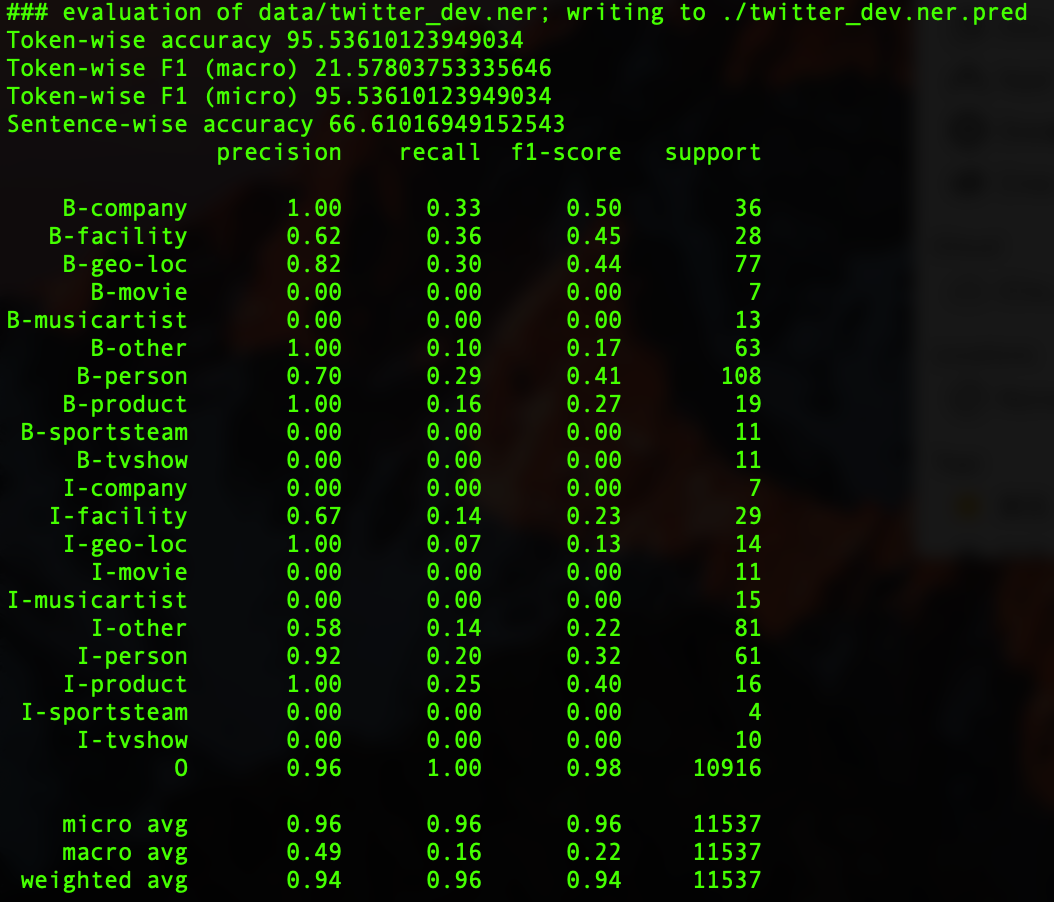
* 1. Feature design explanation

To begin with, I ran the default feature the script given and saved the data as follow:

**Twitter\_dev\_test.ner**



**Twitter\_dev.ner**



**The features I applied to improve F1 score as follow:**

First word

This feature finds out whether the current word is the first one of a sentence or not. In tweet, the first word usually is a named entity.

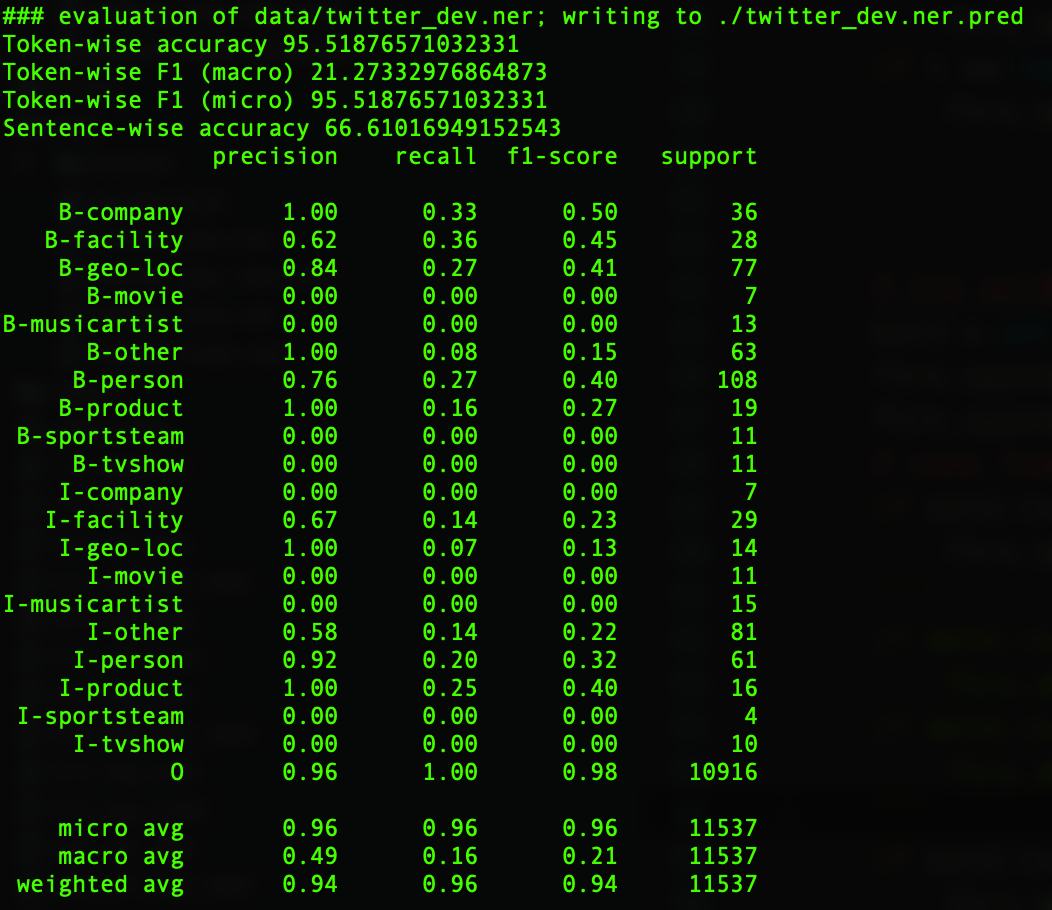
For example, Columbus B-geo-loc， according to this one, the first word ‘Columbus ’is a named entity.

Numeric

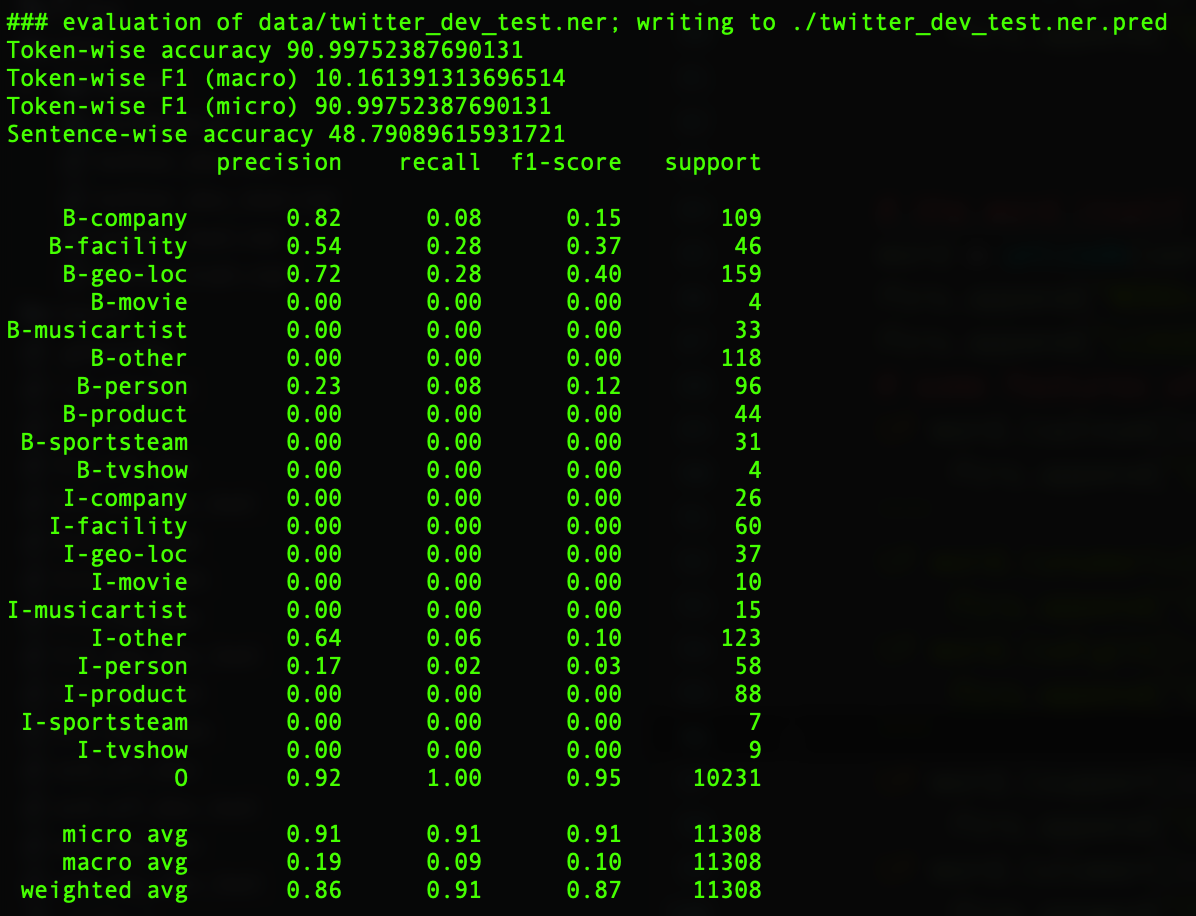
Determine whether current word is a number or not, usually number not a named entity but a count of named entity. For example, “a server 2008 (I-productd) R2”, “the number ”2008”belongs to named entity. Which is product type.

After I delete it from code the result as follow:

**Twitter\_dev\_test.ner**



**Twitter\_dev.ner**



Compared to the default result, without numeric result is worse, so numeric is an useful feature.

Previous word

By ad the previous word as a feature, this can make current word provides more information instead just about itself. For example, “a server 2008 (I-productd) R2” and the current word is ‘2008’, the previous word is ‘server’, which is a named entity.

Next word

By ad the next word as a feature, this can make current word provides more information instead just about itself. For example, “a server 2008 (I-productd) R2” and the current word is ‘2008’, the next word is ‘R2’, which is a named entity.

Upper case

Identify whether current word is in uppercase or not, usually if a word is in uppercase it should be a named entity. Contrast to lowercase feature this one is helpful to tweets. But its not reliable due lots of tweets still not follow this rule.

For example, ‘My Oakland (B-geo-loc) line’, in this tweet there are two fully uppercase words but actually only ‘Oakland’ is named entity instead of ‘My’.

Lower case

Identify whether current word is in lowercase or not, usually if a word is in lowercase its not a named entity.

First word.

This feature finds out whether the current word is the first one of a sentence or not. In tweet, the first word usually is a named entity.

For example, ‘Pedigree (B-company) Donates Dog’, according to this one, the first word ‘Pedigree ’is a named entity.

Last word.

This feature finds out whether the current word is the last one of a sentence or not. In context the last word usually is a named entity. After applied this feature, based on python script evaluation I get follow data:

twitter\_dev.ner:

|  |  |  |
| --- | --- | --- |
|  | basic | advanced |
| accuracy | 95.54 | 95.67 |
| F1(macro) | 21.58 | 22.85 |
| F1(micro) | 95.54 | 95.67 |

twitter\_dev\_test.ner:

|  |  |  |
| --- | --- | --- |
|  | basic | advanced |
| accuracy | 91.02 | 91.52 |
| F1(macro) | 10.92 | 17.18 |
| F1(micro) | 91.02 | 91.52 |

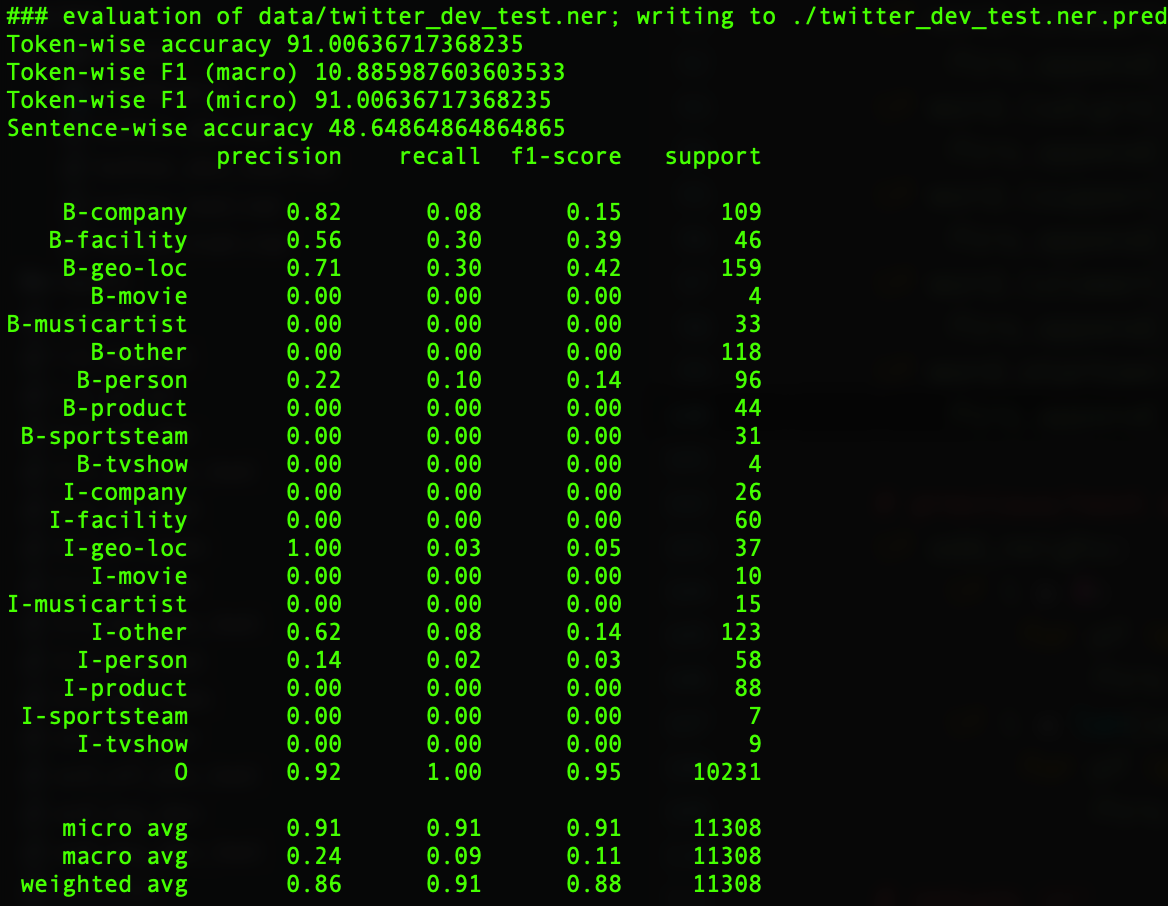
Stop word

This feature is the set of English words that usually appear at the end of an English sentence. These words are not named entities but they are useful to identify named entity.

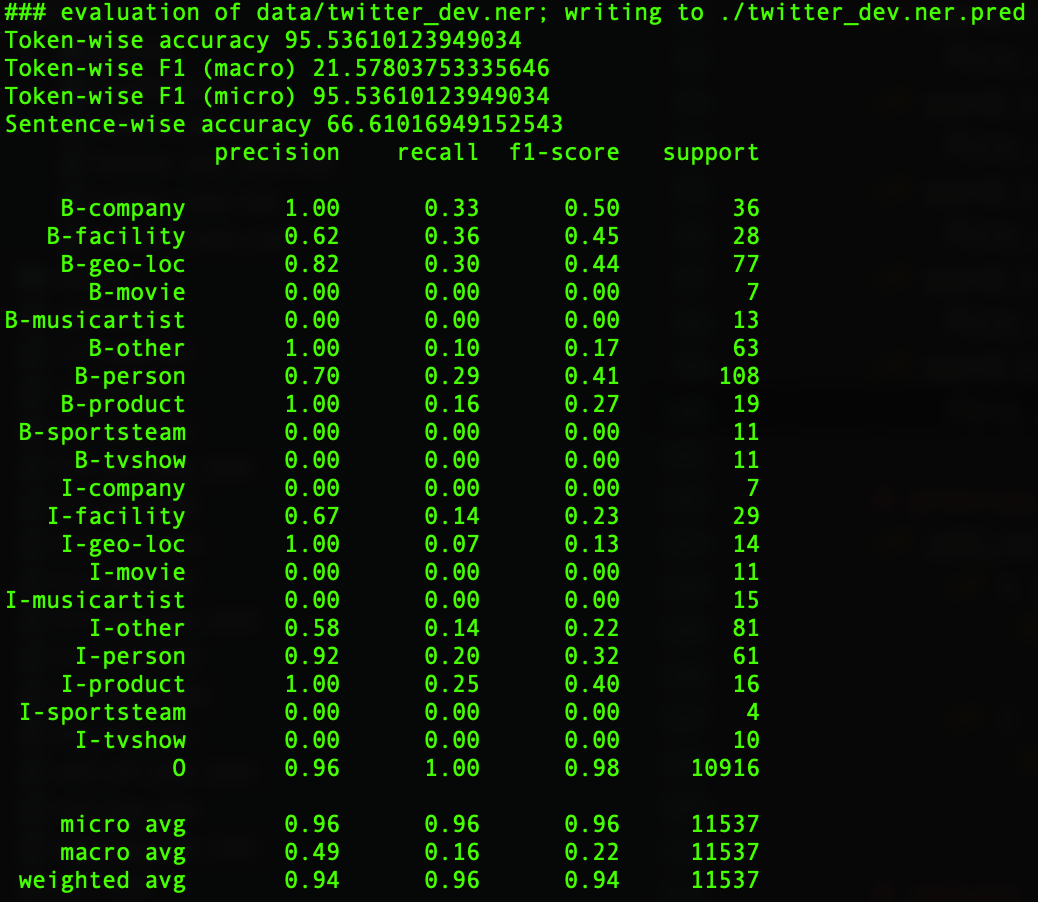
IS\_URL

This feature is used to identify whether the current is a web link or not by the key word ’http , one web link is not a named entity. After adding hyperlink, the result as follow:

**Twitter\_dev\_text.ner**



**Twitter\_dev.ner**



By comparing this result and default, it seems no change at all, but by combine it with lexicon and other useful features, I think it still usefull.

Lexicon word

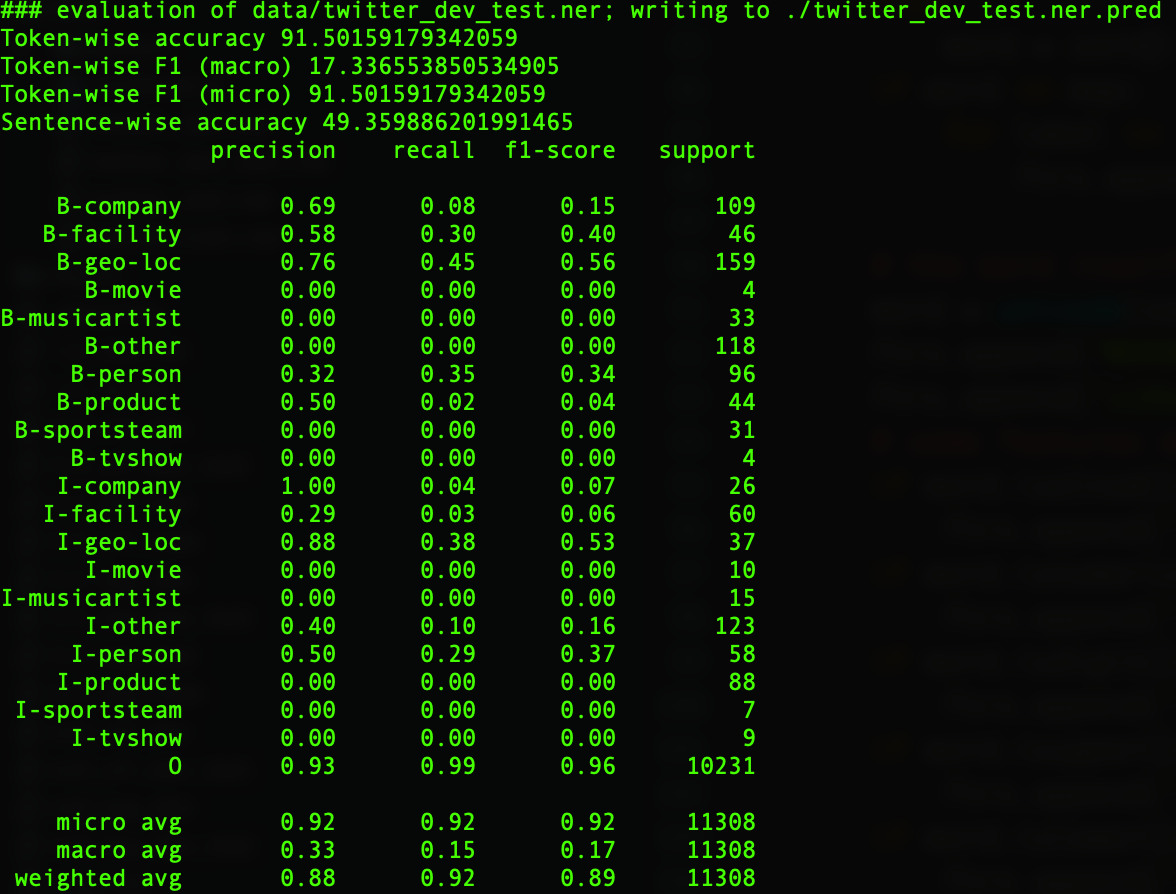
These features includes some sets I added to the code as follow:

Location; Business Product; Product; Tv show; Sport team; Location; Name;

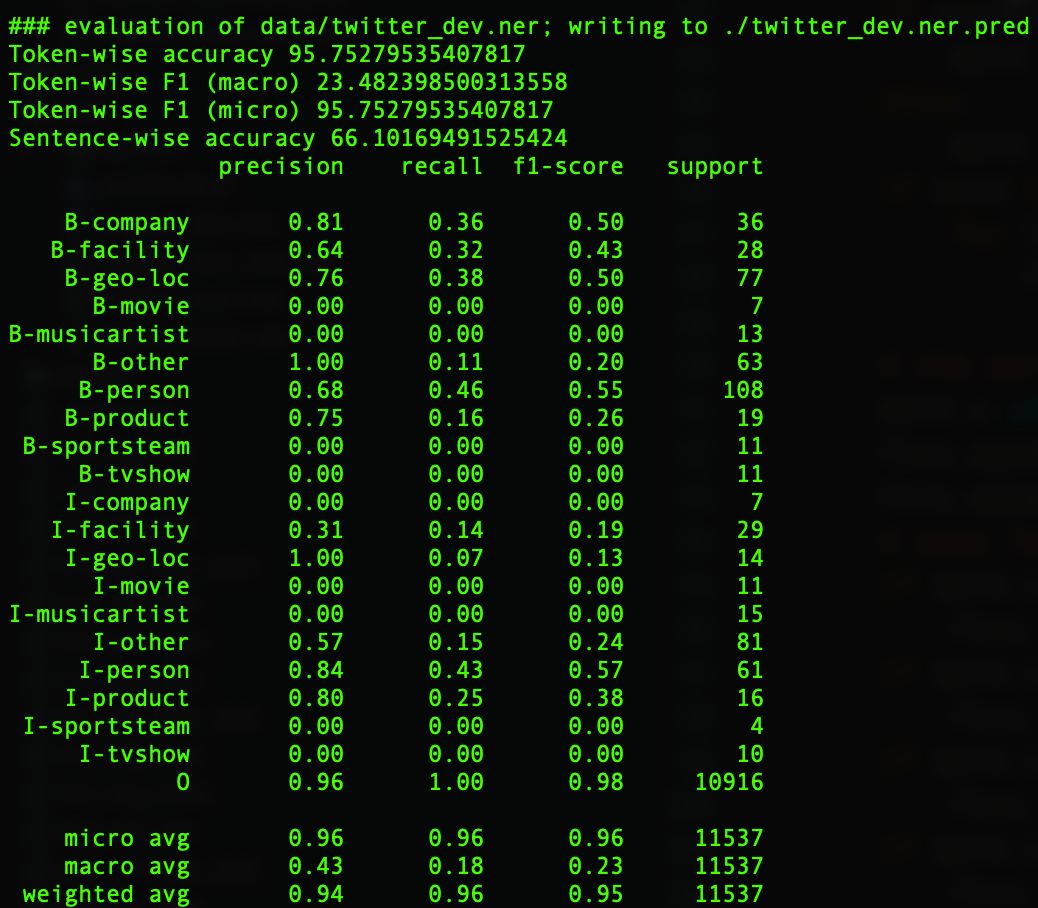
The label ‘name’ Including ‘last name' and ‘first name’ two features, these two features contribute to name, which is named entity.

For example, “LG Venus” in the set ‘product’ and ‘business\_consumer\_product’. so both label ‘Product’ and ‘Business Product’ is attached and others so on.

**twitter\_dev\_test.ner:**



**twitter\_dev.ner:**

****

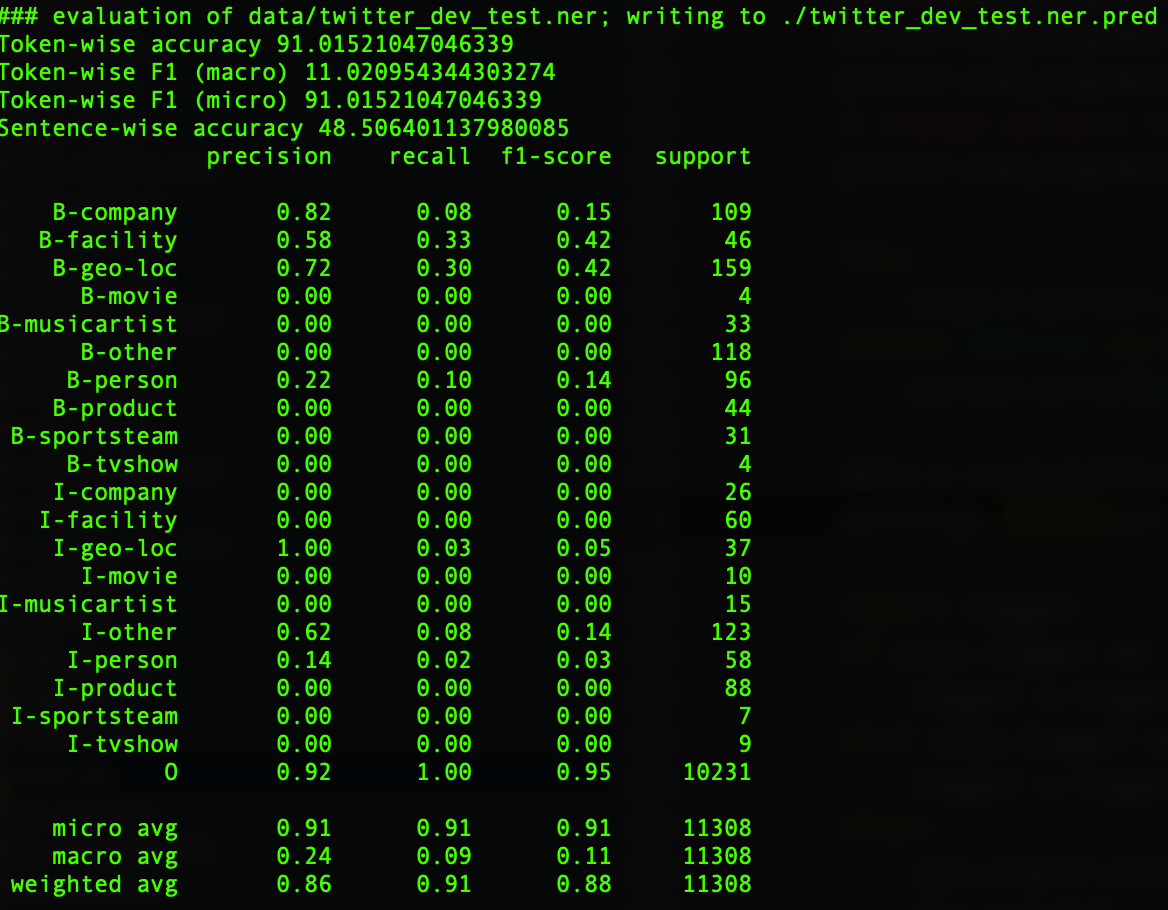
Its obvious that after applied lexicon features the result (especially F1) improved a lot.

**The features I applied not improve F1 score as follow:**

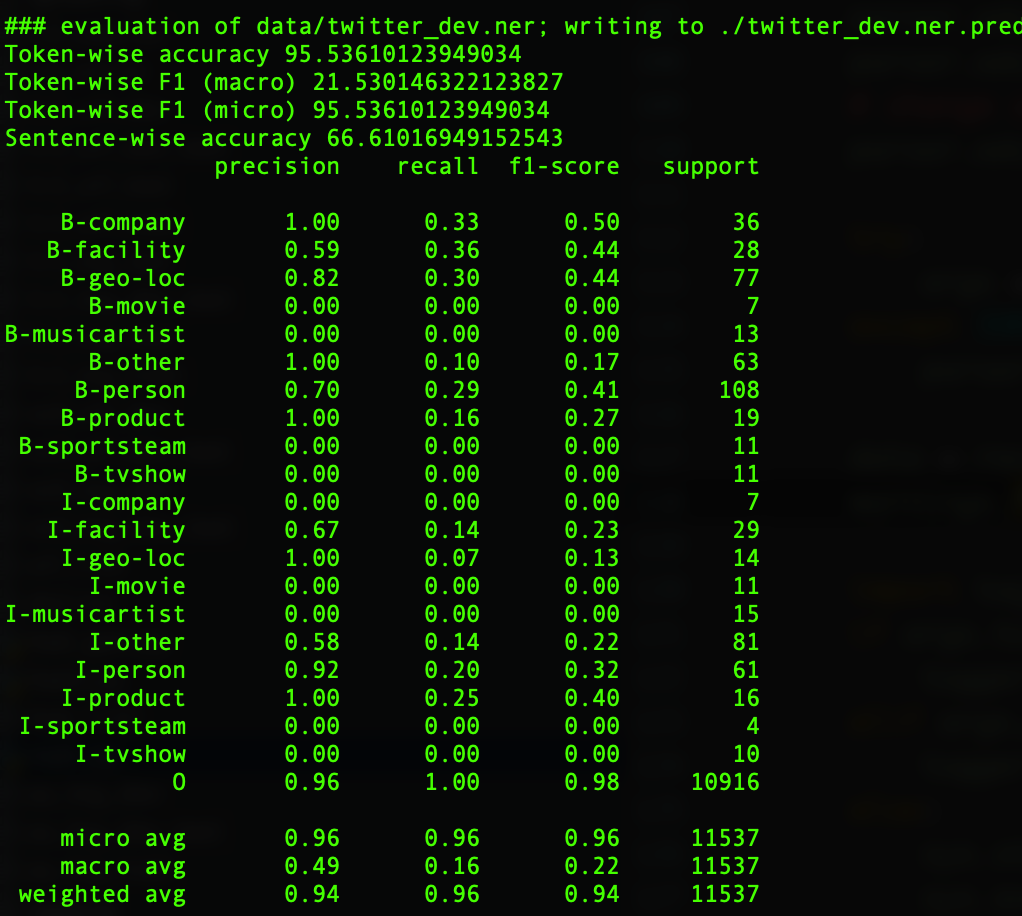
First word is number

Counting the number appear at the beginning of sentence, which usually are the pre part of named entity. But this feature is not work well. According to the follow data, the result gets worse.

**twitter\_dev\_test.ner:**

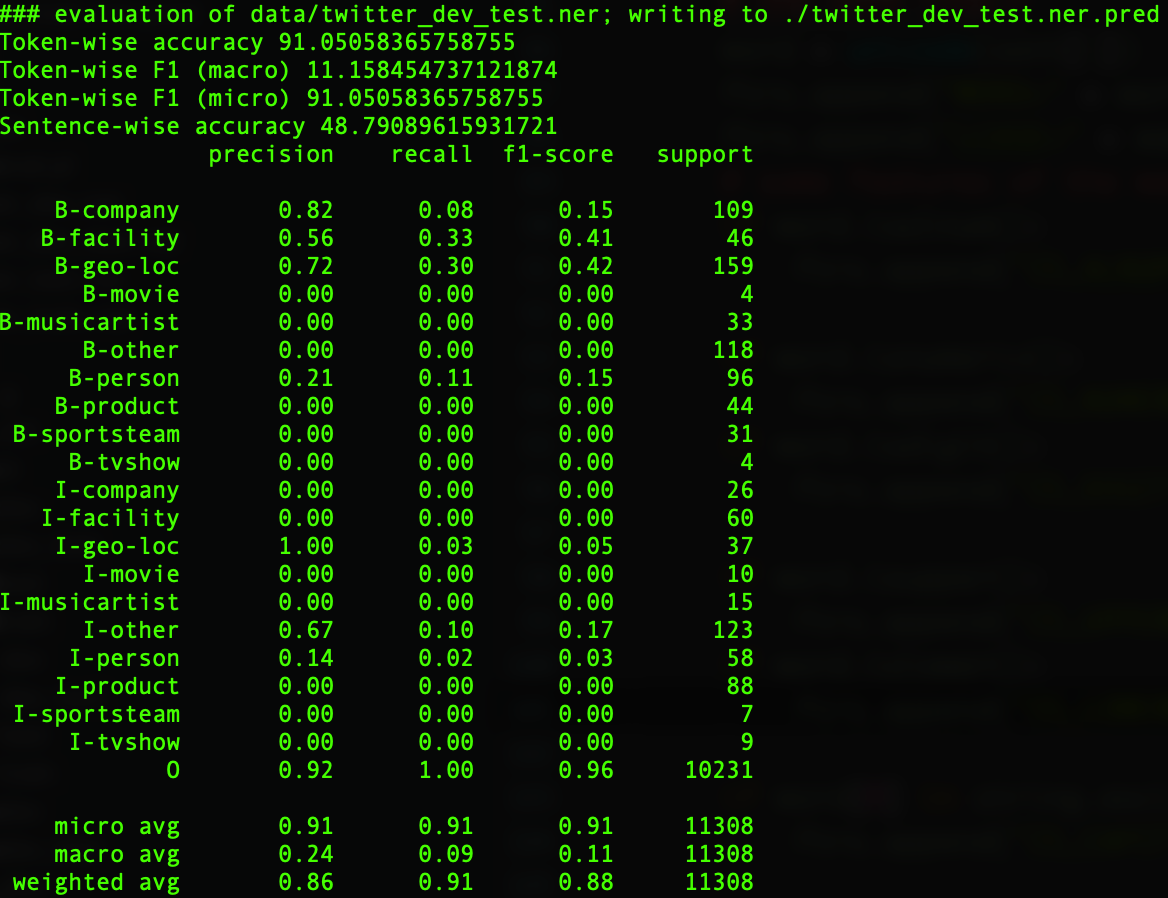
****

**twitter\_dev.ner:**

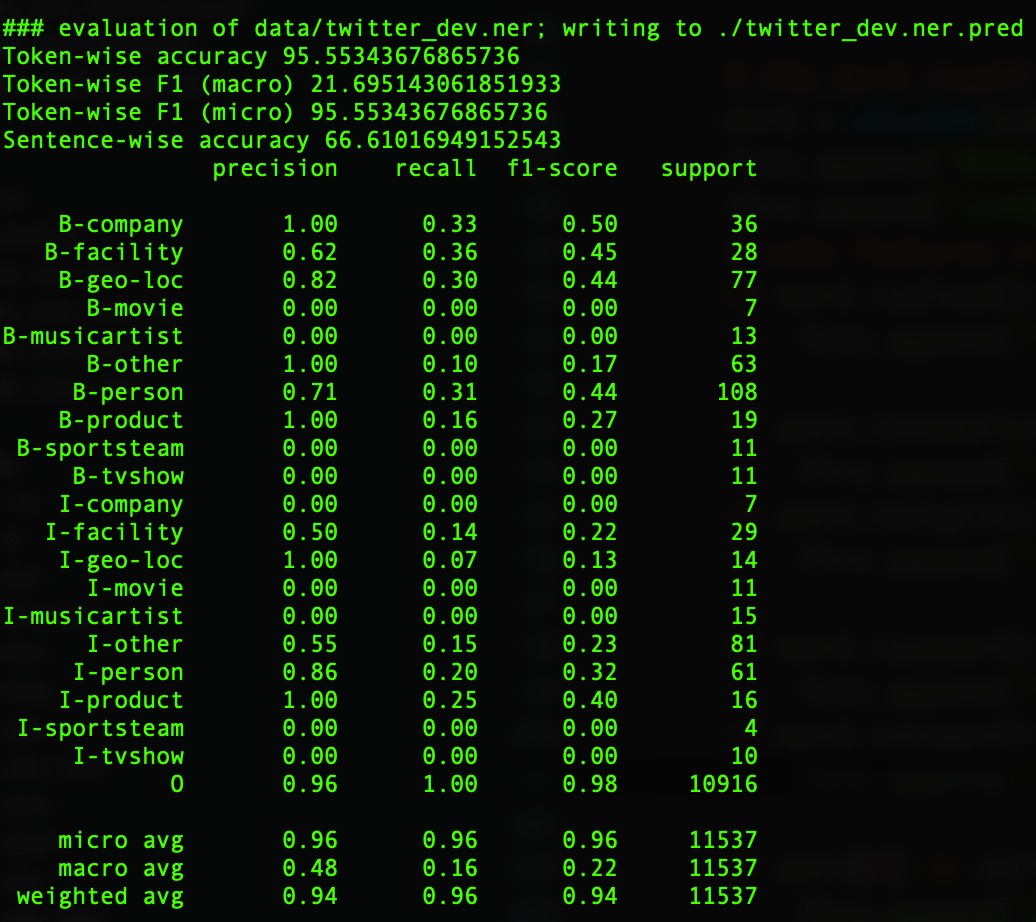
****

First character is uppercase

**twitter\_dev\_test.ner:**

****

**twitter\_dev.ner:**



this result makes no difference with the default’s , so it is not useful.