Tim auf der Landwehr
t.auf.der.landwehr@student.rug.nl
S2548682
Last revision: December 10, 2013

Learning from Data
Assignment 3 - review

Summary

This is an review of the assignment 3 in the course Learning from Data at the University of Groningen.

Review Assignment 3: ned classification task

I reviewed my feauture selection for the task in assignment 3. Attached you find the trained model in the file `model.weka`. The arff file can be generated from the test data using the script `ned_arff_generator_v3.py`

```
python3 ned_arff_generator_v3.py --gen-arff ned.testa ned.testa.arff
```

On basis of this model the arff file can be testet in weka using the following command:

```
java -cp weka.jar weka.classifiers.lazy.IBk -c 1 -l model.weka -T ned.testa.arff
```

This leads for the given ’development-test-data’ in `ned.testa` to the following results:

```
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

=== Error on test data ===
Correctly Classified Instances    1528              74.1748 %
Incorrectly Classified Instances   532              25.8252 %
Kappa statistic                      0.6371
Mean absolute error                  0.1294
Root mean squared error              0.3593
Total Number of Instances         2060

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 395  11  20  53 |   a = LOC
  39  86  26  41 |   b = MISC
  85  46 464  91 |   c = ORG
  62  23  35 583 |   d = PER
```

Knowing this test data, I managed to improve my results on it by 4 percent. This is still not very satisfactory. When I compare the data by the distribution of the classes, I cannot make out a huge difference, so this can’t be the issue:

```
Normalized distribution of classes
ned.train: [(’LOC’, 28), (’MISC’, 10), (’ORG’, 18), (’PER’, 42)] sum=98
ned.test1: [(’LOC’, 23), (’MISC’, 9), (’ORG’, 33), (’PER’, 34)] sum=99
```

Looking at the data, I get the impression that quite a big number of entities in the training data is extracted from inbetween brackets. See the following examples:

```
LOC Zwi Rominger ( )
LOC Spa Bahamontes ( )
LOC Spa Lorono ( )
LOC Ita Chiappucci ( )
LOC Fra Geminiani ( )
LOC Ita Bartali ( )
LOC Fra Hinault ( )
LOC Ita Massignan ( )
[...]
```

As I only find very few of those in the testa-data, this could have an impact on the preceding and subsequent word feautures. Unfortunately, if I remove all features that use these words, the accuracy goes down to 64.9029%.