آرش اعلائی - الهه صادقی

گزارش پروژه چهارم: پردازش زبان های طبیعی(NLP)

<<تاثیر حذف کلمات پرتکرار و کم تکرار در دقت بدست آمده>>

کلمات پرتکرار و کم تکرار باعث میشوند احتمال محاسبه شده از مقدار موثر خود فاصله بگیرد در حالی که این کلمات هیچ نقشی در نتیجه بدست آمده ندارند مثل کلمات پایه، ربط .... پس حذف این کلمات باعث میشود مدل کلاس بندی بهتری ارائه دهد

<<تاثیر مقدار λ و ϵ دقت بدست آمده>>

یک روش خوب برای محاسبه احتمالات استفاده از interpolation که در آن احتمال n-gram های مختلف را جمع میکند و به هر مدل وزن λ را اختصاص می دهد و برای جلوگیری از صفر شدن از یک ϵ استفاده می کند. برای مدل بایگرام بدیهی است مدل Bigram کمی بهتر از مدل unigram عمل می کند پس باید وزن بیشتری داشته باشد یعنی λ(bigram) >λ(unigram) > $\lambda(\epsilon)$ که جمعشان ۱ است در مورد ϵ هم هرچه مقدار به مقدار کمتری میل کند بهتر است.اما باید توجه داشت ممکن است در یک حوزه ای مدل unigramکمی بهتر از مدل Bigram عمل می کند برای بدست آوردن بهترین وزن ها می توان از گرادیان کاهشی استفاده کرد به شرط آنکه تابع ارزیابی محدب باشد.

<<بهترین دقت دستیافته و تحلیل تاثیر پارامترها در آن>>

```
*** Result of evaluating bigram model***
F1-Score: 0.9971848608070066
Accuracy: 0.9971901342491414
Precision: 0.9987468671679198
Recall: 0.995627732667083
specificity: 0.9987515605493134
##############################################
```

شکل بالا بهترین نتایج مدل را نشان می دهد که در آن

λ3=0.75, λ2=0.15, λ1=0.1, ε =0.1

و با توجه به توضیحات داده شده در قسمت ۲ نشان میدهد که مدل Bigram کمی بهتر از مدل unigram عمل می کند پس باید وزن بیشتری داشته باشد یعنی

λ(bigram) >λ(unigram) > $\lambda(\epsilon)$