Write Up – Lab 2

Kriti Sharda

Data Collection:

- Data was collected from different articles on Wikipedia page for both the languages i.e. English and Dutch.
- Paragraphs from different articles were copied into a text file. It was mad sure that a huge number of data is available for training.
- Python Script was used to generate training and testing data that would be given as an input to the classifier.
  Python Script: generateTrainTestData.py
- For training data, each line was appended by the language it belonged to. So, for english, 'en|' was appended to the beginning of the line and for dutch data, 'nl|' was appended to the beginning of the line
- For the testing data, each line contained just the words. The length of a lined was either 10 words, 20 words or 50 words
- Decision Tree and Adaboost Stump both use these training and testing files to train the model or predict using the trained model respectively.

Feature Collection:

Following are the features that I used to check if the language of the given data is in English or Dutch. A total of 11 features have been used.

- Check whether 'het' is present in the record. This is a dutch word, which is equal to the article 'the' in English. It appears in the dutch text very often. Hence, If we come across this word in the record. We can say that the language might be Dutch.
- Check whether 'de' is present in the record. This is a dutch word, which is also equal to the article 'the' in English. It appears in the dutch text very often just like the keyword 'het'. Hence, If we come across this word in the record. We can say that the language might be Dutch.
- Check whether 'een' is present in the record. This is a dutch word, which is equal to the article 'a' in English. It appears in the dutch text very often. Hence, If we come across this word in the record. We can say that the language might be Dutch.
- Check whether 'en' or 'aan' is present in the record. Both of these are dutch words, which are equilant to the words 'and' and 'to' in English language

respectively. It appears in the dutch text very often. Hence, If we come across this word in the record. We can say that the language might be Dutch.

- Check whether ‘ij’ is present in the words. I observed that a lot of dutch words contain the characters ‘ij’ together. If the count of these words is more that 2, then we can say that the language might be Dutch.
- Check whether length of the words are greater that 13. Dutch words usually have a long length. So, Length of all words in the record is calculated and if 3 or more words are found with length greater than 13, then we can say that the language might be Dutch.
- Check whether ‘a’ and ‘an’ are present in the record. These are the articles in the English language and occur a lot in the text. Hence, If we come across these word in the record. We can say that the language might be English.
- Check whether ‘are’ and ‘were’ are present in the record. These are the commonly occurring words in the English language and occur a lot in the text. Hence, If we come across these word in the record. We can say that the language might be English.
- Check whether ‘and’ are present in the record. This is the word that joins sentences in English and occurs a lot in the text. Hence, If we come across these word in the record. We can say that the language might be English.
- Check whether ‘on’ and ‘to’ are present in the record. These are the prepositions in the English language and occur a lot in the text. Hence, If we come across these word in the record. We can say that the language might be English.
- Check whether ‘the’ is present in the record. This is the articles in the English language and occurs in a lot in the text. Hence, If we come across these word in the record. We can say that the language might be English.

Decision Tree:

- First the Training of the decision tree is done and a decision tree model is found. Then prediction is performed using the decision tree model and final classes of the test set it printed on the console.
- For training, the input text file is given to the algorithm. First, the file is broken into records which is a list depicting multiple lines. Each line has the class assigned to it followed by the words present in that line
- Next, features are extracted from each record. Finally a categorical record is obtained which contains a ‘True’ or ‘False’ value for each feature.
- This categorical data is sent to train and build the decision tree.
- The algorithm calculated the Class Entropy and entropy for each attribute. Then Information gain is calculated for each attribute value. The attribute, which has the highest information gain, is used to split the data. Further, new

sub-decision trees are build for the split data and the algorithm runs in recursion.

- There are 2 main stop condition in the algorithm. First, when a pure leaf node is reached, which says whether the language is English or Dutch. Second, when the algorithm runs out of attribute to split or data to test. In both these cases, the class or language that is present for maximum number of records is returned as the predicted language of the record.
- For Prediction, the above decision tree model is used to trace the path that leads to the final class prediction.

AdaBoost:

ALGO:

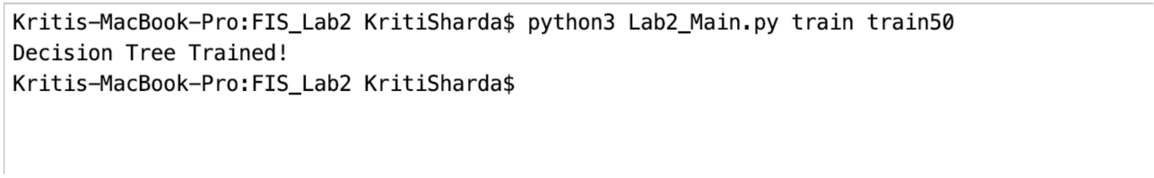
- This boosting algorithm works by assigning weights to each record, while calculating the information gain. Multiple stumps of the decision tree are generated by using weights that get updated in order to increase the accuracy of the classifier.
- The initial weights assigned are 1/number of records in the data. Then all the correct and incorrect prediction is calculated and the error rate is calculated.
- Weights are increased for the records that were classified incorrectly and decreased for the records that were classified correctly.
- The new weights are normalized so that they sum up to 1. Finally these weights are used to update the decision tree and this process is repeated several times to increase the accuracy of the decision stump.

Data Length:

- It was observed that when the length of the line decreased, the classifier made more erroneous predictions as compared to the data when more length of line is considered.
- The result for data that contained 10 words per line was the lease accurate because naturally less words are covered in each record and for this reason less features are learnt by the classifier.
- The result for data that contains 20 words per line showed better accuracy as compared to 10 words per line. This data was able to train the classifier better that the previous data.
- The data that contained 50 words per line trained the model in the best way out of the three.
- The number of lines used to train the classifier also matters. More the number of training records, better is the accuracy of the model.

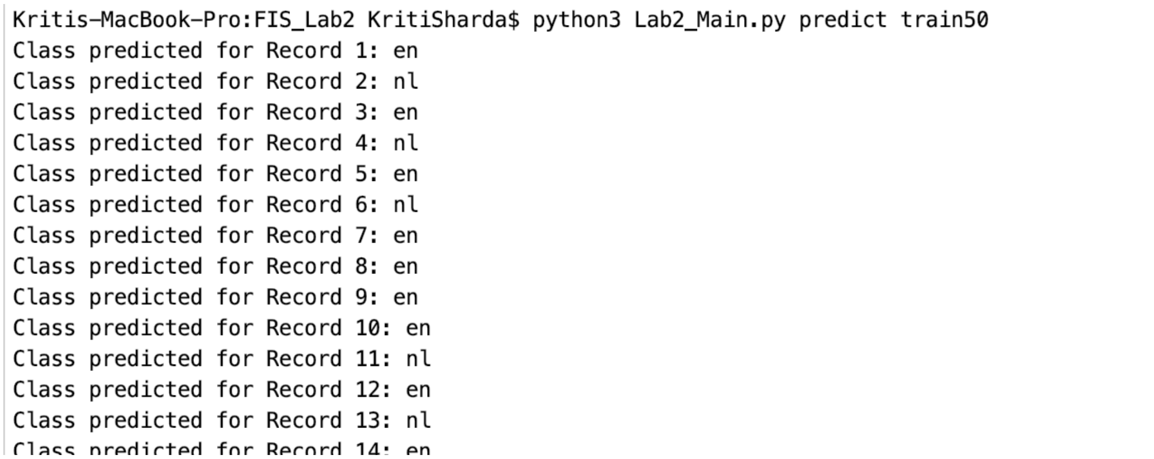
Sample Outputs:

Training:



```
Kritis-MacBook-Pro:FIS_Lab2 KritiSharda$ python3 Lab2_Main.py train train50
Decision Tree Trained!
Kritis-MacBook-Pro:FIS_Lab2 KritiSharda$
```

Prediction:



```
Kritis-MacBook-Pro:FIS_Lab2 KritiSharda$ python3 Lab2_Main.py predict train50
Class predicted for Record 1: en
Class predicted for Record 2: nl
Class predicted for Record 3: en
Class predicted for Record 4: nl
Class predicted for Record 5: en
Class predicted for Record 6: nl
Class predicted for Record 7: en
Class predicted for Record 8: en
Class predicted for Record 9: en
Class predicted for Record 10: en
Class predicted for Record 11: nl
Class predicted for Record 12: en
Class predicted for Record 13: nl
Class predicted for Record 14: en
```