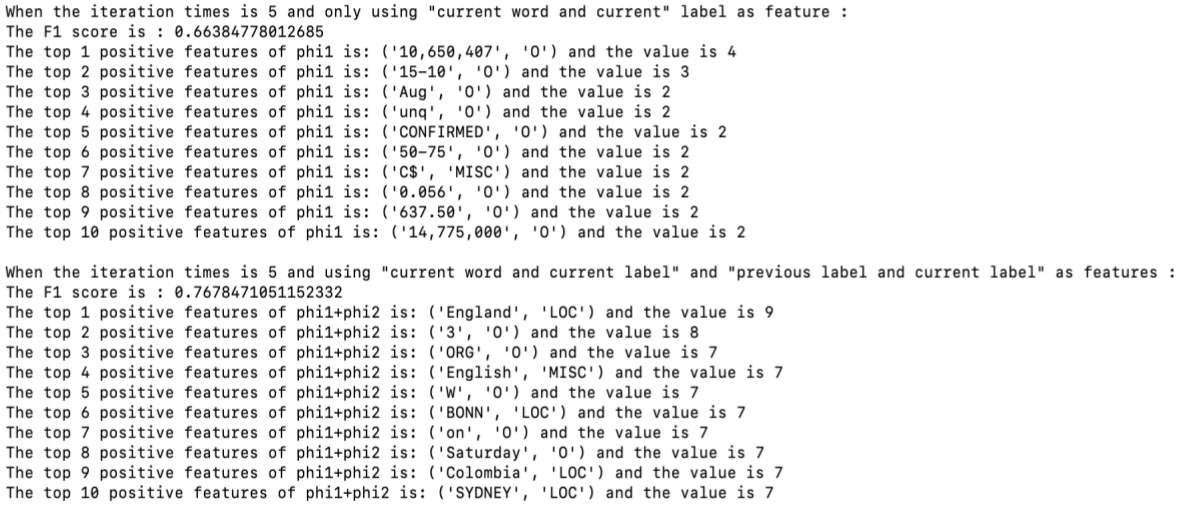
Lab 3 report: Named Entity Recognition with the Structured Perceptron

1. A brief description of the code：

The function 'cw_cl': is to get the counts of current word and current label of all data and 'pl_cl' is to get the counts of previous label and current label of all data.
The function 'phi_1': is to get the given sentence's counts of current word and label and 'phi_2' is to get the given sentence's counts of previous and current labels.
The function 'predict': is to predict the tags for given sentence, the input 'feature' is to determine the feature is phi1 or phi1+phi2. And this function doesn't have parameter 'tags_list', because for every sentence I need to consider all possible labels, so I directly set 5 tags as list in 'predict'. And because I predict two features in one function, so EVERY TIME CALL this 'predict', it must pass both 'cw_cl_counts' and 'pl_cl_counts' as parameters.
The function 'train': is to train model (i.e. Structured Perceptron), the input 'feature' is to determine the feature is phi1 or phi1+phi2.
The function 'average_weight': is to iterate the weight, and the parameter 'iteration' determine how many times it will iterate.
This function 'get_y_predict': is to get the predict tags for test data, the parameter 'all_words_list' is words of all sentence in test data.

2. The result of two features (F1 score & Top 10 positive features) :

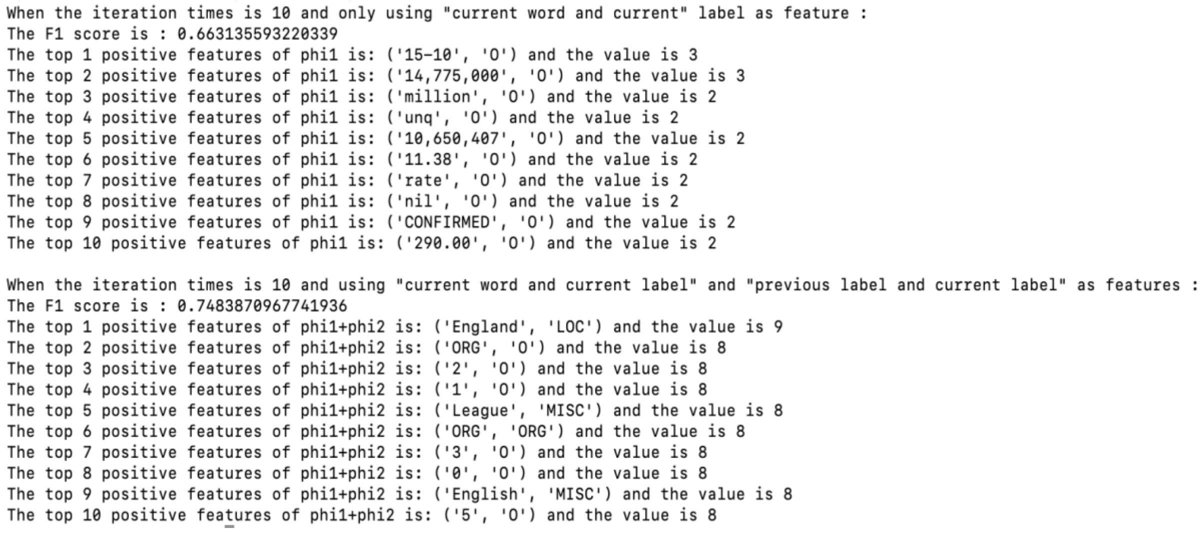
(The result of iteration times is 5)



```
When the iteration times is 5 and only using "current word and current" label as feature :
The F1 score is : 0.66384778012685
The top 1 positive features of phi1 is: ('10,650,407', 'O') and the value is 4
The top 2 positive features of phi1 is: ('15-10', 'O') and the value is 3
The top 3 positive features of phi1 is: ('Aug', 'O') and the value is 2
The top 4 positive features of phi1 is: ('unq', 'O') and the value is 2
The top 5 positive features of phi1 is: ('CONFIRMED', 'O') and the value is 2
The top 6 positive features of phi1 is: ('50-75', 'O') and the value is 2
The top 7 positive features of phi1 is: ('C$', 'MISC') and the value is 2
The top 8 positive features of phi1 is: ('0.056', 'O') and the value is 2
The top 9 positive features of phi1 is: ('637.50', 'O') and the value is 2
The top 10 positive features of phi1 is: ('14,775,000', 'O') and the value is 2

When the iteration times is 5 and using "current word and current label" and "previous label and current label" as features :
The F1 score is : 0.7678471051152332
The top 1 positive features of phi1+phi2 is: ('England', 'LOC') and the value is 9
The top 2 positive features of phi1+phi2 is: ('3', 'O') and the value is 8
The top 3 positive features of phi1+phi2 is: ('ORG', 'O') and the value is 7
The top 4 positive features of phi1+phi2 is: ('English', 'MISC') and the value is 7
The top 5 positive features of phi1+phi2 is: ('W', 'O') and the value is 7
The top 6 positive features of phi1+phi2 is: ('BONN', 'LOC') and the value is 7
The top 7 positive features of phi1+phi2 is: ('on', 'O') and the value is 7
The top 8 positive features of phi1+phi2 is: ('Saturday', 'O') and the value is 7
The top 9 positive features of phi1+phi2 is: ('Colombia', 'LOC') and the value is 7
The top 10 positive features of phi1+phi2 is: ('SYDNEY', 'LOC') and the value is 7
```

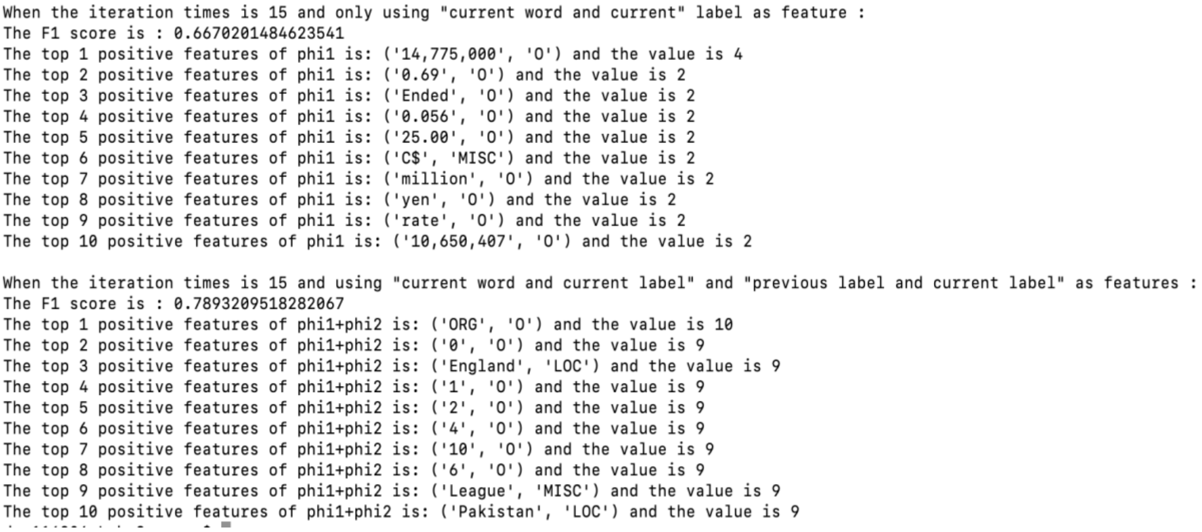
(The result of iteration times is 10)



```
When the iteration times is 10 and only using "current word and current" label as feature :
The F1 score is : 0.663135593220339
The top 1 positive features of phi1 is: ('15-10', 'O') and the value is 3
The top 2 positive features of phi1 is: ('14,775,000', 'O') and the value is 3
The top 3 positive features of phi1 is: ('million', 'O') and the value is 2
The top 4 positive features of phi1 is: ('unq', 'O') and the value is 2
The top 5 positive features of phi1 is: ('10,650,407', 'O') and the value is 2
The top 6 positive features of phi1 is: ('11.38', 'O') and the value is 2
The top 7 positive features of phi1 is: ('rate', 'O') and the value is 2
The top 8 positive features of phi1 is: ('nil', 'O') and the value is 2
The top 9 positive features of phi1 is: ('CONFIRMED', 'O') and the value is 2
The top 10 positive features of phi1 is: ('290.00', 'O') and the value is 2

When the iteration times is 10 and using "current word and current label" and "previous label and current label" as features :
The F1 score is : 0.7483870967741936
The top 1 positive features of phi1+phi2 is: ('England', 'LOC') and the value is 9
The top 2 positive features of phi1+phi2 is: ('ORG', 'O') and the value is 8
The top 3 positive features of phi1+phi2 is: ('2', 'O') and the value is 8
The top 4 positive features of phi1+phi2 is: ('1', 'O') and the value is 8
The top 5 positive features of phi1+phi2 is: ('League', 'MISC') and the value is 8
The top 6 positive features of phi1+phi2 is: ('ORG', 'ORG') and the value is 8
The top 7 positive features of phi1+phi2 is: ('3', 'O') and the value is 8
The top 8 positive features of phi1+phi2 is: ('0', 'O') and the value is 8
The top 9 positive features of phi1+phi2 is: ('English', 'MISC') and the value is 8
The top 10 positive features of phi1+phi2 is: ('5', 'O') and the value is 8
```

(The result of iteration times is 15)



```
When the iteration times is 15 and only using "current word and current" label as feature :
The F1 score is : 0.6670201484623541
The top 1 positive features of phi1 is: ('14,775,000', 'O') and the value is 4
The top 2 positive features of phi1 is: ('0.69', 'O') and the value is 2
The top 3 positive features of phi1 is: ('Ended', 'O') and the value is 2
The top 4 positive features of phi1 is: ('0.056', 'O') and the value is 2
The top 5 positive features of phi1 is: ('25.00', 'O') and the value is 2
The top 6 positive features of phi1 is: ('C$', 'MISC') and the value is 2
The top 7 positive features of phi1 is: ('million', 'O') and the value is 2
The top 8 positive features of phi1 is: ('yen', 'O') and the value is 2
The top 9 positive features of phi1 is: ('rate', 'O') and the value is 2
The top 10 positive features of phi1 is: ('10,650,407', 'O') and the value is 2

When the iteration times is 15 and using "current word and current label" and "previous label and current label" as features :
The F1 score is : 0.7893209518282067
The top 1 positive features of phi1+phi2 is: ('ORG', 'O') and the value is 10
The top 2 positive features of phi1+phi2 is: ('0', 'O') and the value is 9
The top 3 positive features of phi1+phi2 is: ('England', 'LOC') and the value is 9
The top 4 positive features of phi1+phi2 is: ('1', 'O') and the value is 9
The top 5 positive features of phi1+phi2 is: ('2', 'O') and the value is 9
The top 6 positive features of phi1+phi2 is: ('4', 'O') and the value is 9
The top 7 positive features of phi1+phi2 is: ('10', 'O') and the value is 9
The top 8 positive features of phi1+phi2 is: ('6', 'O') and the value is 9
The top 9 positive features of phi1+phi2 is: ('League', 'MISC') and the value is 9
The top 10 positive features of phi1+phi2 is: ('Pakistan', 'LOC') and the value is 9
```

3. The answer of three questions:

①The result of this question I've shown in above pictures.

②The result of this question I've shown in above pictures.
And I observe that for feature 'current word and label', the reason why the top 1 is **('10,650,407', 'O')** is {(10,650,407 10,650,407 10,650,407 10,650,407)(O O O O)} in one sentence and it can't find any other matched tags, so when this word '10,650,407' first time to be predicted, because this is the first time this word appears, so the score must be zero for all possible tags and it has high probability of error prediction, so it will go to update, the weight will equal to real phi minus predict phi, **the value of real phi depends on the frequency of this word in the given sentence, so one word has higher frequency in sentence and will have higher value of weight.**

For feature 'previous label and current label', the top 10 positive features look like make more sense, it proves combine the feature 'previous label and current label' and feature 'current word and label' can improve the result.

③For F1 score, the phi1 + phi2 is higher than phi1, which makes sense, **because two features can better describe the data set, so as to carry out more accurate training and obtain better weight.** But for iteration times, I expected the more iterations, the higher the value of f1, but it doesn't work like this, **it looks like the iteration times can't change too much(for feature phi_1),** even sometimes because the randomly return and input, the value has even gone down, which is not what I expected. And I find when iteration times is 15, the F1 score of feature 'phi1+phi2' is very higher**,** which is what I expected, but I find the top 10 features is changing.