WIKIPEDIA LANGUAGE CLASSIFIER

By Ravikiran Jois Yedur Prabhakar

- *Description:*
    - This is a program that classifies an input file consisting of English and Dutch sentences with 15 words in it into its respective language

- *Usage*:
    - The training dataset has to be used to train the Decision Tree and the Adaboost
    - The command line argument for Decision Tree:
        - Training:
            - train train_2500.txt hype_out_dt dt
        - Predict:
            - predict hype_out_dt test.dat

    - The command line argument for Adaboost:
        - Training:
            - train train_2500.txt hype_out_ada ada
        - Predict:
            - predict hype_out_ada test.dat

    - While using the predict method, the user will be prompted to enter the option for predicting the data i.e., if the user wants to predict the results of the Decision Tree, the user has to enter 'dt' and if the user wants to predict the results of the Adaboost, the user has to enter 'ada'

    - The training has to be executed before the predict execution so that the node written to the 'hype_out' by the training method is fed to the predict method

- *Features used:*
    - *Definite articles:*
        - The words 'de' and 'het' are definite articles in Dutch
        - Thus, these words are very frequent in the Dutch sentences
    - *Vowels that occur together:*
        - Words in Dutch commonly have vowels that appear together like 'aa', 'ee' and 'oo' in them
    - *Average length of words:*
        - The average length of words in a sentence in Dutch is greater than that in English
        - In this exercise, the average length of words in a sentence for Dutch is taken to be greater than 7
    - *Common words in Dutch:*
        - The common words in Dutch are taken into consideration here
        - The words considered in this exercise are: 'van', 'dat', 'en' and 'niet'
    - *The words 'You' and 'I' in Dutch:*
        - The word for 'You' in Dutch is 'je' and 'I' in Dutch in 'ik'
        - These words are common in any language and so, this has been taken into consideration
    - *The sentence with word with 'ij':*
        - The words in Dutch commonly have the letters 'i' and 'j' occurring together, like 'ij'
        - In English words, we do not find this pattern commonly
    - *Common words in English:*
        - These are the words that occur very commonly in English: 'the', 'I', 'a', 'be', 'and', 'to' and 'of'
        - These words are not as frequent in Dutch
    - *Words that begin with 'z' in it:*
        - In Dutch, it is common to have words that begin with 'z'

        - In English, however, this is far less common
    - *Words containing 'oe' or 'cht' in it:*
        - In Dutch, the words commonly contain 'oe' or 'cht' in it
        - This is not a frequent occurrence in English

- ***Decision Tree:***
    - For creating the decision tree, the information gain is calculated based on the entropy of the attributes
    - For each attribute, the entropy is calculated, and the information gain is calculated for each attribute
    - The attribute with the highest information gain is selected and the node is created with that attribute's name
    - For the next recursion step, that particular attribute would not be in consideration for the calculation of information gain
    - This, at the end, creates a decision tree
    - The training data is used to train the decision tree and the node is written to the file entered in the command line argument using the pickle package
    - The testing data makes use of the node object written by the trainer and the predict method prints out the predicted language per line

- ***Adaboost:***
    - The weights for each line of the training data is calculated and assigned
    - The entropy and information gain are calculated for each of these attributes and this calculation, unlike that in case of decision tree, is based on the weights
    - For the attribute with the highest information gain, the true and false branches are created
    - The above steps are performed for *N* number of times, *N* being the number of stumps that is to be created
    - N has to be less than the length of the set of attributes
    - The set of weights and stumps are sent to pickle to write the node object to a file
    - This file is sent to the predict method which prints out the predicted language for each line in the test file
    - The most useful attributes are: (In descending order of usefulness: most useful being first in list which is based on the correct prediction for the given training and test data attached)
        - *definite article 'de'*: 100%
        - *word contains 'ij'*: 90%
        - *vowels in a word together: 'aa'*: 80%
        - *definite article 'het'*: 80%
        - *common Dutch word 'van'*: 70%
        - *common Dutch word 'dat'*: 70%
    - These percentages are calculated manually by looking at the actual data and comparing it with each stump
    - The test data has 10 lines of English and Dutch

- The dataset used for both Decision Tree and Adaboost training and testing is attached along with the code and report for the Lab 2