به نام خدا

محمدمهدی آقاجانی

۹۳۳۱۰۵۶

تمرین دوم

استاد : دکتر ممتازی

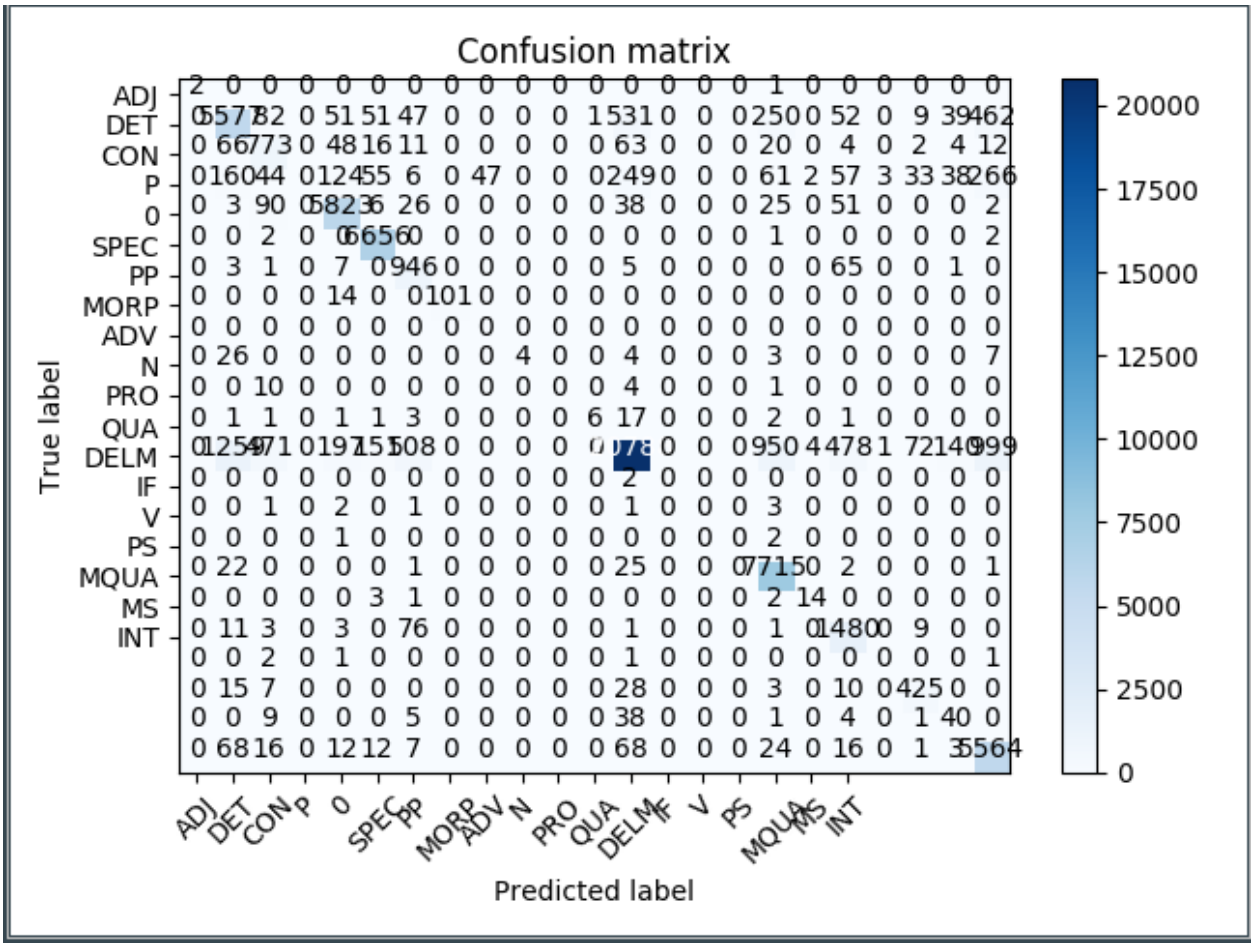
تمرین اول

در این تمرین از ابزار NLTK استفاده شد . برای عمل آموزش از متد HiddenMarkovModelTagger استفاده کردیم که آن را با داده هایی که در اختبار داشتیم train کردیم و سپس با استفاده از متد tag اقدام به تگ گذاری داده ها نمودیم.

نکته بسیار مهم این است که در این نوع یادگیری باید حتما داده ها را بر اساس جمله و نه کلمه له کلمه در احتبار ابزار بگذاریم تا دقت بهتری حاصل شود.

با استفاده از روش بالا دقت ۸۵ درصد به دست آمد که دقت مناسبی میتواند باشد البته می توان با تنظیم پارامتر ها به دقت های بهتری دست پیدا کرد. ماتریس درهم[۱] این این سوال نیز در گزارش آمده است

[۱] Confusion matrix


Confusion matrix
True label
Predicted label
ADJ
DET
CON
P
0
SPEC
PP
MORP
ADV
N
PRO
QUA
DELM
IF
V
PS
MQUA
MS
INT
20000
17500
15000
12500
10000
7500
5000
2500
0

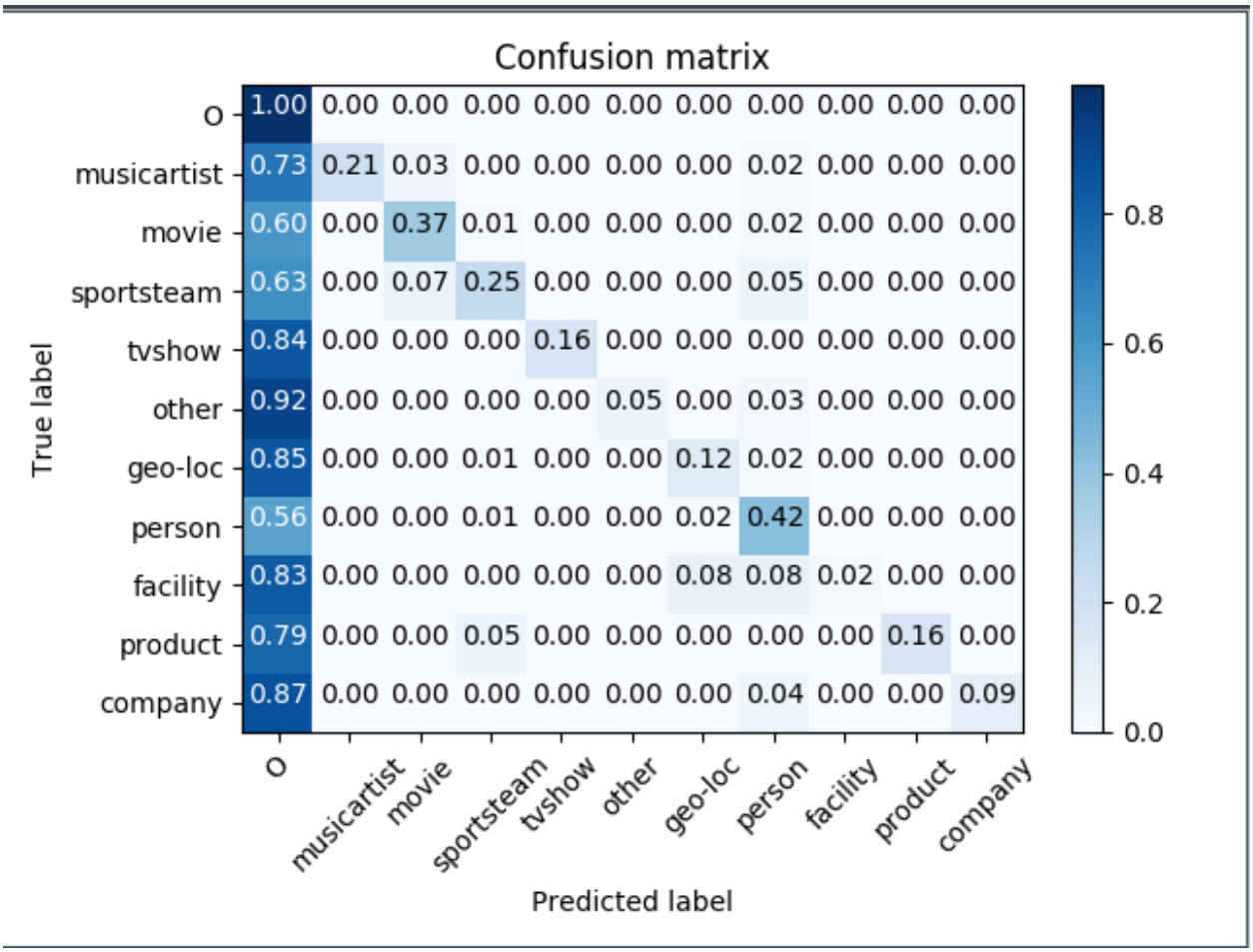
تمرین دوم

در این تمرین از ابزار Stanford استفاده نمودیم. ابتدا لازم بود که مدل مربوط به خودمان را با داده های آماده شده بسازیم که برای این کار باید یک فایل prop. آماده میشد که مسیر داده و پارامتر های مورد نظر در آن ست شده بود (این فایل درون پوشه بارگذاری شده موجود می باشد). سپس با استفاده از دستور زیر مدل خود را ساختیم :

```
java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -prop filename.prop
```

با این دستور یک فایل ser.gz. تولید میشود که باید در کد از آن استفاده کنیم.

سپس در کد با استفاه از متد tag آن ها را دسته بندی کردیم. در این حالت دقت کار ۹۵ درصد شد که قابل قبول است اما نکته مهم توجه به confusion matrix است که در گزارش هم آمده است و در این نمودار میبینیم که خیلی از کلاس ها به کلاس PERSON نگاشت شده اند. البته تا حدودی نشان از این بحث چالش برانگیز می باشد که خیلی از این دسته بندی ها با دسته بندی person اشتراک زیادی دارند زیرا اسامی اشخاص میتواند عمومیت بیشتری نسبت به ما بقی کلاس ها داشته باشد.


Confusion matrix
True label
Predicted label
O
musicartist
movie
sportsteam
tvshow
other
geo-loc
person
facility
product
company
1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.73 0.21 0.03 0.00 0.00 0.00 0.00 0.02 0.00 0.00 0.00
0.60 0.00 0.37 0.01 0.00 0.00 0.00 0.02 0.00 0.00 0.00
0.63 0.00 0.07 0.25 0.00 0.00 0.00 0.05 0.00 0.00 0.00
0.84 0.00 0.00 0.00 0.16 0.00 0.00 0.00 0.00 0.00 0.00
0.92 0.00 0.00 0.00 0.00 0.05 0.00 0.03 0.00 0.00 0.00
0.85 0.00 0.00 0.01 0.00 0.00 0.12 0.02 0.00 0.00 0.00
0.56 0.00 0.00 0.01 0.00 0.00 0.02 0.42 0.00 0.00 0.00
0.83 0.00 0.00 0.00 0.00 0.00 0.08 0.08 0.02 0.00 0.00
0.79 0.00 0.00 0.05 0.00 0.00 0.00 0.00 0.00 0.16 0.00
0.87 0.00 0.00 0.00 0.00 0.00 0.00 0.04 0.00 0.00 0.09
0.8
0.6
0.4
0.2
0.0