Foundations Artificial Intelligence
Lab - 2

Byreddy Vishnu Saketh

Running the code:

The code is divided into two parts : train.py , predict.py

To run train.py : **python train.py <trainData> <modelName>**

Ex : python train.py train.dat DTreeModel

The model is saved as a .pkl file

To run predict.py file : **python <modelName.pkl> <testData>**

Ex : python predict.py DTreeModel.pkl test.dat

Each prediction is printed in a new line.

Feature Selection:

- AverageWordCount > 5.1: The average word count of commonly used words English language is 4.9 - 5 while in Ducth it is slightly higher.
- Letter ‘e’ frequency > 15%: Letter E frequency in Dutch words is 19% whereas in English it is 11%. Source : https://www.sttmedia.com/characterfrequency-dutch
- Letter ‘n’ frequency > 8.5%: Letter N frequency in Ducth words is 10% whereas in English it is 8.5%.
- Does it contain the word ‘the’ : ‘The’ is the most commonly occurring word in the English language.
- Does it contain the word ‘de’ : ‘De’ is the most commonly occuring word in the Ducth language.
- Do words have two same letters consecutively: The Dutch language frequently contains words which have two letters back to back. For example : gepubliceerd has two e’s.
- Do words contain letters ‘ij’ in them: The Dutch language frequently contains words with letters ‘ij’ together.
- VowelCount > 14: On an average the words in Dutch language contain more vowels than the words in English language. Especially due to the high occurrence of the vowel ‘e’.
- Does it contain the letter ‘q’: Dutch language has a very few words with the letter q in them, whereas English language has comparatively better number of words with q.
- Does it contain words of length 1: English language uses a lot of length 1 words whereas the Ducth language does not.

Decision Tree:

A Decision Tree was built using all the above features with **max_depth as 7.** Each node was selected by using information gain i.e, attribute with the highest info gain was selected at each level.

Useful features based on Information Gain:

1) Does it contain the word the
2) Do words contain letters 'ij' in them
3) Does it contain the word de
4) Does it contain words of length 1
5) Letter 'e' frequency > 15%
6) AverageWordCount > 5.1
7) Letter 'n' frequency > 8.5 %

Some of the other features were not getting selected when using a train data of 115 sentences.

With my test data of 20 sentences the model was classifying 19/20 sentences correctly.

AdaBoost:

Decision stumps were created by selecting the best feature which classified the data based on weights. Weights were consecutively updated with each attribute by calculating the significance value from the weights of the incorrect predictions.

For number_of_estimators(decision stumps) = 8, the algorithm had optimal performance.

The following attributes were selected in order of their significance value:

1) Does it contain the word the
2) Do words contain letters 'ij' in them
3) Do words have two same letters consecutively
4) Does it contain the word de
5) Does it contain words of length 1
6) Letter 'e' frequency > 15%
7) Letter 'n' frequency > 8.5 %
8) VowelCount > 14