Sungmin Kim

Intro to IS Lab 2: Wikipedia Language Classification

Features

To split up the data, 12 features were selected based on frequent patterns (common words, etc.) in the training set. 5 other generic features (e.g. number of letter pairs and average word length) were also selected for a total of 17 features.

Feature	Explanation	Justification
cv-ratio	A range. The ratio of consonants to vowels in a sentence.	In some languages, sentences might have a disproportionate number of vowels or consonants.
av-len	A range. The average length of words in a sentence.	Some languages might have longer or shorter average word lengths.
v-pairs	A range. Number of vowel pairs in the sentence. A vowel pair is two consecutive appearances of the same vowel e.g. “ee” or “uu”.	Certain languages seem to have a higher frequency of vowel pairs in sentences, including Dutch.
c-pairs	A range. Number of consonant pairs in the sentence. A consonant pair is two consecutive appearances of the same consonant e.g. jj or pp	Certain languages might have a higher frequency of consonant pairs in sentences.
l-pairs	A range. Number of letter pairs in the sentence. A letter pair is a consonant or	Certain languages might have a higher frequency of

	vowel pair e.g “oo” and “qq” are letter pairs.	letter pairs in sentences.
ends-en	Boolean. Does the sentence contain a word that ends in “en”?	A lot of Dutch sentences contain words that end in “en”.
ends-e	Boolean. Does the sentence contain a word that ends in “e”	A lot of Dutch sentences contain words that end in “e”.
has-aa	Boolean. Is “aa” a substring of the sentence?	“aa” tends to be a substring of many dutch sentences.
has-ee	Boolean. Is “ee” a substring of the sentence?	“ee” tends to be a substring of many dutch sentences.
has-word-het	Boolean. Does the sentence contain the word “het”?	Many dutch sentences contain the word “het”.
has-word-een	Boolean. Does the sentence contain the word “een”?	Many dutch sentences contain the word “een”.
has-word-en	Boolean. Does the sentence contain the word “en”?	Many dutch sentences contain the word “en”.
has-word-de	Boolean. Does the sentence contain the word “de”?	Many dutch sentences contain the word “de”.
has-word-the	Boolean. Does the sentence contain the word “the”?	Many English sentences contain the word “the”.
has-word-and	Boolean. Does the sentence contain	Many English sentences

	the word “and”?	contain the word “and”.
has-word-in	Boolean. Does the sentence contain the word “in”?	Many English sentences contain the word “in”.
has-word-of	Boolean. Does the sentence contain the word “of”?	Many English sentences contain the word “of”.

Decision Tree

The data consisted of 102 instances of English and Dutch (51 each), set aside exclusively for training and 32 instances of both languages (16 each), used exclusively for testing. The decision tree accepts a parameter “depth” which is an upper bound for the depth of the tree. I experimented with trees of varying heights and checked their error rates. A minimum depth of 5 was required to correctly classify all the training and test data. I decided to go with a tree that was 8 levels deep to possibly account for variance in a larger test set.

Adaboost

The data used for the decision tree was also used to train and test the boosting model. In this case, however, a weight was assigned to each instance of the data, forming a distribution of weights. The boosting trainer accepts a single parameter “ensemble_size” which is the number of decision stumps to be used in the ensemble. I experimented with various ensemble sizes and checked the error rates. An ensemble of 5 decision stumps was enough to accurately classify the training and test data. I went with 100 to account for larger test sets.