CS561 - Assignment-4

2111cs14-2111cs18

Question classification using Decision Tree classifier:

Training Data Sample:

:

	Class	Questions	CleanedQuestions	Length	Words	Lexical	Tagged
0	DESC	How did serfdom develop in and then leave Russ...	how serfdom develop leave russia	9	[how, serfdom, develop, leave, russia]	[how]	[WRB, JJ, VB, JJ, NN]
1	ENTY	What films featured the character Popeye Doyle ?	what films featured character popeye doyle	7	[what, films, featured, character, popeye, doyle]	[what, character]	[WP, VBD, JJ, NN, NN, NN]
2	DESC	How can I find a list of celebrities ' real na...	how i find list celebrities real names	10	[how, i, find, list, celebrities, real, names]	[how, i, find, list, real, names]	[WRB, JJ, VBP, JJ, NNS, JJ, NNS]
3	ENTY	What fowl grabs the spotlight after the Chines...	what fowl grabs spotlight chinese year monkey	12	[what, fowl, grabs, spotlight, chinese, year, ...	[what, chinese, year]	[WP, NN, NN, NN, JJ, NN, NN]
4	ABBR	What is the full form of .com ?	what full form com	7	[what, full, form, com]	[what, full, form, com]	[WP, JJ, NN, NN]

Numerical Value:

Out[63]:

	Length	Lexical	POSTag	Class
0	9	27.631021	10.305454	DESC
1	7	27.631021	10.305454	ENTY
2	10	34.736081	11.891937	DESC
3	12	27.631021	10.305454	ENTY
4	7	35.509271	10.305454	ABBR

In [64]:

```
# test data preparation

category = []
subcategory = []
questions = []
length =[]

with open('TREC_10.label',mode='r',encoding = "ISO-8859-1") as f:
    for line in f:
        #print(line)
        split_index1 = line.index(":")
        split_index2 = line.index(" ")
```

Gini Index:

```
Training Data Report

Accuracy : 68.0

              precision    recall  f1-score   support

        DESC       1.00      0.25      0.40         8
         NUM       0.43      0.34      0.38       118
        ENTY       0.28      0.51      0.36       118
         LOC       0.28      0.28      0.28       130
         HUM       0.38      0.25      0.30        81
        ABBR       0.25      0.14      0.18        91

    accuracy                           0.32       546
   macro avg       0.44      0.30      0.32       546
weighted avg       0.33      0.32      0.31       546

Test Data Report

Accuracy: 51.0

              precision    recall  f1-score   support

        DESC       0.83      0.56      0.67         9
         NUM       0.64      0.85      0.73       138
        ENTY       0.37      0.68      0.48        94
         LOC       0.28      0.28      0.28        65
         HUM       0.42      0.12      0.19        81
        ABBR       0.58      0.27      0.36       113

    accuracy                           0.49       500
   macro avg       0.52      0.46      0.45       500
weighted avg       0.50      0.49      0.45       500
```

Entropy:

```
Training Data Report

Accuracy : 68.0

              precision    recall  f1-score   support

        DESC       0.71      0.62      0.67         8
         NUM       0.39      0.27      0.32       119
        ENTY       0.27      0.50      0.35       116
         LOC       0.33      0.31      0.32       141
         HUM       0.29      0.17      0.21        76
        ABBR       0.41      0.29      0.34        86

    accuracy                           0.32       546
   macro avg       0.40      0.36      0.37       546
weighted avg       0.34      0.32      0.32       546

Test Data Report

Accuracy: 52.0

              precision    recall  f1-score   support

        DESC       0.83      0.56      0.67         9
         NUM       0.64      0.86      0.74       138
        ENTY       0.36      0.62      0.46        94
         LOC       0.27      0.28      0.27        65
         HUM       0.37      0.16      0.22        81
        ABBR       0.60      0.27      0.37       113

    accuracy                           0.48       500
   macro avg       0.51      0.46      0.45       500
weighted avg       0.49      0.48      0.46       500
```

Miss classification:

```
Training Data Report

Accuracy : 68.0

              precision    recall  f1-score   support

        DESC       1.00      0.25      0.40         8
         NUM       0.43      0.34      0.38       118
        ENTY       0.28      0.51      0.36       118
         LOC       0.28      0.28      0.28       130
         HUM       0.38      0.25      0.30        81
        ABBR       0.25      0.14      0.18        91

    accuracy                           0.32       546
   macro avg       0.44      0.30      0.32       546
weighted avg       0.33      0.32      0.31       546

Test Data Report

Accuracy: 51.0

              precision    recall  f1-score   support

        DESC       0.83      0.56      0.67         9
         NUM       0.64      0.85      0.73       138
        ENTY       0.37      0.68      0.48        94
         LOC       0.28      0.28      0.28        65
         HUM       0.42      0.12      0.19        81
        ABBR       0.58      0.27      0.36       113

    accuracy                           0.49       500
   macro avg       0.52      0.46      0.45       500
weighted avg       0.50      0.49      0.45       500
```

Entropy classification is providing better results.