گزارش تمرین سوم درس NLP

هدف از انجام این تمرین مقایسه ی دو روش naive bayes و maximum entropy با ابزار mallet می باشد. بدین منظور مراحل زیر را انجام می دهیم:

1. ساخت فایل ورودی متناسب با ابزار mallet

یک فایل می سازیم که هر خط از این فایل یک نمونه از داده ی train است که در واقع یک جمله است. فرمت هر خط به شکل زیر است:

INSTANCE_NAME LABEL FEATURE1 FEATURE2 ...

Instance_name را برابر با یک عدد که نشان دهنده ی چندمین نمونه است قرار دادیم. مقدار label هم یکی از دو کلاس politics یا sports خواهد بود. Feature هایی که انتخاب کردیم سه مورد است:

الف) 22 کلمه که هر کدام در یکی از کلاس ها تکرار زیادی داشته است و در کلاس دیگر یا اصلا نبوده یا تعداد کمی داشته است.

ب) part of speech هر کدام از کلمات جمله را هم به عنوان feature داده ایم که با استفاده از کتابخانه ی Nhazm این تگ ها را بدست آوردیم.

ج) bigram های موجود در هر جمله را هم feature در نظر گرفتیم.

2. ساخت فایل mallet. از روی txt.

دستور زیر را در command prompt وارد می کنیم:

```
C:\mallet>bin\mallet import-file --input dataTrain.txt --output dataTrain.mallet
C:\mallet>
```

3. مقایسه ی دو روش naive bayes و maximum entropy

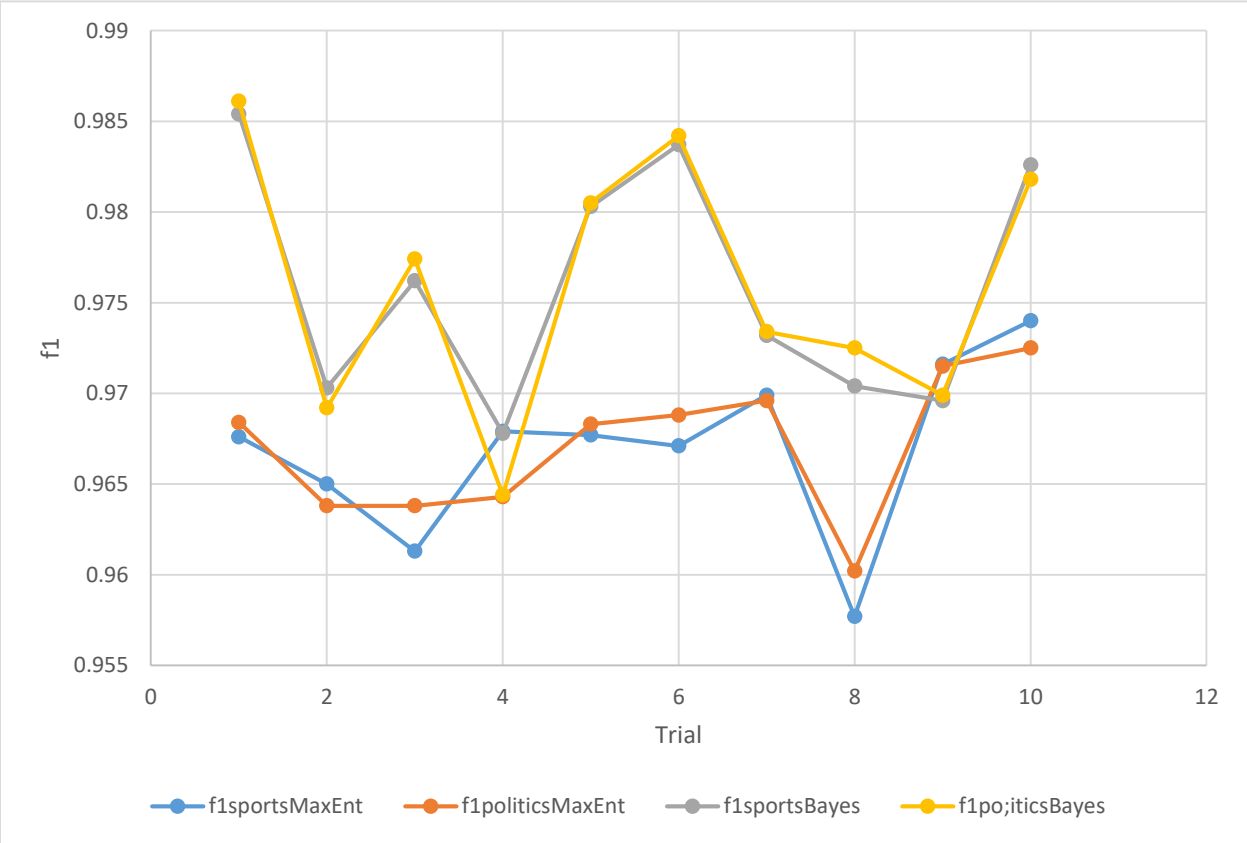
پس از وارد کردن دستور زیر در command prompt نتیجه برای یک trailبه صورت زیر بدست می آید:

```
C:\mallet>bin\mallet train-classifier --input dataTrain.mallet --training-portion 0.9 --trainer MaxEnt --trainer NaiveBa
yes --cross-validation 10.
```

```
------------------- Trial 0  -------------------

Trial 0 Training MaxEntTrainer,gaussianPriorVariance=1.0 with 5065 instances
Value (labelProb=105.26321360947485 prior=127.80390904386117) loglikelihood = -233.067122653336
Exiting L-BFGS on termination #1:
value difference below tolerance (oldValue: -233.08707625163936 newValue: -233.067122653336
Value (labelProb=105.31391199632974 prior=127.72710365933979) loglikelihood = -233.04101565566953
Exiting L-BFGS on termination #1:
value difference below tolerance (oldValue: -233.05695796202895 newValue: -233.04101565566953

Trial 0 Training MaxEntTrainer,gaussianPriorVariance=1.0 finished
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 training data accuracy = 0.998025666337611
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 Test Data Confusion Matrix
Confusion Matrix, row=true, column=predicted  accuracy=0.9680284191829485 most-frequent-tag baseline=0.5079928952042628
       label   0   1  |total
  0 politics 276  10  |286
  1   sports   8 269  |277

Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data precision(politics) = 0.971830985915493
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data precision(sports) = 0.96415770609319
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data recall(politics) = 0.965034965034965
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data recall(sports) = 0.9711191335740073
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data F1(politics) = 0.968421052631579
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data F1(sports) = 0.9676258992805756
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data accuracy = 0.9680284191829485
Trial 0 Training NaiveBayesTrainer with 5065 instances
Trial 0 Training NaiveBayesTrainer finished
Trial 0 Trainer NaiveBayesTrainer training data accuracy = 0.9889437314906219
Trial 0 Trainer NaiveBayesTrainer Test Data Confusion Matrix
Confusion Matrix, row=true, column=predicted  accuracy=0.9857904085257548 most-frequent-tag baseline=0.5079928952042628
       label   0   1  |total
  0 politics 285   1  |286
  1   sports   7 270  |277

Trial 0 Trainer NaiveBayesTrainer test data precision(politics) = 0.976027397260274
Trial 0 Trainer NaiveBayesTrainer test data precision(sports) = 0.996309963099631
Trial 0 Trainer NaiveBayesTrainer test data recall(politics) = 0.996503496503496
Trial 0 Trainer NaiveBayesTrainer test data recall(sports) = 0.9747292418772563
Trial 0 Trainer NaiveBayesTrainer test data F1(politics) = 0.986159169550173
Trial 0 Trainer NaiveBayesTrainer test data F1(sports) = 0.9854014598540146
Trial 0 Trainer NaiveBayesTrainer test data accuracy = 0.9857904085257548
```

	Trail 0	Trail 1	Trail 2	Trail 3	Trail 4	Trail 5	Trail 6	Trail 7	Trail 8	Trail 9
F1 sports MaxEnt	0.9676	0.9650	0.9613	0.9679	0.9677	0.9671	0.9699	0.9577	0.9716	0.9740
F1 politics MaxEnt	0.9684	0.9638	0.9638	0.9643	0.9683	0.9688	0.9696	0.9602	0.9715	0.9725
F1 sports Bayes	0.9854	0.9703	0.9762	0.9678	0.9803	0.9837	0.9732	0.9704	0.9696	0.9826
F1 politics Bayes	0.9861	0.9692	0.9774	0.9644	0.9805	0.9842	0.9734	0.9725	0.9699	0.9818

```
MaxEntTrainer,gaussianPriorVariance=1.0
Summary. train accuracy mean = 0.9982231737511054 stddev = 1.5288097863905536E-4 stderr = 4.83452103414963 96E-5
Summary. test accuracy mean = 0.9671264135319811 stddev = 0.004013487191867437 stderr = 0.001269176088621432
Summary. test precision(politics) mean = 0.9662773850552238 stddev = 0.004974090628849755 stderr = 0.0015729455675264465
Summary. test precision(sports) mean = 0.9678600454393511 stddev = 0.007941843718472355 stderr = 0.00251143149714737
Summary. test recall(politics) mean = 0.9681245699566782 stddev = 0.006765525948383172 stderr = 0.002139447156586159
Summary. test recall(sports) mean = 0.9662427103059127 stddev = 0.004679369890756848 stderr = 0.0014797466869204928
Summary. test f1(politics) mean = 0.9671777356227415 stddev = 0.003708101634444576 stderr = 0.0011726046960237936
Summary. test f1(sports) mean = 0.9670275718631253 stddev = 0.0045038639751714135 stderr = 0.0014242468433121713

NaiveBayesTrainer
Summary. train accuracy mean = 0.9904248519736905 stddev = 7.18765570180707E-4 stderr = 2.2729363054806 37E-4
Summary. test accuracy mean = 0.9760121489478706 stddev = 0.006468396035630173 stderr = 0.0020454864280595
Summary. test precision(politics) mean = 0.973034589704222 stddev = 0.009777087549089751 stderr = 0.003091786553799693
Summary. test precision(sports) mean = 0.9789277623902859 stddev = 0.00944622776221125 stderr = 0.002987159502530298
Summary. test recall(politics) mean = 0.9790640688683302 stddev = 0.009164337459710209 stderr = 0.002898017961908583
Summary. test recall(sports) mean = 0.9731696908867464 stddev = 0.008849457211984525 stderr = 0.0027984440846074614
Summary. test f1(politics) mean = 0.9759955856806384 stddev = 0.0068038827747944845 stderr = 0.002151576650113704 3
Summary. test f1(sports) mean = 0.9759942886143907 stddev = 0.00622179994379696 stderr = 0.0019675058968306003
C:\mallet>bin
```



همانطور که در نمودار بالا مشخص است در حالت کلی نمودار مربوط به naiveBayes بالاتر از Max Entropy است یعنی f1 برای naive Bayes بهتر است. همانطور که در عکس بالا هم مشخص است

میانگین پارامتر های precision,recall, f1 برای naive Bayes تقریبا96 درصد است و برای Max Entropy تقریبا 97 درصد است.