
NB	
DT	
MLP	
MyNB	

Challenge Results and Discussion

Conclusions

The tokens used in our classifier are formed from single words. Therefore, it will not analyse common consecutive words that are found in spam emails, leading to the failure of our classifier from detecting these emails.

By taking into account permutations of consecutive words, or words that appear within a specified distance of each other, the accuracy of our Bayesian classifier could be increased.

Reflection

Instructions: How to run code