آرمین رحمتی
۴۰۳۱۳۱۰۷۲

پردازش زبان طبیعی دکتر ممتازی

تمرین 1

کد به همراه کامنت و همچنین توضیحات به صورت markdown در فایل nlp-tamrin1.ipynb مشاهده است در اینجا نتیجه ران شده کد را نیز به صورت اسکرین شات نشان میدهم.

نکته:

لپ تاپ بنده کارت گرافیک درست حسابی نداره همینطور کامپیوتر های آزمایشگاه هم همین طور که بنده کد را روی آنها اجرا کردم که با صرفا تست 10 درصد از داده های دیتاست برای cnn نتایج حدودی زیر به دست آمده که شاید خیلی طبیعی نباشد اما همین نتایج حدود ۱۰ ساعت بعد ران شدن پروژه بدست آمده برای مثال انتظار می رود bigram و trigram میزان پرپلکسیتی کمتری رو داشته باشند و مدل هایی بهتری باشند اما به خاطر نداشتن توان پردازشی لازم و استفاده از بخشی از دیتاست و مواردی چون sparsity نتایج زیر را مشاهده کردیم.

همین طور شایان ذکر است که نتیجه پیش بینی قسمت اخر در فایل

sample_output.csv که دو ستون پیشبینی تنها یک کلمه و پیش بینی چند کلمه به آن اضافه شده اضافه شده است قابل رویت است

```
[Unigram] best alpha = 0.01, val perplexity = 1308.1129901488932
[Bigram] best discount = 0.75, val perplexity = 204.19341350665897
[Trigram] best discount = 0.75, val perplexity = 6082.749435511439

Train perplexities: 1651.5449061453303 73.5026099910924 3.9270501360394254
```

همان طور که قابل حدس بود بایگرم در اینجا نتیجه بهتری از یونیگرم دارد اما ترایگرم بر خلاف انتظار و به دلیل حجم کم دیتاست val نتیجه خوبی را نشان نمیده که طبیعی است و به داده بیشتری نیاز داریم.

نتایج خط اخر حالت نرمال صرفا روی داده های train میباشد.

```
=== TEST PERPLEXITIES ===
Unigram model perplexity: 1312.1047027555612
Bigram  model perplexity: 204.06683231216311
Trigram model perplexity: 5738.741642152637

Bigram is best on test set.
```

For n=1, best LR=0.01, perplexity=34388.00000004098

For n=2, best LR=0.01, perplexity=34387.999931053615

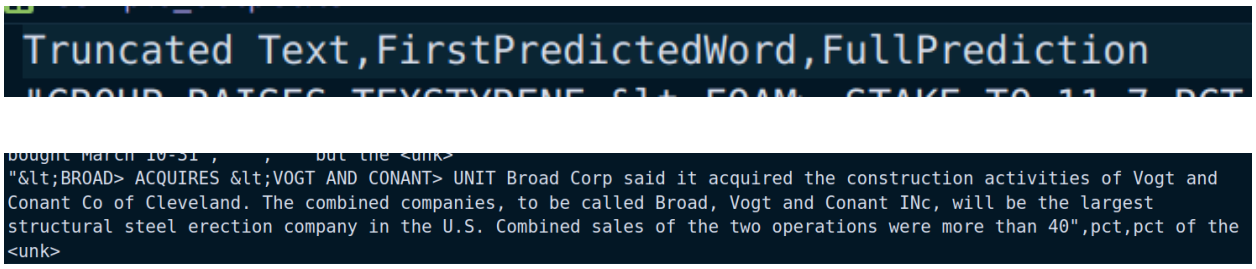
For n=3, best LR=0.1, perplexity=34387.99996559879

```
=== TEST PERPLEXITIES (Feed-Forward) ===
Unigram   => 34388.00000038776
Bigram    => 34387.99991768191
Trigram   => 34387.999920531416


Now comparing with best n-gram results from Section 1:
Unigram n-gram perplexity:  1312.1047027555612
Bigram n-gram perplexity:   204.06683231216311
Trigram n-gram perplexity:  5738.741642152637
```

```
=> The best overall model is: Bigram N-gram with perplexity = 204.06683231216311
```

Sample_output.csv:



```
Truncated Text,FirstPredictedWord,FullPrediction
"&lt;BROAD> ACQUIRES &lt;VOGT AND CONANT> UNIT Broad Corp said it acquired the construction activities of Vogt and
Conant Co of Cleveland. The combined companies, to be called Broad, Vogt and Conant INc, will be the largest
structural steel erection company in the U.S. Combined sales of the two operations were more than 40",pct,pct of the
<unk>
```

همان طور که مشخص است مدل ما تداوم حرف pct و حرف ناشناخته unk را در سطر آخر برای این تیکه از متن پیش بینی کرده است.