بسمه تعالی

مقایسه naiveBayes classifier و maximum entropy classifier

ساختار و پوشه بندی کل پروژه :

پوشه src : شامل تمامی کد های اجرایی است.
پوشه data :شامل تمامی داده های جمع آوری شده یا ساخته شده است .
پوشه tools : شامل ابزار های استفاده شده در کد برنامه است.
پوشه analyze : شامل نمودار ها و نتایج بدست آمده از در هر فاز و نیز داکیومنت ها است.

فایل های مربوط به این فاز :

فایل p3_mallet_labeler.py کد مربوط به تولید داده ی برچسب خورده مناسب برای mallet میباشد که خروجی آن p3_all_comments_labeled.txt است که نمونه ای از خروجی آن را در زیر میبینیم :

خسته نباشی تاتر هم عالی بود قای کارگردان fr94 friendly
پیجتون خیلی قشنگههه fr95 friendly
دوستان عزیزم به پیج سلامتی ما هم یر بزنید fr97 friendly
دارم تلاش می کنم پروسس کنم ولی انصافا شبیه نیس به الانت... پیر شدی میلاد fr75 friendly

معمولا چه موقع هایی میذاره PS کسی میدونی دیجی کالا تخفیفای خوبشو برای celeb209 celebrity
معمولا چه موقع هایی میذاره PS کسی میدونی دیجی کالا تخفیفای خوبشو برای celeb210 celebrity
رو فقط تومن ارزون تر کن من بتونم بگیرم PS slim G دیجی جانه من celeb211 celebrity
رو فقط تومن ارزون تر کن من بتونم بگیرم PS slim G دیجی جانه من celeb212 celebrity

فایل p3_mallet_trainer حاوی دستور بش زیر برای آموزش دیتا هست :

```
../tools/mallet-2.0.8/bin/mallet train-classifier  --input ../data/
all_comments.mallet --trainer MaxEnt --trainer NaiveBayes   --training-portion
0.9
```

عدد جلوی trainning-portion نشان میدهد چه میزان از داده ها برای آموزش و چه میزان برای تست باشد .۹ یعنی .۱ تست داریم.
همچنین در این دستور میگوییم هم naive bayes و هم maxent را اموزش دهد تا بتوانیم مقایسه کنیم .

بسمه تعالی

نحوه اجرای کد :

وارد پوشه src میشویم

دستورات زیر را به ترتیت وارد میکنیم تا دیتا لیبل بخورد و لرن شود :

python3 p3_mallet_labeler.py. ->

Sudo ./p3_mallet_trainer.sh

نتیجه ی خروجی بر روی دیتای تست بدون 10 fold cross validation:

برای maxEnt

```
train accuracy mean = 0.9170518122111408 : دقت روی داده آموزشی
test accuracy mean = 0.8194748358862144 : دقت روی داده تست

MaxEntTrainer test precision(friendly) = 0.7631578947368421
MaxEntTrainer test precision(celebrity) = 0.8245823389021479
MaxEntTrainer test recall(friendly) = 0.28292682926829266
MaxEntTrainer test recall(celebrity) = 0.9746121297602257
MaxEntTrainer test F1(friendly) = 0.4128113879003559
MaxEntTrainer test F1(celebrity) = 0.8933419521654816
```

برای naive bayes

```
train accuracy mean = 0.8926052055460958 :دقت روی داده آموزشی
test accuracy mean = 0.8030634573304157 : دقت روی داده تست

test precision(friendly) mean = 0.5875
test precision(celebrity) mean = 0.8521970705725699
test recall(friendly) mean = 0.4585365853658537
test recall(celebrity) mean = 0.9026798307475318
test f1(friendly) mean = 0.5150684931506849
test f1(celebrity) mean = 0.8767123287671234
```

نتیجه ی خروجی بر روی دیتای تست با 10 fold cross validation:

برای max ent :

```
Summary. train accuracy mean = 0.9170315206858634
Summary. test accuracy mean = 0.8268373673728133
```

بسمه تعالی

```
Summary. test precision(friendly) mean = 0.8182229103103307
Summary. test precision(celebrity) mean = 0.8277738116588409
Summary. test recall(friendly) mean = 0.308198247002325
Summary. test recall(celebrity) mean = 0.9797536073232151
Summary. test f1(friendly) mean = 0.4465958866077601
Summary. test f1(celebrity) mean = 0.8972966000996706
```

برای naive bayes :

```
Summary. train accuracy mean = 0.8937658026964431
Summary. test accuracy mean = 0.8069159071136347
Summary. test precision(friendly) mean = 0.5999341995733974
Summary. test precision(celebrity) mean = 0.8527945138773365
Summary. test recall(friendly) mean = 0.4670944095517127
Summary. test recall(celebrity) mean = 0.9068449689979609
Summary. test f1(friendly) mean = 0.5249590454856257
Summary. test f1(celebrity) mean = 0.8789638389767257
```

مقایسه :

در دوستانه ها naivebayes . F1 بهتری دارد .

در سلبریتی ها maxent . F1 بهتری دارد.

در کل به نظرم naive bayes عملکرد بهتری داشته است.

در کل با 10 fold cross validation بهتر عمل کرده که بدیهی است.

خروجی نمونه بدون cross valiadtion:

```
Training portion = 0.9
Unlabeled training sub-portion = 0.0
Validation portion = 0.0
Testing portion = 0.09999999999999998

-------------------- Trial 0  --------------------

Trial 0 Training MaxEntTrainer,gaussianPriorVariance=1.0 with 8222
instances
Value (labelProb=5748.815170297192 prior=0.5000000000000548)
loglikelihood = -57Value (labelProb=4979.8322012052295
prior=1.3020917593817196) loglikelihood = -4Value
(labelProb=4628.190802834337 prior=2.2150179682578366) loglikelihood =
-46Value (labelProb=4356.796649309284 prior=3.9326426006393316)
loglikelihood = -43Value (labelProb=4116.896311686929
prior=8.387235664059467) loglikelihood = -412Value
(labelProb=3888.4448414240355 prior=16.512611239592104) loglikelihood
= -3Value (labelProb=3669.15770512907 prior=36.796701292356566)
loglikelihood = -370Value (labelProb=3393.259806492413
```

بسمه تعالی

prior=85.12861927324317) loglikelihood = -347Value
(labelProb=3130.8415309861584 prior=186.49697579074717) loglikelihood
= -3Value (labelProb=2962.3139648956358 prior=257.4795140056982)
loglikelihood = -32Value (labelProb=3096.666221917322
prior=363.39370482578164) loglikelihood = -34Value
(labelProb=2926.0888555303313 prior=278.0936319342232) loglikelihood =
-32Value (labelProb=2851.1125056734727 prior=315.4195266508251)
loglikelihood = -31Value (labelProb=2754.0514801870177
prior=361.26830675123364) loglikelihood = -3Value
(labelProb=2673.814872427057 prior=394.95391768247197) loglikelihood =
-30Value (labelProb=2560.414312264846 prior=429.1108954747781)
loglikelihood = -298Value (labelProb=2546.5504234614286
prior=418.62297477593904) loglikelihood = -2Value
(labelProb=2440.1334815681257 prior=435.56246118247486) loglikelihood
= -2Value (labelProb=2453.242465753934 prior=409.58500754655597)
loglikelihood = -28Value (labelProb=2430.7712918893944
prior=384.4170598012717) loglikelihood = -28Value
(labelProb=2397.973200401434 prior=380.1159865998201) loglikelihood =
-277Value (labelProb=2376.920135758225 prior=381.5702999425193)
loglikelihood = -275Value (labelProb=2353.246736891487
prior=390.300498820112) loglikelihood = -2743Value
(labelProb=2312.826555898296 prior=396.78268221238255) loglikelihood =
-27Value (labelProb=2290.598001209369 prior=403.29180244897356)
loglikelihood = -26Value (labelProb=2272.42974661409
prior=411.80471599224944) loglikelihood = -268Value
(labelProb=2251.960413466092 prior=419.3611765337051) loglikelihood =
-267Value (labelProb=2183.066729651215 prior=447.6210945912519)
loglikelihood = -263Value (labelProb=2154.150806789476
prior=479.0140093918152) loglikelihood = -263Value
(labelProb=2155.7706190100453 prior=461.91984650544464) loglikelihood
= -2Value (labelProb=2114.008212580218 prior=483.5564555614802)
loglikelihood = -259Value (labelProb=2111.308393918082
prior=480.8773108701163) loglikelihood = -259Value
(labelProb=2098.034448026632 prior=486.5506008221182) loglikelihood =
-258Value (labelProb=2097.150558806155 prior=483.487707358744)
loglikelihood = -2580Value (labelProb=2088.8227423750177
prior=487.0448851582701) loglikelihood = -25Value
(labelProb=2133.6268807740007 prior=508.85190869307536) loglikelihood
= -2Value (labelProb=2086.054479085769 prior=489.11793285465006)
loglikelihood = -25Value (labelProb=2078.0012240141873
prior=495.0190400027815) loglikelihood = -25Value
(labelProb=2073.8906543896496 prior=497.7807928622119) loglikelihood =
-25Value (labelProb=2068.7461123945172 prior=501.17706000862194)
loglikelihood = -2Value (labelProb=2064.4824464842172
prior=503.94901088737464) loglikelihood = -2Value
(labelProb=2055.7202602508382 prior=510.8169917385251) loglikelihood =
-25Value (labelProb=2049.2192815101807 prior=515.643491241807)
loglikelihood = -256Value (labelProb=2048.374407242819
prior=515.3242359577148) loglikelihood = -256Value

بسمه تعالی

```
(labelProb=2046.8445459070904 prior=515.6240393095479) loglikelihood =
-25Value (labelProb=2041.7985268001319 prior=519.2285183287275)
loglikelihood = -25Value (labelProb=2032.8316637371174
prior=527.596424244224) loglikelihood = -256Value
(labelProb=2031.6541908142588 prior=529.0457246724778) loglikelihood =
-25Value (labelProb=2031.713029088696 prior=528.2361410776576)
loglikelihood = -255Value (labelProb=2033.1159818963679
prior=526.655698098899) loglikelihood = -2559.771679995267
Exiting L-BFGS on termination #1:
value difference below tolerance (oldValue: -2559.9491701663537
newValue: -2559.771679995267
Value (labelProb=2211.4862412774696 prior=526.8940953394621)
loglikelihood = -27Value (labelProb=2034.1866223446084
prior=526.6345378229698) loglikelihood = -25Value
(labelProb=2033.1378278114278 prior=526.643867960963) loglikelihood =
-255Value (labelProb=2032.9974967471244 prior=526.6494705299446)
loglikelihood = -25Value (labelProb=2032.82952276942
prior=526.6352077080843) loglikelihood = -2559.4647304775044
Exiting L-BFGS on termination #1:
value difference below tolerance (oldValue: -2559.646967277069
newValue: -2559.4647304775044

Trial 0 Training MaxEntTrainer,gaussianPriorVariance=1.0 finished

No examples with predicted label !
No examples with true label !
No examples with predicted label !
No examples with true label !
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 training data
accuracy = 0.9170518122111408
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 Test Data
Confusion Matrix
Confusion Matrix, row=true, column=predicted
accuracy=0.8194748358862144 most-frequent-tag
baseline=0.7757111597374179
        label   0   1   2  |total
  0  friendly  58   . 147  |205
  1             .   .   .  |0
  2 celebrity  18   . 691  |709

Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
precision(friendly) = 0.7631578947368421
No examples with predicted label !
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
precision() = 1.0
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
precision(celebrity) = 0.8245823389021479
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
recall(friendly) = 0.28292682926829266
```

بسمه تعالی

```
No examples with true label !
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
recall() = 1.0
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
recall(celebrity) = 0.9746121297602257
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
F1(friendly) = 0.4128113879003559
No examples with predicted label !
No examples with true label !
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data F1()
= 1.0
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
F1(celebrity) = 0.8933419521654816
Trial 0 Trainer MaxEntTrainer,gaussianPriorVariance=1.0 test data
accuracy = 0.8194748358862144
Trial 0 Training NaiveBayesTrainer with 8222 instances
Trial 0 Training NaiveBayesTrainer finished
No examples with true label !
No examples with true label !
Trial 0 Trainer NaiveBayesTrainer training data accuracy =
0.8926052055460958
Trial 0 Trainer NaiveBayesTrainer Test Data Confusion Matrix
Confusion Matrix, row=true, column=predicted
accuracy=0.8030634573304157 most-frequent-tag
baseline=0.7757111597374179
        label   0   1   2  |total
  0  friendly  94   . 111  |205
  1             .   .   .  |0
  2 celebrity  66   3 640  |709

Trial 0 Trainer NaiveBayesTrainer test data precision(friendly) =
0.5875
Trial 0 Trainer NaiveBayesTrainer test data precision() = 0.0
Trial 0 Trainer NaiveBayesTrainer test data precision(celebrity) =
0.8521970705725699
Trial 0 Trainer NaiveBayesTrainer test data recall(friendly) =
0.4585365853658537
No examples with true label !
Trial 0 Trainer NaiveBayesTrainer test data recall() = 1.0
Trial 0 Trainer NaiveBayesTrainer test data recall(celebrity) =
0.9026798307475318
Trial 0 Trainer NaiveBayesTrainer test data F1(friendly) =
0.5150684931506849
No examples with true label !
Trial 0 Trainer NaiveBayesTrainer test data F1() = 0.0
Trial 0 Trainer NaiveBayesTrainer test data F1(celebrity) =
0.8767123287671234
Trial 0 Trainer NaiveBayesTrainer test data accuracy =
0.8030634573304157
```

بسمه تعالی

```
MaxEntTrainer,gaussianPriorVariance=1.0
Summary. train accuracy mean = 0.9170518122111408 stddev = 0.0 stderr
= 0.0
Summary. test accuracy mean = 0.8194748358862144 stddev = 0.0 stderr =
0.0
Summary. test precision(friendly) mean = 0.7631578947368421 stddev =
0.0 stderr = 0.0
Summary. test precision() mean = 1.0 stddev = 0.0 stderr = 0.0
Summary. test precision(celebrity) mean = 0.8245823389021479 stddev =
0.0 stderr = 0.0
Summary. test recall(friendly) mean = 0.28292682926829266 stddev = 0.0
stderr = 0.0
Summary. test recall() mean = 1.0 stddev = 0.0 stderr = 0.0
Summary. test recall(celebrity) mean = 0.9746121297602257 stddev = 0.0
stderr = 0.0
Summary. test f1(friendly) mean = 0.4128113879003559 stddev = 0.0
stderr = 0.0
Summary. test f1() mean = 1.0 stddev = 0.0 stderr = 0.0
Summary. test f1(celebrity) mean = 0.8933419521654816 stddev = 0.0
stderr = 0.0

NaiveBayesTrainer
Summary. train accuracy mean = 0.8926052055460958 stddev = 0.0 stderr
= 0.0
Summary. test accuracy mean = 0.8030634573304157 stddev = 0.0 stderr =
0.0
Summary. test precision(friendly) mean = 0.5875 stddev = 0.0 stderr =
0.0
Summary. test precision() mean = 0.0 stddev = 0.0 stderr = 0.0
Summary. test precision(celebrity) mean = 0.8521970705725699 stddev =
0.0 stderr = 0.0
Summary. test recall(friendly) mean = 0.4585365853658537 stddev = 0.0
stderr = 0.0
Summary. test recall() mean = 1.0 stddev = 0.0 stderr = 0.0
Summary. test recall(celebrity) mean = 0.9026798307475318 stddev = 0.0
stderr = 0.0
Summary. test f1(friendly) mean = 0.5150684931506849 stddev = 0.0
stderr = 0.0
Summary. test f1() mean = 0.0 stddev = 0.0 stderr = 0.0
Summary. test f1(celebrity) mean = 0.8767123287671234 stddev = 0.0
stderr = 0.0
```