Normalization

برای این کار ابتدا از ماژول hazm استفاده کردیم و متد hazm.Normalizer().normalize، سپس خروجی آن را دوباره چک کردیم و علائم نگارشی، حروف اضافه و برخی از ضمایر را نیز حذف کردیم.

Tokenization

برای این کار نیز از ماژول hazm استفاده کردیم و دو متد hazm.word_tokenize ،hazm.sent_tokenize که یکی به صورت کلمه به کلمه توکن می‌کند و دیگری به صورت جمله.
از اولی برای ساخت دیکشنری کلمات استفاده کردیم و از دومی برای ساخت داده‌های train و test.

Naive Bayes

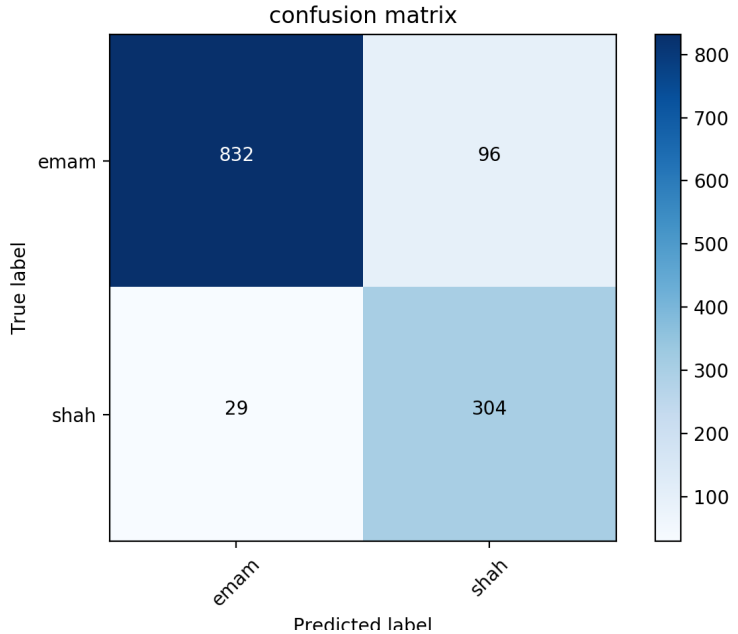
در ابتدا داده‌های test و train را به صورت یک document کامل به کلسیفایر می‌دادیم و دقت تقریبا ۱۰۰٪ می‌گرفتیم ولی بعد دیتا را به صورت جمله دادیم و نتایج منطقی تر شد.

نتایج :

number of emam test sentences -> 928
emam precision -> 0.9663182346109176
emam recall -> 0.896551724137931
emam fscore -> 0.9301285634432644

number of shah test sentences -> 333
shah precision -> 0.76
shah recall -> 0.9129129129129129
shah fscore -> 0.8294679399727148

Accuracy :(number of true predicts/ total predictions) = 0.9008723235527359



کیارش آقاکثیری
۹۴۵۲۳۰۴۵

Vowpal Wabbit

برای استفاده از vowpal wabbit ابتدا داده‌ها را به فرمت مناسب vowpal wabbit تبدیل کردیم و سپس با استفاده از دستور زیر آن‌ها را train کردیم:

```
vw -d Train.txt -c —passes 10 -f predictor.vw --ngram n —loss_function quantile
```

البته هر دفعه که برنامه اجرا می‌شود ابتدا cache را خالی می‌کنیم تا train دوباره انجام شود.

برای test نیز از دستور زیر استفاده کردیم:

```
vw -d Test.txt -t -i predictor.vw -p prediction.txt
```

خروجی پیش‌بینی‌ها در فایل prediction.txt اعدادی بین ۰ و ۱ (بستگی به کلاس‌هایی که خودمان نام‌گذاری کردیم دارد).
سپس خروجی‌ها را با ۰.۵ (یا هر عدد دیگری) جدا می‌کنیم و برچسب می‌زنیم و نتایج به صورت زیر شد.

1-gram :

number of emam test sentences vw-> 333
emam precision vw -> 0.90625
emam recall vw-> 0.26126126126126126
emam fscore vw-> 0.40559440559440557

number of shah test sentences vw-> 333
shah precision vw-> 0.5684210526315789
shah recall vw-> 0.972972972972973
shah fscore vw-> 0.717607973421927

Accuracy :(number of true predicts/ total predictions) = 0.6171171171171171

2-gram:

number of emam test sentences vw-> 305
emam precision vw -> 0.8533333333333334
emam recall vw-> 0.2098360655737705
emam fscore vw-> 0.3368421052631579

number of shah test sentences vw-> 305
shah precision vw-> 0.5495327102803739
shah recall vw-> 0.9639344262295082
shah fscore vw-> 0.7000000000000001

Accuracy :(number of true predicts/ total predictions) = 0.5868852459016394

3-gram:

number of emam test sentences vw-> 235
emam precision vw -> 0.8783783783783784
emam recall vw-> 0.2765957446808511
emam fscore vw-> 0.4207119741100324

number of shah test sentences vw-> 235
shah precision vw-> 0.5707070707070707
shah recall vw-> 0.9617021276595744
shah fscore vw-> 0.7163232963549919

Accuracy :(number of true predicts/ total predictions) = 0.6191489361702127

متاسفانه خروجی‌ای که از vowpal wabbit گرفتیم ضعیف‌تر از naive bayes بود.

برای کار کردن با برنامه ابتدا پارامترهای ورودی کلاس را تعریف کنید و سپس یک instance از آن بسازید و برنامه را اجرا کنید.