Language Classification

Feature List:

1. Definite articles in Dutch - De, and Het
   Definite articles are commonly found in the Dutch language.

2. Prepositions in English - in, at, on, of, to
   A sentence in English generally comprises prepositions.

3. Article in English - The
   Only one definite article in English, thus usually found in sentences.

4. Consecutive vowels in Dutch - aa, ee, ii, oo, uu
   The dutch language generally consists of consecutive vowels in words.

5. Consecutive letters in Dutch - ij
   The use ‘ij’ in the Dutch words in generally present.

6. Conjunctions in English - and, or, but, yet
   Conjunctions are used to combine the sentences in English and thus found with a high probability.

Decision Tree:

A. Train

1. Read the input from the command line
2. Generated 6 features to test the input lines and generate a boolean results
3. Passed input line(each row) through the Feature test, generated 6 boolean values for each input line and inserted them into a 2D ArrayList
4. Calculated the Remainder value for each column of the 2D List
5. Picked the column with the minimum remainder value and set it as a root of a binary tree
6. Split the above column into two lists according to boolean values - true and false
7. Updated the 2D ArrayList according to the boolean values above and recursively called the true splits and false splits on points 4, 5, 6 and 7
   Considerations -
   Left Branch - True
   Right Branch - False

8. The maximum depth of the decision tree is each when the minimum value of the remainder is 1 and the count of nl and en is stored at the leaf nodes
9. Thus a complete Decision Tree (binary tree object) is created
10. The Decision Tree Object is serialized and saved

B. Predict

1. The trained Decision Tree model is tested on a Test Data Set
2. The serialized object is deserialized and the input Test file is read one line at a time
3. Each input line(each row) is passed through the feature test and 6 boolean values are generated for each row
4. The boolean value is selected on the basis of the Decision Tree root node and follows the path till it reaches the leaf node
5. Prediction for each line is displayed on the basis of the count of nl and en at the leaf node. The one with the greater count is displayed
6. The process in points 3, 4, and 5 are repeated until the prediction is made for all the lines in the test data

C. Results

The Decision Tree model was trained on a training data set and tested on a test data set of 200 lines.(100 English and 100 Dutch)
Below are the predictions -
English 100 lines
Correct - 93
Incorrect - 7
Accuracy - 93%

Dutch 100 lines
Correct - 97
Incorrect - 3
Accuracy - 97%

Adaboost

A. Train

1. Read the input from the command line
2. Generated 6 features to test the input lines and generate a boolean results
3. Passed input line(each row) through the Feature test, generated 6 boolean values for each input line and inserted them into a 2D ArrayList

4. Calculated the Remainder value for each column of the 2D List
5. Assigned equal weights initially to all the example according to the total number of examples
6. Picked the column with the minimum remainder value
7. Calculated the error count for the column, updated the weights of each row by normalization(normalized weights)
8. Created a Decision Stump, calculated the alpha value(weight factor of each stump) for the respective Decision Stump and stored it in linkedlist object
9. Updated the dataset according to the cumulative weights
10. Continued the process in 4, 5, 6, 7, 8, and 9 for a few iterations

B. Predict

1. The trained Adaboost model is tested on a Test Data Set
2. The serialized object is deserialized and the input Test file is read one line at a time
3. Each input line(each row) is passed through the feature test and 6 boolean values are generated for each row
4. The generated boolean values are passed through the stumps and the total value is calculated using alpha of each stump and the labels nl and en are displayed accordingly

C. Results

The Adaboost model was trained on a training data set and tested on a test data set of 200 lines.(100 English and 100 Dutch)
Below are the predictions -
English 100 lines
Correct - 97
Incorrect - 3
Accuracy - 97%

Dutch 100 lines
Correct - 93
Incorrect - 7
Accuracy - 93%