تاثیر حذف کلمات پرتکرار و کم تکرار در دقت بدست آمده چیست؟

به طور کلی اگر به تعداد مناسب کلمات خیلی پرتکرار یا خیلی کم تکرار را حذف کنیم و داده های train به اندازه کافی باشند، باعث افزایش دقت بدست آمده میشود. اما اگر بیش از حد کلمات را حذف کنیم باعث کاهش دقت خواهد شد.همچنین به طور کلی حذف کلمات پرتکرار باعث افزارش سرعت میشود.

تاثیر مقدار λ و ε دقت بدست آمده چیست؟

با توجه به فرمول زیر هرچه مقدار **ε** کمتر باشد بهتر است و در واقع مدل بیشتر احتمالات بدست آمده در قسمت train را در تصمیم گیری اثر میدهد و کمتر شانسی تصمیم میگیرد.

همچنین هرچه مقدار لاندا۳ بزرگتر ار دو لاندای دیگر باشد و هرچه مقدار لاندا۲ بزرگتر از لانداا۱ باشد، دقت بالاتر خواهد بود. چون هرچه لانداا۳ بیشتر باشد یعنی به وابستگی های بین کلمات توجه بیشتری شده است و در محاسبه احتمال نهایی تاثیر بیشتری دارد. و هرچه لانداا۲ بیشتر از لانداا۱ باشد یعنی کمتر شانسی تصمیم گرفته ایم و به احتمال حضور تک کلمه ها در یک جمله اهمیت بیشتری دادیم .

$$P(w_i|w_{i-1}) = \lambda_3 P(w_i|w_{i-1}) + \lambda_2 P(w_i) + \lambda_1 \epsilon$$

$$\lambda_3 + \lambda_2 + \lambda_1 = 1$$

$$0 < \epsilon < 1$$

بهترین دقت دست یافته و تحلیل تاثیر پارامتر ها در آن چیست؟

```
BIGRAM
percision:  73.58490566037736
Clean
percision:  69.81132075471697
UNIGRAM:
percision:  65.09433962264151
Clean
percision:  64.15094339622641
```

در این حالت در مدل بایگرام:

ε = 0.00001

λ1= 0.005

Λ2=0.1

λ3= 0.895

تحلیل : همانطور که در سوال بالا توضیح داده شد، هرچه ضریب احتمال کلمات دوتایی بیشتر باشد و همچنین اپسیلون کوچک تر باشد به دقت بالاتری میرسیم.