CSCI 630 Foundation of Artificial Intelligence

Lab 2 : Wikipedia Language Classification

Sarthak Thakkar (st4070)

1. Introduction:

In this assignment, We had to write a code to create language classification model to classify two languages ‘English’ and ‘Dutch’ based on the training dataset and testing data set formats.Here we’ve implemented 2 approaches ‘Decision Trees’ and ‘ADA boosting with decision trees’.

The code is written in python 3.The usage of the code can be stated as :

```
python3 main_class.py train <training_file> <hypothesis_file>.oj
<learning_type>
```

- Training_file : Consists of labelled examples for making a decision tree.
- Hypothesis_file : Consists of serialized object to store mode generated by the given training data.
- Learning_type : Specifies the type ‘dt’ for decision trees and ‘ada’ of ada boost approach.

```
python3 main_class.py predict <hypothesis_file>.oj <test_file>
```

- Hypothesis_file : Consists of generated model used to be used by prediction program to predict language of given statements
- Test_file : Consists list of unlabelled sentences to be classified.

2. Description of Features and Justification :

There are 11 boolean attributes used to train our model the description of each attribute and how they are treated as Boolean is given below:

2.1. Repeated Constants:

In Dutch statements while collecting data I observed that many dutch words contain repeated consonants like ‘pp’,’tt’,’kk’,’rr’ more often then they are observed in a general english statement.
So I selected this as attribute.If there are two or more pairs of consonants in a statement than this flag will be *True*.

2.2. Repeated occurrence of ‘ee’:

The usage of ‘ee’ was also observed to be more frequent in words of dutch language.
So if the occurrence of ‘ee’ was observed in a statement this flag would be *True*.

2.3. Repeated occurrence of ‘aa’:

The usage of ‘aa’ was also observed to be more frequent in words of dutch language.
So if the occurrence of ‘ee’ was observed in a statement this flag would be *True*.

2.4. Occurrence of ‘ij’:

Dutch statements contained frequent use of ‘ij’ which is very less likely to be observed in english statements except few words like ‘hijack’ etc.
So if ‘ij’ was to be detected in a statement it would be more likely to be a Dutch statement and the flag would be *True.*

2.5. Ending with ‘en’:

The frequency of words in dutch words ending in ‘en’ was also observed to be higher compared to the frequency of english words.
So if a statement contains two or more words ending in ‘en’ the flag would be marked as *True.*

2.6. Repeated occurrence of ‘oo’:

The usage of ‘oo’ was also observed to be more frequent in words of dutch language.

So if the occurrence of ‘ee’ was observed in a statement this flag would be *True*.

2.7. Ending with ‘ing’:
The frequency of words in english words ending in ‘ing’ was observed to be higher compared to the frequency of dutch words.
So if a statement contains words ending in ‘ing’ the flag would be marked as *True.*

2.8. Non-alphanumeric characters :
I observed that there were some non alpha numeric characters present in dutch words which are not found in english sentence. So words containing non alphanumeric characters are more likely to be dutch sentences.
So If a non-alphanumeric character is found this flag is marked *True.*

2.9. Vowel to consonant ratio :
While testing for multiple english statements i observed the Ratio of vowels to consonants was more than 0.7 in Dutch and was in between 0.5-0.7 in English statements.
So if the Vowels to consonants ratio was 0.7 or higher this flag would be marked *True.*

2.10. Ending in ‘ae’:
The frequency of words in dutch words ending in ‘ae’ was also observed to be higher compared to the frequency of english words.
So if a statement contains two or more words ending in ‘ae’ the flag would be marked as *True.*

2.11. Repeated occurrence of ‘de’:
Dutch statements contained frequent use of ‘de’ which is very less likely to be observed in english statements as ‘de’ is article for *the* in dutch .
So if dej’ was to be detected in a statement more than once it would be more likely to be a Dutch statement and the flag would be *True.*

3. Decision Tree Approach and results:

For Decision Tree based learning approach. First I made a sentence object and stored values of all the attributes in the respective objects.After getting attribute values of sentences I started training of model by creating the decision tree by

find the best split attribute based on the information gain from the split using the formula

$H(V) = \sum_K P(v_k) \log_2(1/ P(v_k))$

On training data recursively till we meet the three terminating conditions

- There are no more attributes to split.
- There are no sentences in a specific attribute split and have same tag.
- There are no more sentence data to train and form a combination of attributes which might be possible if we had more data to train.

Thus, By using this approach I implemented decision tree based language classifier and results for different depth were as follows.

Depth	Accuracy
2	0.75
4	0.75
10	0.75
12	0.75
16	0.83
20	0.83

4. ADA boosting Approach and results:

For ADA boosting using decision tree instead of using multilevel complex decision trees we use the decision trees of single attribute (stumps). Combine them according to the weight and their ability to give the correct output. Approach we follow is similar, we first create sentence objects and find the best split for even weights and then look for further splits based on weights updated by previous splits and look for the appropriate splits to create an Ensemble to be used for prediction by combining the stumps based on weight and order.

No. of stumps	Accuracy

2	0.75
3	0.75
4	0.75
5	0.75
10	0.75