LAB-3

HIMA BINDU KRONUIDI
(HK4233)

DESCRIPTION OF FEATURES :

Based on trained data 5 Features has been used to initialize.

...> Features have been choosen based on frequency, i.e most common occuring words

i) common coords in english like "which", "if", "is", "at", "then" ... etc

ii) Dutch alphabets that are unique ä, ë, ï, ö, ü

iii) Dutch numbers like "een", "twee", ...

iv) English numbers like "one", "two", ...

v) Dutch common words "goed", "zij", ...

DECISION TREE :

i) Four parameters have been used level, maximum level, already visited nodes and output file

ii) Level with one has been treated as root node. Visited node has been initialized. Using the total count of features a matrix has been created using which entropy of each column has been calculated.

iii) A decision tree will be created based on the entropy values. A visited set is also maintained to avoid redudency.

iv) Maximum level of 5 has been used as operation beyond level 5 has not produced anything different.

v) Decision tree outputs have been included in the examples folder within the folder submitted.

ADABOOST :-

i) Condition being consider as :

$$\text{Weight} \propto \frac{1}{\text{length of data frame}}$$

Weight is inversely proportional to length of dataframe.

ii) Here 10 decision tree have been created, trained and based on that it guess either true or fake which leads us to the actual answer.

iii) Weight values will be updated based on the error value calculated.

iv) If the decision tree gives the expected output, the updated weight will be less than the actual weight, where as the updated weight will be more for corong decisions.

v.) This process keeps operating for all the decision trees. Error rate's will be updated based on the training of every decision tree, and based on the error rates ada-boost gives the output through the comparsion of values that is english or Dutch. Whcve all the decision trees conclusively provide adaboost result.