Wei Zeng
Report of Lab2
11/27/2018

Report of Lab 2

1. Data collection

   In this lab, I selected sentences from 10 English articles and 10 Dutch articles as samples for every training file. Thus, each training file has 20 instances. Then I selected "Does it contain the letter Q/X/Y?", "Does it contain Is/The/Van/Voor/De/It/And" as features to use for learning. Because letters Q/X/Y are used in English commonly but barely used in Dutch, I selected "Does it contain the letter Q/X/Y?". Similarly, the words Is/The/It/And are usually used in English and the words Van/Voor/De are used in Dutch, so I selected "Does it contain Is/The/Van/Voor/De/It/And". However, I found the word Is is used in both English and Dutch even in almost the same frequency after finish data collection, so the feature "Does it contain Is" does not work in this lab.

2. Experimentation

   To find out the model, including decision tree and decision stumps, which gives the best results on my test data. I tried to adjust training instances for many times. Firstly, all instances I found in English articles contain the letter Y and all instances in Dutch articles do not contain, so the model has only one feature "Does it contain Y". Apparently, the model can not make a correct decision about whether an article is English or Dutch if an English article does not contain Y. After using some special instances to replace some approximate instances, the model is strong enough to judge the type of article.

3. Exploration

   To explore the performance of two types of models which are generated by using 10, 20 or 50 words training file on short and long test articles, I collected 10 English articles and 10 Dutch articles which include more than 50 words as long test file, and another 10 English and 10 Dutch article which include less than 20 words as short test file. Then test these files on decision tree and decision stumps separately. Finally, I use the following table to record the accuracy of predictions which are tested under different circumstance.

   Accuracy of Decision Tree

	10 words training file	20 words training file	50 words training file
Long Test File (words ≥ 50)	100%	90%	100%
Short Test File (words < 20)	80%	80%	70%

   Accuracy of Decision Stumps

	10 words training file	20 words training file	50 words training file
Long Test File (words ≥ 50)	100%	100%	100%
Short Test File (words < 20)	100%	100%	100%

According to the tables, the general performance of model of decision stumps is better than decision tree especially on short test file. Besides, the model of decision tree has low performance on short test file no matter which training file the model selected for learning.