بسم الله الرحمن الرحیم

فاز سوم تمرین دوم درس NLP

برای این فاز ابتدا چند فیچر را در نظر گرفتیم. این فیچرها شامل کلمه اول مصرء اول و دوم بیت اول هر نمونه، بیست کلمه به دست آمده از وردمپ فاز اول پروژه و نقش کلمه اول و آخر هر نمرنه هستند.

سپس با فرمت زیر این اطلاعات را در فایل data.txt می نویسیم:

InstanceName label Feature1Name=Feature1Value
Feature2Name=Feature2Value …

برای مثال یک خط این فایل به شکل زیر می باشد :

19 molana f1= خامش f2= کاغذ f3=0 f4=0 f5=0 f6=0 f7=0 f8=0 f9=0
f10=0 f11=0 f12=0 f13=0 f14=0 f15=0 f16=0 f17=0 f18=0 f19=1
f20=0 f21=0 f22=0 f23=N f23=N

اعداد صفر و یک نشان دهنده وجود و یا عدم وجود واژه در نظر گرفته در نمونه می باشد.

پس از ساختن این فایل در cmd با زدن دستور زیر فایلی با فرمت mallet می سازیم :

bin\mallet import-file --input data.txt --output data.mallet

و بعد از آن با زدن دستور زیر نتایج precision و recall و f1 را برای دو الگوریتم NaiveBayes و MaxEnt به دست می آوریم :

bin\mallet train-classifier --input data.mallet --training-portion 0.9
--trainer MaxEnt --trainer NaiveBayes

نتایج به دست آمده بدون در نظر گرفتن دو فیچر نقش کلمات اول آخر به شکل زیر می باشد :

```
Trial 0 Trainer NaiveBayesTrainer test data precision(hafez) = 0.6415094339622641
Trial 0 Trainer NaiveBayesTrainer test data precision(molana) = 0.7083333333333334
Trial 0 Trainer NaiveBayesTrainer test data recall(hafez) = 0.7445255474452555
Trial 0 Trainer NaiveBayesTrainer test data recall(molana) = 0.5985915492957746
Trial 0 Trainer NaiveBayesTrainer test data F1(hafez) = 0.6891891891891891
Trial 0 Trainer NaiveBayesTrainer test data F1(molana) = 0.648854961832061
Trial 0 Trainer NaiveBayesTrainer test data accuracy = 0.6702508960573477

MaxEntTrainer,gaussianPriorVariance=1.0
Summary. train accuracy mean = 0.8354531001589826 stddev = 0.0 stderr = 0.0
Summary. test accuracy mean = 0.6523297491039427 stddev = 0.0 stderr = 0.0
Summary. test precision(hafez) mean = 0.6234567901234568 stddev = 0.0 stderr = 0.0
Summary. test precision(molana) mean = 0.6923076923076923 stddev = 0.0 stderr = 0.0
Summary. test recall(hafez) mean = 0.7372262773722628 stddev = 0.0 stderr = 0.0
Summary. test recall(molana) mean = 0.5704225352112676 stddev = 0.0 stderr = 0.0
Summary. test f1(hafez) mean = 0.6755852842809364 stddev = 0.0 stderr = 0.0
Summary. test f1(molana) mean = 0.6254826254826253 stddev = 0.0 stderr = 0.0

NaiveBayesTrainer
Summary. train accuracy mean = 0.8211446740858506 stddev = 0.0 stderr = 0.0
Summary. test accuracy mean = 0.6702508960573477 stddev = 0.0 stderr = 0.0
Summary. test precision(hafez) mean = 0.6415094339622641 stddev = 0.0 stderr = 0.0
Summary. test precision(molana) mean = 0.7083333333333334 stddev = 0.0 stderr = 0.0
Summary. test recall(hafez) mean = 0.7445255474452555 stddev = 0.0 stderr = 0.0
Summary. test recall(molana) mean = 0.5985915492957746 stddev = 0.0 stderr = 0.0
Summary. test f1(hafez) mean = 0.6891891891891891 stddev = 0.0 stderr = 0.0
Summary. test f1(molana) mean = 0.648854961832061 stddev = 0.0 stderr = 0.0
```

	Precision (hafez)	Precision (molana)	Recall (hafez)	Recal (molana)	F1 (hafez)	F1 (molana)
Naïve Bayes	0.64	0.70	0.74	0.59	0.68	0.64
MaxEnt	0.62	0.69	0.73	0.57	0.67	0.62

مشاهده می شود مقادیر در الگوریتم Naïve Bayes از MaxEnt بیشتر است.

و در آخر برای cross validation دستور زیر را میزنیم :

bin\mallet train-classifier --input data.mallet --training-portion 0.9 --trainer MaxEnt --trainer NaiveBayes --cross-validation 10

با این کار 10 بار مقادیر f1 و precision و recall را به دست می آوریم که نتایج به شکل زیر می باشد :

مقادیر f1 برای کلاس مولانا

	1	2	3	4	5	6	7	8	9	10
Naïve Bayes	0.68	0.59	0.66	0.66	0.65	0.64	0.66	0.61	0.62	0.65
MaxEnt	0.70	0.59	0.61	0.68	0.64	0.62	0.64	0.61	0.61	0.61

مقادیر f1 برای کلاس حافظ

	1	2	3	4	5	6	7	8	9	10
Naïve Bayes	0.73	0.68	0.72	0.72	0.69	0.69	0.66	0.64	0.70	0.66
MaxEnt	0.74	0.69	0.72	0.72	0.69	0.71	0.69	0.62	0.68	0.64