1 导入相关的库

In [4]:

```
import pandas as pd
from sklearn.metrics import f1_score
import fasttext
```

2 读取数据

In [2]:

```
train_df = pd.read_csv('./data/train_set.csv', sep='\t')
train_df
```

Out[2]:

	label	text
0	2	2967 6758 339 2021 1854 3731 4109 3792 4149 15...
1	11	4464 486 6352 5619 2465 4802 1452 3137 5778 54...
2	3	7346 4068 5074 3747 5681 6093 1777 2226 7354 6...
3	2	7159 948 4866 2109 5520 2490 211 3956 5520 549...
4	3	3646 3055 3055 2490 4659 6065 3370 5814 2465 5...
...	...	...
199995	2	307 4894 7539 4853 5330 648 6038 4409 3764 603...
199996	2	3792 2983 355 1070 4464 5050 6298 3782 3130 68...
199997	11	6811 1580 7539 1252 1899 5139 1386 3870 4124 1...
199998	2	6405 3203 6644 983 794 1913 1678 5736 1397 191...
199999	3	4350 3878 3268 1699 6909 5505 2376 2465 6088 2...

200000 rows × 2 columns

In [3]:

```
train_df['label_ft']='__label__'+train_df['label'].astype(str)
```

In [4]:

```
train_df[['text','label_ft']].iloc[:-5000].to_csv('trian_fast_195000.csv', index=None, header=
```

In [5]:

```
fs1= fasttext.train_supervised(input='trian_fast_195000.csv', dim=200, epoch=25,
                                    lr=0.1, wordNgrams=2, minCount=1,loss='softmax')
```

In [6]:

```
result=fs1.test('trian_fast_195000.csv')
```

In [7]:

```
result
```

Out[7]:

(195000, 0.9786461538461538, 0.9786461538461538)

In [8]:

```
val_pred = [fs1.predict(x)[0][0].split('__')[-1] for x in train_df.iloc[-5000:]['text']]
print(f1_score(train_df['label'].values[-5000:].astype(str), val_pred, average='macro'))
```

0.9274703163040121

fasttext参数选下列时 input='trian_fast_195000.csv', dim=200, epoch=25,lr=0.1, wordNgrams=2, minCount=1,loss='softmax'
分数达到0.927

In [9]:

```
test_df = pd.read_csv('./data/test_a.csv', sep='\t')
```

In [10]:

```
test_pred_ft = [fs1.predict(x)[0][0].split('__')[-1] for x in test_df['text']]
```

In [11]:

```
test_pred_ft=pd.DataFrame(test_pred_ft)
test_pred_ft.columns=['label']
```

In [12]:

```
test_pred_ft.to_csv('./output/test_a_pred_ft.csv', index=None, encoding='utf8')
```

In [6]:

```
fs2= fasttext.train_supervised(input='trian_fast_195000.csv', dim=150, epoch=25,
                                        lr=0.1, wordNgrams=2, minCount=1,loss='hs')
```

In [7]:

```
val_pred = [fs2.predict(x)[0][0].split('__')[-1] for x in train_df.iloc[-5000:]['text']]
print(f1_score(train_df['label'].values[-5000:].astype(str), val_pred, average='macro'))
```

0.9112899510591518

In [ ]:

```
将dim改为150后，分数降低，说明dim对结果影响挺大
```