Amirkabir University of Technology
(Tehran Polytechnic)

Spring 2023

NLp-HW2

Zahra Zanjani

Zahra.zanjani2@gmail.com

Zahra.zanjani99@aut.ac.ir

Student id: 401131025

بخش اول

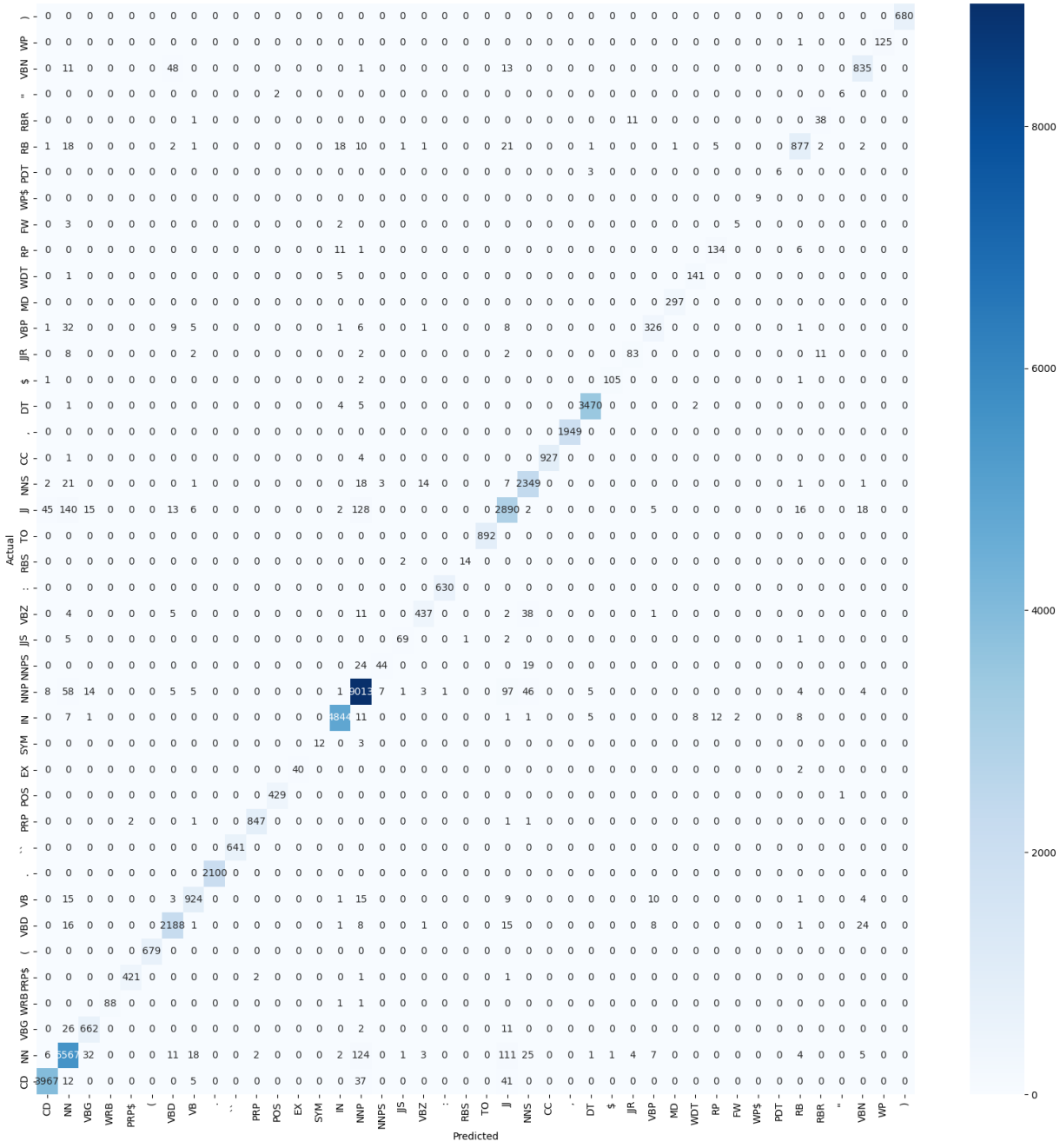
قسمت الف

در این بخش با استفاده از کتابخانه‌ی sklearn_crfsuite الگوریتم crf را پیاده سازی شده است.برای بهبود عملکرد crf تابع word2features ویژگی های هر کلمه را به دست امده است این تابع اساس مستندات کتابخانه ی sklearn_crfsuite نوشته شده‌است.نتایج مدل بعد از ۲۰۰ ایتریشن[1] به شرح زیر می باشد

	score
accuracy_score	0.96
precision_score	0.91
recall_score	0.95

[1] iteration

خروجی ماتریس درهمریختگی به شرح زیر می باشد.



بیشترین خطا برای خروجی JJ (adjective) به جای NN (noun)و IN(preposition)می باشد. به طور کلی مدل روی صفت بایاس هست و تعداد زیادی از تگ ها را به طور اشتباه صفت پیش بینی میکند. جون صفت ها بعد از اسم پرتکرار ترین تگ هستند[2] و تشخیص prepositions پیچیده است.

برای ایجاد بردار ویژگی کلمات با استفاده از pos کلمات مهم هر جمله(اسم ، فعل و صفت) را پیدا می‌کنیم و بردار ویژگی این کلمات را میانگین میگیریم.

[2] according to https://academicguides.waldenu.edu/writingcenter/grammar/prepositions

قسمت ب

در این بخش با استفاده از کتابخانه‌ی sklearn_crfsuite الگوریتم crf را پیاده سازی شده است.برای بهبود عملکرد crf تابع word2features ویژگی های هر کلمه را به دست امده است این تابع اساس مستندات کتابخانه ی sklearn_crfsuite نوشته شده‌است.

برای این استفاده از این مدل نیاز به pos tag کلمات داریم که با استفاده از مدل قسمت قبل برای متون این بخش POS tags را به دست می اید.

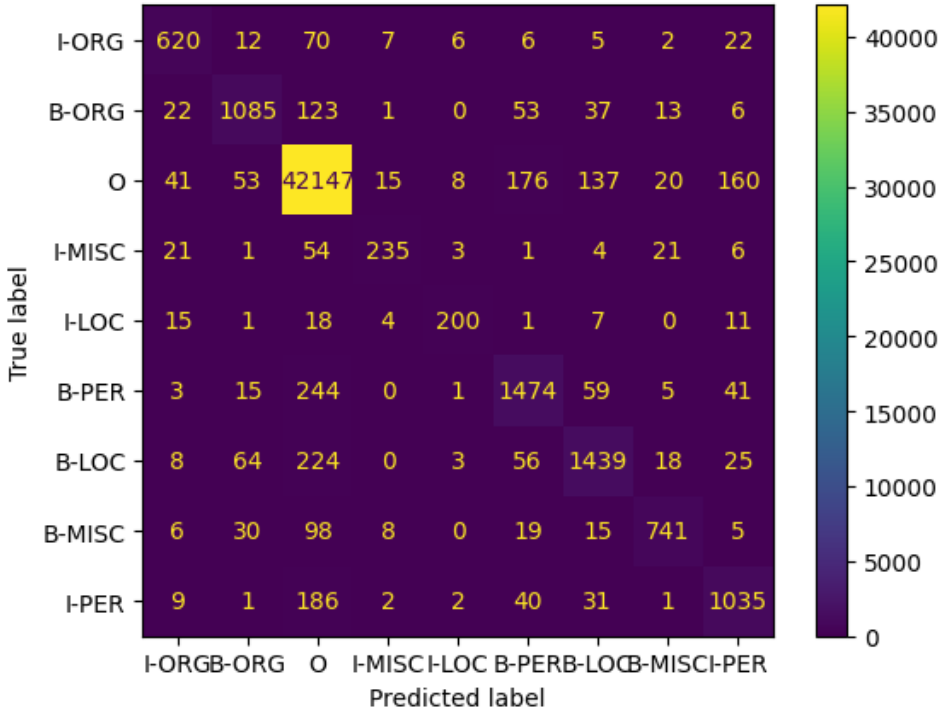
نتایج مدل بعد از ۲۰۰ ایتریشن[3] به شرح زیر می باشد.

	score
accuracy_score	0.95
precision_score	0.80
recall_score	0.80

Label	Precision	Recall	F1	Support
B-LOC	0.702	0.684	0.693	1837
B-MISC	0.815	0.723	0.767	922
B-ORG	0.709	0.533	0.608	1340
B-PER	0.743	0.708	0.725	1842
I-LOC	0.673	0.560	0.612	257
I-MISC	0.733	0.523	0.611	346
I-ORG	0.668	0.624	0.645	750
I-PER	0.732	0.722	0.727	1307
O	0.965	0.983	0.974	42757

[3] iteration

خروجی ماتریس درهمریختگی به شرح زیر می باشد.

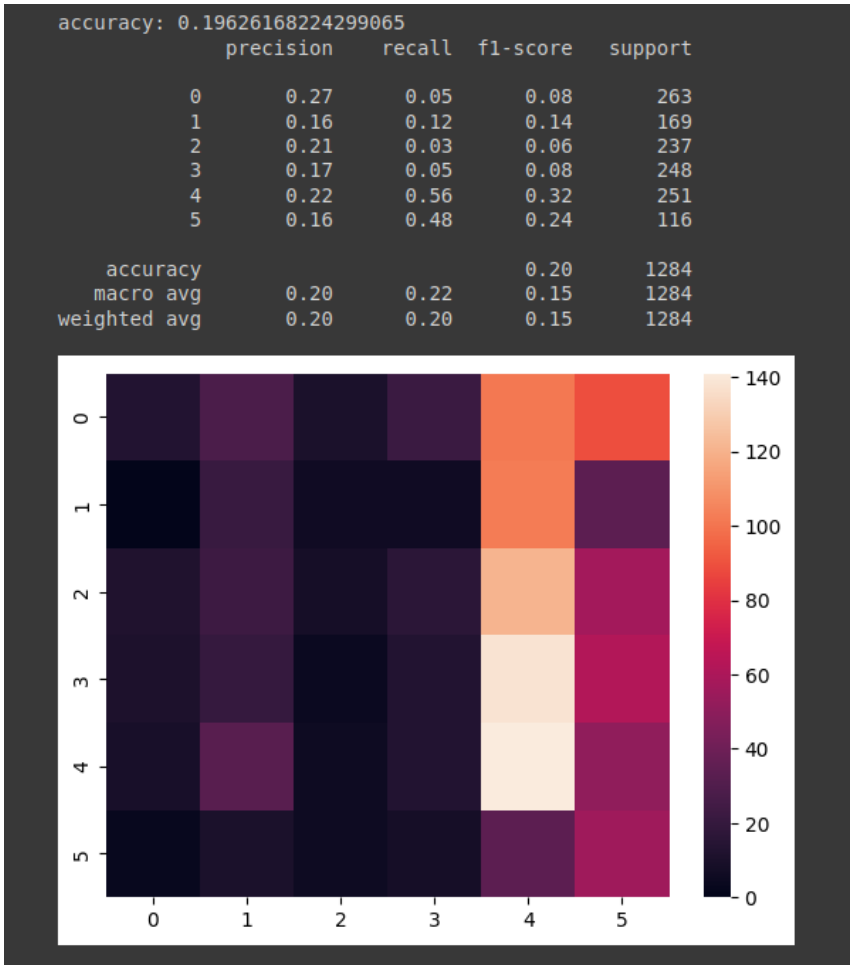


بیشترین تعداد خطا مربوط به تشخیص o هست و مدل به کلماتی که Named Entity نیستند برچسب فرد یا مکان نسبت داده است. چون برچسب بیشتر کلمات o می باشد. پایین ترین دقت در مربوط تشخیص ادامه ی اسم مکان , miscellaneous و organisation می باشد زیرا تشخیص مرز یک Named Entity برای مدل مشکل است. مدل روی برچسب O بایاس هست زیرا بیشتر حجم دادگان اموزش برچسب O دارند.

برای ایجاد بردار ویژگی کلمات با استفاده از NERکلمات مهم هر جمله(اسم فرد ،اسم مکان و اسم) را پیدا می‌کنیم و بردار ویژگی این کلمات را میانگین میگیریم.

قسمت ج

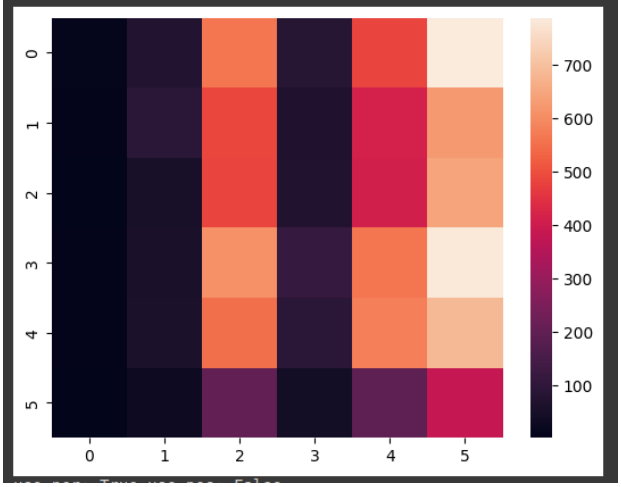
در ابتدا با حذف کلمات stopwords , بردن کلمات به حالت مصدری (lemmenize) با استفاده از word2vec بردار هر جمله را به دست می اوریم و سپس با استفاده از مدل نایو بیز دسته بندی میکنیم (برای بهبود نتایج میتوان از مدل های از پیش اموزش دیده شبکه ی عصبی استفاده کرد)



نتایج استفاده از POS , NER به شرح زیر می باشد.

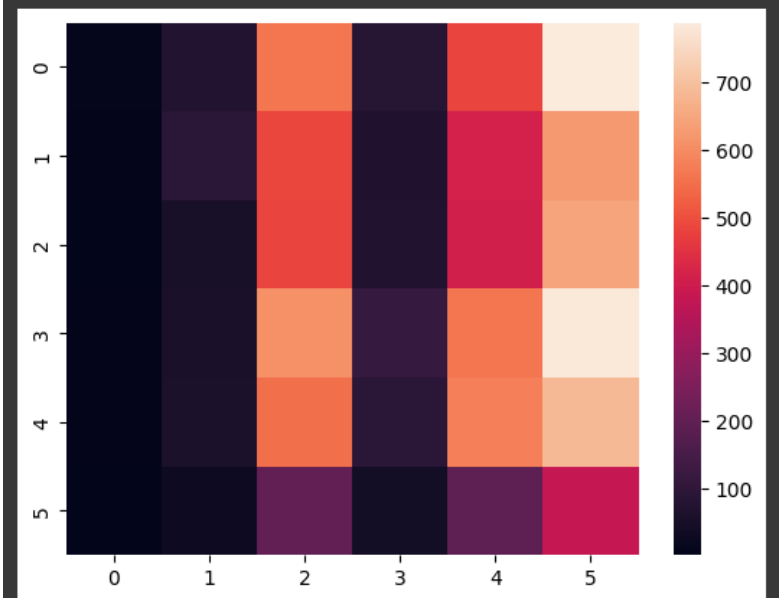
استفاده ی همزمان از pos و ner

```
use_ner: True use_pos: True
accuracy: 0.1609699094361671
              precision    recall  f1-score   support

           0       0.29      0.00      0.01      1998
           1       0.25      0.05      0.09      1683
           2       0.17      0.29      0.21      1657
           3       0.25      0.05      0.09      2123
           4       0.22      0.29      0.25      1966
           5       0.10      0.46      0.16       842

    accuracy                           0.16     10269
   macro avg       0.21      0.19      0.13     10269
weighted avg       0.23      0.16      0.13     10269
```

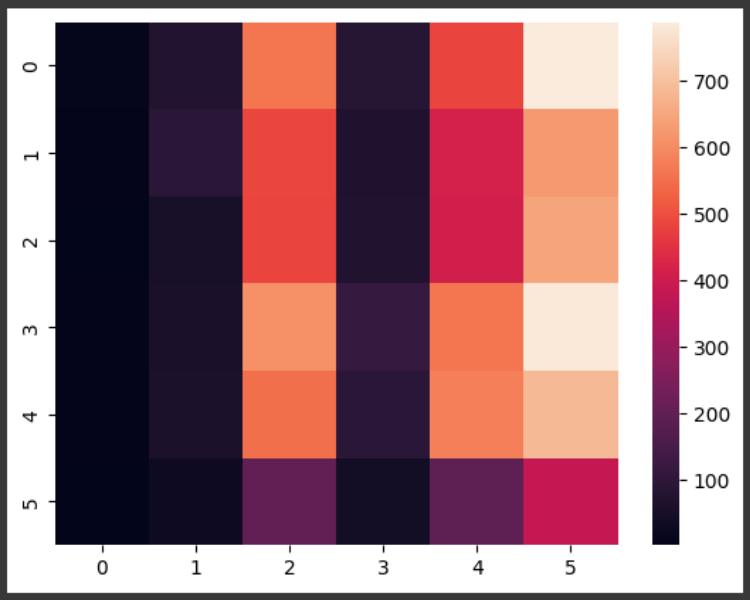


استفاده ی ner

```
use_ner: True use_pos: False
accuracy: 0.1609699094361671
              precision    recall  f1-score   support

           0       0.29      0.00      0.01      1998
           1       0.25      0.05      0.09      1683
           2       0.17      0.29      0.21      1657
           3       0.25      0.05      0.09      2123
           4       0.22      0.29      0.25      1966
           5       0.10      0.46      0.16       842

    accuracy                           0.16     10269
   macro avg       0.21      0.19      0.13     10269
weighted avg       0.23      0.16      0.13     10269
```



```
use_ner: False use_pos: True
accuracy: 0.1609699094361671
              precision    recall  f1-score   support

           0       0.29      0.00      0.01      1998
           1       0.25      0.05      0.09      1683
           2       0.17      0.29      0.21      1657
           3       0.25      0.05      0.09      2123
           4       0.22      0.29      0.25      1966
           5       0.10      0.46      0.16       842

    accuracy                           0.16     10269
   macro avg       0.21      0.19      0.13     10269
weighted avg       0.23      0.16      0.13     10269
```



نتایج استفاده از pos