CSCI331

Lab2 Writeup

Lyan Ye (yxy6465)

Example Collection:

1. The training data is extended by 60 more lines with total 70 examples of English/ Dutch sentences (35 each).
2. The testing data is the original file downloaded from the website of the lab guideline.

Although we all know that the more training data we collect, the more accurate prediction model we will get, I had tried my best to collection the training data since there is no sharing data on the discussion post.

Features

1. Boolean: usage of the word “and” in daily English sentences
2. Boolean: usage of the word “en” in daily Dutch sentences
3. Boolean: usage of the word “the” in daily English sentences
4. Boolean: usage of the word “de” in daily Dutch sentences
5. Boolean: usage of the word “enn” in daily Dutch sentences
6. Boolean: usage of the word “het” in daily Dutch sentences
7. Boolean: contains the substring of “ij” in daily Dutch words
8. Range: Words in Dutch tend to be longer than words in English
9. Range: Frequency of the usage of double vowels (consecutive two same vowels) in Dutch is more than in English
10. Range: Frequency of the usage of double consonants (consecutive two same consonants) in Dutch is more than in English
11. Range: Frequency of letters in words such as “j, k, v, z” in Dutch is more than in English

Decision Tree Learning

* A decision tree model was built using the training data with 70 example entries(sentences).
* Entropy is used to find the impurity for each level of classification.
* The information gain algorithm is used to classify the entries by the best feature for each level as well.
* A maximum depth of 15 was set in order to handle larger training data and testing data. However, for my default training set, I found out that 5 is enough to generate a good trained model.

Adaboost

* A boosted ensemble model was built using the same training data, and a weight for each entry(sentence) was assigned in order to use the Adaboost algorithm to adjust the weight of each entry before going to the next stump.
* A maximum size of stumps was set to 5 and it is enough for my default training data.

To run the program:

Default training and testing files locate in data directory, and best decision tree model called tree.o and best adaboost model called ensemble.o were generated in out directory. The output of the program while using the predict action is simply the predicted label(en/nl) for each sentence from the test file.

* *To train*: **python classify.py train <examples> <hypothesisOut> <learning-type>**

P.S. for entering files, full location is necessary

Example：classify.py train data\train.dat out\tree.o dt

* *To predict*: **python classify.py predict <hypothesis> <file>**