COM6513, Lab7

Registration Number 170224545

17/04/2018

1. Description

In this lab, the program successfully trains and evaluate a structured perceptron with following features:

1. current word-current label
2. current word-current label and previous label-current label
3. current word-current label and previous word-previous label

The program can be executable by running:

python3 lab7.py train.txt test.txt

And the result is as follows:

```
C:\Users\canon\Desktop>python lab7.py train.txt test.txt
First method: The micro-F1 score with the current word-current label feature is:
0.780615586117
Second method: The micro-F1 score with the current word-current label and previous label-current label feature is:
0.690589031821
Third method: The micro-F1 score with the current word-current label and previous word-previous label feature is:
0.732106339468
The top 10 for the first method is:
{'O': ['.', ':', ',', '1', '1996-08-28', '0', '2', '1996-08-22', '1996-08-29', '1996-08-27'], 'PER': ['Peter', 'Younis',
'R.', 'Mark', 'Adrian', 'Yoshikawa', 'Paul', 'Martin', 'Malik', 'Warner'], 'LOC': ['LONDON', 'NEW', 'YORK', 'England',
'PARIS', 'AMSTERDAM', 'BONN', 'WASHINGTON', 'MOSCOW', 'BRUSSELS'], 'ORG': ['Newsroom', 'CHICAGO', 'TEXAS', 'St', 'NEW',
'YORK', 'CINCINNATI', 'OAKLAND', 'Sydney', 'MINNESOTA'], 'MISC': ['DIVISION', 'League', 'EASTERN', 'LEAGUE', 'CENTRAL',
'WESTERN', 'Major', 'English', 'Dutch', 'NATIONAL']}
The top 10 for the second method is:
{'O': ['1996-08-28', '1996-08-27', ',', '.', ':', '4', ')', '1996-08-22', '0', 'points', 'Amount'], 'PER': ['Law', 'Ian', 'Pa
ul', 'David', 'Scott', 'Gilford', 'Ganguly', 'Wessels', 'Adrian', 'Karen'], 'LOC': ['Bank', 'Korea', 'Africa', 'HEMISPHE
RE', 'DELHI', 'KONG', 'England', 'N.J.', 'Lebanon', 'LISBON'], 'ORG': ['Newsroom', 'Christian', 'Ham', 'POST', 'Commodit
ies', 'Newsdesk', 'Vienna', 'FRANCISCO', 'Norwich', 'United'], 'MISC': ['League', 'Democratic', 'DIVISION', 'GMT', 'Kore
an', 'German', 'LEAGUE', 'Rep', 'C$', 'English']}
The top 10 for the third method is:
{'O': ['"', 'seconds', '"', ',', '.', ':', '(', 'second', 'day', '6-3', '6-4'], 'PER': ['Steve', 'Davis', 'Stranksy', 'Shah',
'Ganguly', 'Crosson', 'Vasilopoulos', 'R.', 'Wessels', 'Westwood'], 'LOC': ['Uganda', 'Congo', 'Mauritius', 'Botswana',
'.', 'Tegel', 'Tempelhof', 'ASIA', 'MED', 'Schoenefeld'], 'ORG': ['Akron', 'Weather', 'Hampshire', 'NBH', 'AL', 'Subsid
iaries', 'Tyrrell', 'Preston', 'Texas', 'BUILDING'], 'MISC': ['Yellow', 'Scottish', 'WESTERN', 'German', 'McLaren', 'C$'
, 'DIVISION', 'English', 'Ferrari', 'Super']}
```

2. Evaluation

(1) the top 10 features:

For example，the top 10 for the ‘PER’ by the first method is:

'PER': ['Peter', 'Younis', 'R.', 'Mark', 'Ahmed', 'Corser', 'Fogarty', 'Kocinski', 'Fine', 'Adrian']

These top 10 words can be make sense and are the most positively-weighted features

(2) Discussion of micro-F1 score：

The F1 scores for the three method are 0.780615586117, 0.690589031821, 0.732106339468.

The features I propose improve the results because there is a significant connection between the current word-current label and previous word-previous label feature. The perceptron could work well using the current word-current label and previous word-previous label feature.

(3) other features：

There are other good features:

current word-current label and previous label-current label and current label- next label

current word-current label and previous word-previous label and previous label-current label

previous label-current label and current label- next label