به نام خدا

محمدمهدی آقاجانی

۹۳۳۱۰۵۶

تمرین سوم

استاد : دکتر ممتازی

گزارش

در ابتدای کار برای راحتی کار با دیتاست داده شده آن را کمی تمیز کردیم به طوریکه اطلاعات اضافی از آن پاک شد و تنها جملات به همراه قطبیت آنها در فایل دیتاست موجود می باشند . همچنین این دو موجودیت با علامت @ از یکدیگر جدا شده اند. گزارش این تمرین برای موارد خواسته شده به صورت زیر است:

الف) baysian unigram

هنگام خواندن داده و موارد بی استفاده جمله ها از قبیل علامات . یا ، یا : یا ; حذف شدند که همین امر باعث بهبود نسبی نتایج گردید زیرا در غیر آن صورت برخی کلمات یکبار به همراه برخی از این علامات و یکبار هم به طور جداگانه در محاسبات ظاهر می شدند. مورد دیگر که موجب بهبود نتایج گردید استفاده از تکنیک smoothing به این صورت بود که به جای حالاتی که مقدار unigram برابر صفر میشد عددی ثابت را ضرب کردیم که این عدد هم به طور دستی و با جست و جوی دو دویی حریصانه بدست آمد و مقدار آن برای این حالت 0.00024 می باشد.

نتایج بدست آمده برای این روش به شرح زیر است :

```
NEG:

Precision :  0.8565400843881856
Recall :  0.8864628820960698
F-measure :  0.871244635193133
----------------------------
POS:

Precision :  0.9078014184397163
Recall :  0.8827586206896552
F-measure :  0.8951048951048951
============================
Accuracy :  0.884393063583815
```

همانطور که مشاهده می شود دقت بدست آمده حدود دو درصد از دقتی که در مقاله داده شده بدست آمده بود بهتر است

ب) Baysian Bigram

در این حالت هم همانند حالت قبل موارد رعایت شده است. برای حالتی که BIGRAM برابر صفر می شود ابتدا مقدار UNIGRAM آن را قرار دادیم و دقت ۸۲ درصد حاصل شد ولی در حالتی که همانند حالت قبل مقدار ثابتی که با جست و جوی دودویی بدست آمده را جایگزین کردیم دقت ۸۵ درصد حاصل شد(عدد ثابت برابر با 0.0068 می باشد).

نتایج حاصله در این حالت به صورت زیر است:

```
NEG:

Precision :  0.8275862068965517
Recall :  0.8384279475982532
F-measure :  0.8329718004338393
----------------------------
POS:

Precision :  0.8710801393728222
Recall :  0.8620689655172413
F-measure :  0.8665511265164645
============================
Accuracy :  0.8516377649325626
```

ج) SVM with Unigram

برای این حالت هم همایننند حالت های قبلی دیتا را می خوانیم و بعد برای انتخاب ویژگی ها از mutual information استفاده می کنیم . این کار در مقایسه با حالتی که ۱۰۰ کلمه پرتکرار را انتخاب میکنیم دقت بهتری را بدست می دهد. در حالت اول دقت حدود ۸۸ درصد بود ولی با انتخاب حالت mutual information و انتخاب ۱۰۰ کلمه ای که بیشترین امتیاز را کسب کرده اند دقت به نزدیک ۹۱ درصد می رسد که این عدد در مقایسه با مقاله داده شده حدود ۴ درصد بهتر است.

نتایج این روش به صورت زیر است :

```
NEG:

Precision :  0.8922413793103449
Recall    :  0.9039301310043668
F-measure :  0.8980477223427332
----------------------------
POS:

Precision :  0.9233449477351916
Recall    :  0.9137931034482759
F-measure :  0.9185441941074522
============================
Accuracy :  0.9094412331406551
```

د) SVM with Bigram

این حالت مانند قبلی می باشد با این تفاوت که هنگام انتخاب ویژگی ها هم داده های unigram و هم bigram را به صورت mutual info حساب میکنیم و از این بین ۱۰۰ عبارتی که بیشترین امتیاز را داشته باشند انتخاب میکنیم. این روش هم همانند روش قبلی دقتی در حدود ۹۱ درصد بدست می دهد.

نتایج این روش به صورت زیر است:

```
NEG:

Precision :  0.8951965065502183
Recall :  0.8951965065502183
F-measure :  0.8951965065502182
----------------------------
POS:

Precision :  0.9172413793103448
Recall :  0.9172413793103448
F-measure :  0.9172413793103448
============================
Accuracy :  0.9075144508670521
```