Language Classifier

Arjun Ravikumar

Usage

The program can be used as below:

Train

1. **python3 train.py <examples> <hypothesisOut> <learning-type>**

- **examples** is a file containing labeled examples.
- **hypothesisOut** specifies the filename to write your model to.
- **learning-type** specifies the type of learning algorithm you will run, it is either "dt" or "ada".

Predict

2. **python3 predict.py <file> <hypothesis> <model type>**

- **hypothesis** is a trained decision tree or ensemble created by your train program
- **file** is a file containing lines of 15 word sentence fragments in either English or Dutch.
- **model type** specifies the type of algorithm you will run, it is either "dt" or "ada".

Features Selected

To split data, 10 features are used based on online journals and other documents comparing the languages

Feature	Explanation	Expection
Presence ‘a’, ‘and’, ‘the’	English has these alphabets but dutch doesn't contain any of them	True if something present else False
Presence of Dutch articles	Articles specific to dutch	True if nothing present else False
Average word length	Greater than 9 considered to be dutch, English has lower than dutch average	True if less than 9 else false
Most common dutch words	Words common in dutch	False if any of them present else true
Filler words specific to Dutch	These words are not part of english vocabulary	False if any of them present else true
Common dutch letters at the end together	The letters which are frequently found together in dutch at the end but not in english	False if any of them present else true
Combination words in dutch	Dutch has common letters used together english does not	False if any of them present else true
Long words	Dutch has long words more than 8 character long 5 is kept as the	False if any of them present else true

more than 8 character long	threshold here	
Frequent ending characters in dutch	Characters that are frequently found at the end in dutch but not so much in english	False if any of them present else true
Occurence of oo, aa, ii, ee together	Dutch has a lot of these togther but english not so much	False if any of them present else true

Decision Tree

The data consisted of 5300 instances of English and Dutch (2650 each). I found that a larger sample works better for my code as it gives a near 100% correctness for all data. Didn’t limit the depth of the tree as when I tested it got the best results in that case. Limiting case is if the number of rows or columns are minimal for calculation.

Adaboost

The data consisted of 20 English and Dutch lines of (15 words each). Smaller sample less than 30 works better on the code as the number increases precision also decreases error in the data. The entropy gain was used in the adaboost technique.
50 classifiers were considered for the adaboost as I found that this number is optimal.

For training please use the testData.dat for Decision tree as hypothesis. (The large file mentioned above)
And use any small file for training Adaboost.

Best hypothesis file for DT - dt
Best hypothesis file for Adaboost - ad