Lab 2 Wikipedia Language Classification

-Sidhartha Gopi Amperayani

INTRODUCTION:

This lab explores the use of decision trees and Adaboost decision stumps to classify given sentences into their respective languages.

FEATURES:

Following are the features used:

1. Common Dutch words.
2. Common English words.
3. Presence of ‘aa’ in the word, since many dutch words contain this sequence.
4. Presence of ‘ij’ in the word, since many dutch words contain this sequence.
5. Presence of ‘oo’ in the word, since many dutch words contain this sequence.
6. Presence of ‘ee’ in the word, since many dutch words contain this sequence.
7. Presence of ‘q’ in the word, since many dutch wordsdo not contain this letter.
8. Presence of ‘de’ in the word, since many dutch words contain this sequence.
9. Presence of ‘een’ in the word, since many dutch words contain this sequence.
10. Presence of ‘en’ in the word, since many dutch words contain this sequence.
11. Presence of ‘van’ in the word, since many dutch words contain this sequence.

DECISION TREE:

Using entropy and information gain calculate the best feature which divides the given sentences into Dutch and English. Make that attribute the root node.

Similarly, for both left and right tree find the next best attribute which classifies the given sentence’s language.

AdaBoost:

A machine learning algorithm which improves the results. Uses decision stumps, which essentially means decision tree with depth 1 (in this lab). Adaboost decision stumps reweights the decision tree and produces efficient results.

STEPS TO RUN:

1. enter 'train' or 'predict’.
2. If train; enter the training file path.
3. enter hypothesis file path
4. enter 'dt' for 'decision tree' or 'ada' for 'adaboost.

After step 4 the training is completed.

Now rerun the program and enter ‘predict’

5. enter the hypothesis file path, same path as training.
6. enter testing file path.

Results of either Decision Tree or AdaBoost are returned as output