



Amirkabir University of Technology
(Tehran Polytechnic)

NLP

hw-1

Zahra Zanjani

401131025

بخش اول.....	3
پیش پردازش دادگان.....	3
laplace smooth ارزیابی مدل های اماری با.....	3
mask پیش بینی تکمیل فایل.....	4
بررسی نتایج.....	9
بخش دوم.....	10
پیش پردازش دادگان.....	10
incomplete و mask نتایج پیش بینی کلمات فایل.....	11
بخش سوم.....	13
پیش پردازش.....	13
شبکه ی عصبی.....	14
نتایج.....	16
incomplete و mask نتایج پیش بینی کلمات فایل.....	17
بررسی کلی نتایج.....	18

پیش بینی تکمیل فایل mask

پیش بینی مدل unigram with laplace smoothing برای جای خالی فایل mask به صورت زیر می باشد.

و predicted_word:
سارتر به عنوان روشن فکری فعال از نظر و

و predicted_word:
این تیم در سال ۱۸۹۹ و

و predicted_word:
شد و تا به حال موفق به کسب یک عنوان و

و predicted_word:
بزرگترین کلیسای و

و predicted_word:
کریم خان پس از و

و predicted_word:
بر دشمنان خود و

و predicted_word:
عمومی ترین و

و predicted_word:
جانوران که طی مراحل و

و predicted_word:
نیز پیش از همه و

پیش بینی مدل bigram with laplace smoothing برای جای خالی فایل mask به صورت زیر می باشد.

predicted_word: ان
سارتر به عنوان روشن فکری فعال از نظر ان

predicted_word: های
این تیم در سال ۱۸۹۹ های

predicted_word: سال
شد و تا به حال موفق به کسب یک عنوان سال

predicted_word: شهر
بزرگترین کلیسای شهر

predicted_word: از
کریم خان پس از از

predicted_word: محمدشاه
بر دشمنان خود محمدشاه

predicted_word: و
عمومی ترین و

predicted_word: ان
جانوران که طی مراحل ان

predicted_word: ان
نیز پیش از همه ان

پیش بینی مدل trigram with laplace smoothing برای جای خالی فایل mask به صورت زیر می باشد.

predicted_word: Null

سارتر به عنوان روشن فکری فعال از نظر

predicted_word: Null

این تیم در سال ۱۸۹۹

predicted_word: Null

شد و تا به حال موفق به کسب یک عنوان

predicted_word: Null

بزرگترین کلیسای

predicted_word: Null

کریم خان پس از

predicted_word: Null

بر دشمنان خود

predicted_word: نشان

عمومی ترین نشان

predicted_word: Null

جانوران که طی مراحل

predicted_word: این

نیز پیش از همه این

خروجی درست فایل mask به صورت زیر می باشد.

سارتر به عنوان روشن فکری فعال از نظر سیاسی شناخته می شد


این تیم در سال ۱۸۹۹ تاسیس شد و تا به حال موفق به کسب یک عنوان قهرمانی شد

بزرگترین کلیسای مسیحیت در جهان است

کریم خان پس از بیروزی بر دشمنان خود شهرت زیادی کسب نمود

عمومی ترین صفت های جانوران که طی مراحل رشد نیز پیش از همه ظاهر می شود.

پیش بینی مدل unigram with laplace smoothing برای فایل incomplete به صورت زیر می باشد

predicted_word: و	
در جریان انقلاب مشروطه ابتدا به عنوان یکی از نیروهای محمدعلی شاه با و	

predicted_word: و
شرکت خدمات مالی و و

predicted_word: و
شخص موسی تورات را و

predicted_word: و
نام آمازون را از یک لغت نامه و

predicted word: و
تیم سپاهان اصفهان که در ابتدا شعبه و

پیش بینی مدل bigram with laplace smoothing برای فایل incomplete به صورت زیر می باشد

عباس
در جریان انقلاب مشروطه ابتدا به عنوان یکی از نیروهای محمدعلی شاه با
شرکت خدمات مالی و
به
شخص موسی تورات را
ایتالیایی
نام آمازون را از یک لغت نامه ایتالیایی
به
تیم سپاهان اصفهان که در ابتدا شعبه به

پیش بینی مدل trigram with laplace smoothing برای فایل incomplete به صورت زیر می باشد

Null
در جریان انقلاب مشروطه ابتدا به عنوان یکی از نیروهای محمدعلی شاه با
شرکت خدمات مالی و
Null
شخص موسی تورات را
Null
نام آمازون را از یک لغت نامه
Null
تیم سپاهان اصفهان که در ابتدا شعبه

خروجی درست فایل incomplete به صورت زیر می باشد.

در جریان انقلاب مشروطه، ابتدا به عنوان یکی از نیروهای محمدعلی شاه با مشروطه خواهان جنگید
شرکت خدمات مالی و بانکداری
شخص موسی تورات را نوشته
نام آمازون را از یک لغت نامه انتخاب کرد
تیم سپاهان اصفهان که در ابتدا شعبه شاهین تهران

بررسی نتایج

در unigram بدون توجه به کلمات قبلی خروجی تولید میشود و در این مجموعه کلمات 'و' بیشترین احتمال را داشته است و خروجی تمام جملات 'و' است.

علت اینکه بیشتر خروجی های trigram به صورت null است این است که ترکیب دو کلمه ی قبل از ماسک با هریک از کلمه های دیتاست در trigram وجود ندارد .

بخش دوم

پیش پردازش دادگان

کد این بخش در فایل NLP_HW2.ipynb می باشد.

پیش پردازش دادگان این بخش مانند بخش اول است.

در این بخش kneser-ney برای trigram پیاده سازی و بررسی شده است.

هایپر پارامتر d که discount نامیده میشود و میتواند مقادیر بین صفر و یک داشته باشد با استفاده از 50 جمله ی اول فایل valid.txt پیدا شده است. تاثیر این پارامتر بر perplexity به شرح زیر است. که نشان میدهد تغییر متغیر discount تاثیری بر روی perplexity ندارد.

discount	perplexity
0.5	4154.021894131072
0.7	4154.021894131072
0.9	4154.021894131072

و نتایج به شرح زیر است:

	perplexity
kneser ney smoothing+ trigram	6530.13

Perplexity: 6530.133257787546

فایل *mask*

predicted_word: Null

سارتر به عنوان روشن فکری فعال از نظر

predicted_word: Null

این تیم در سال ۱۸۹۹

predicted_word: Null

شد و تا به حال موفق به کسب یک عنوان

predicted_word: Null

بزرگترین کلیسای

predicted_word: Null

کریم خان پس از

predicted_word: Null

بر دشمنان خود

predicted_word: Null

عمومی ترین

predicted_word: Null

جانوران که طی مراحل

predicted_word: Null

نیز پیش از همه

فایل *incomplete*

predicted_word: Null

در جریان انقلاب مشروطه ابتدا به عنوان یکی از نیروهای محمدعلی شاه با

predicted_word: Null

شرکت خدمات مالی و

predicted_word: Null

شخص موسی تورات را

predicted_word: Null

نام آمازون را از یک لغت نامه

predicted_word: Null

تیم سپاهان اصفهان که در ابتدا شعبه

بخش سوم

کد این بخش در فایل NLP_HW3_1.ipynb میباشد.

پیش پردازش

پیش پردازش دادگان این بخش با بخش های قبلی متفاوت است.

در این بخش تمام کاراکتر هایی که جزو حروف فارسی نیستند با تابع findall از کتابخانه re حذف شده اند. همچنین با استفاده از تابع replace کاراکتر های u200c\ (نیم فاصله) و \n (خط جدید) با فاصله جاگذاری شده اند.

جملات متن برحسب علائم نگارشی مانند نقطه , علامت سوال و.. جدا شده اند و کلمات برحسب فاصله جدا شده اند.

علاوه بر این هنگام استخراج کلمات vocab کلماتی که کمتر از ۵ تکرار داشته اند از vocab حذف شده اند و مانند کلمات

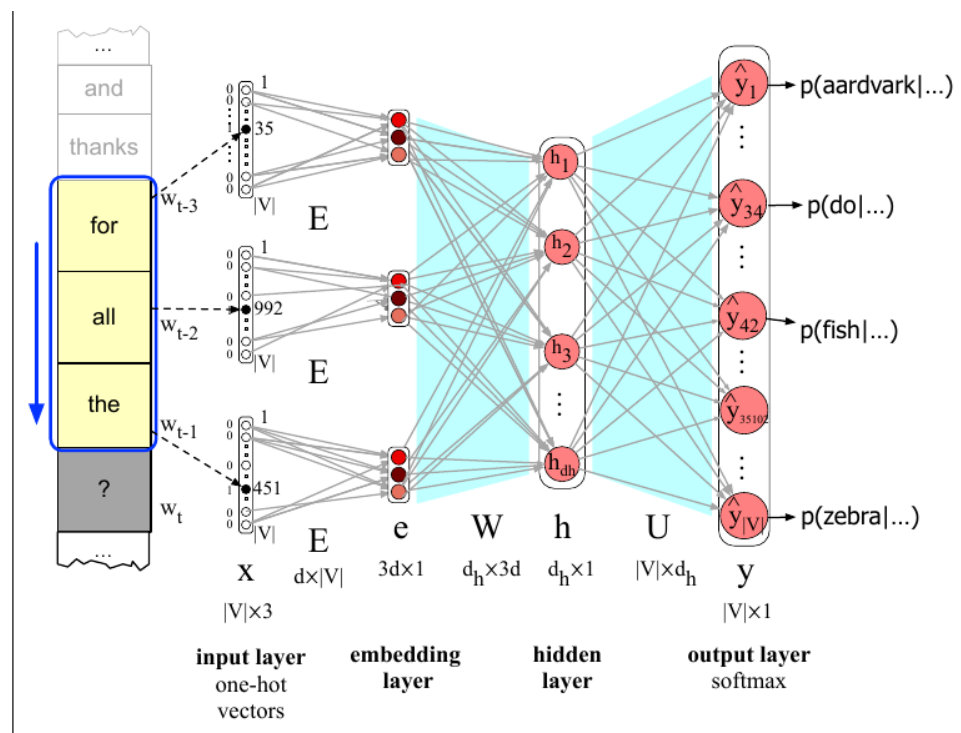
UNK¹

با آنها برخورد میشود. این موضوع باعث کاهش تعداد نوروں های لایه ورودی و خروجی شده و تاثیر خوبی در افزایش سرعت یادگیری مدل و سرعت کاهش loss داشته است. (بدون حذف کردن کلمات با فرکانس پایین میزان کاهش loss در طول آموزش بسیار کمتر بود.)

¹ unknown word

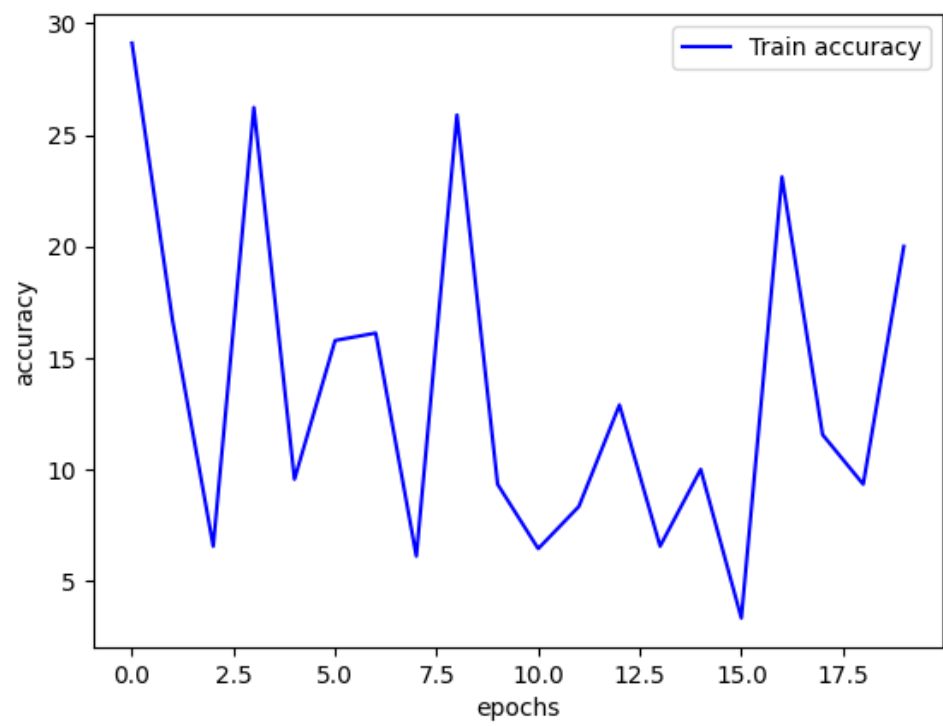
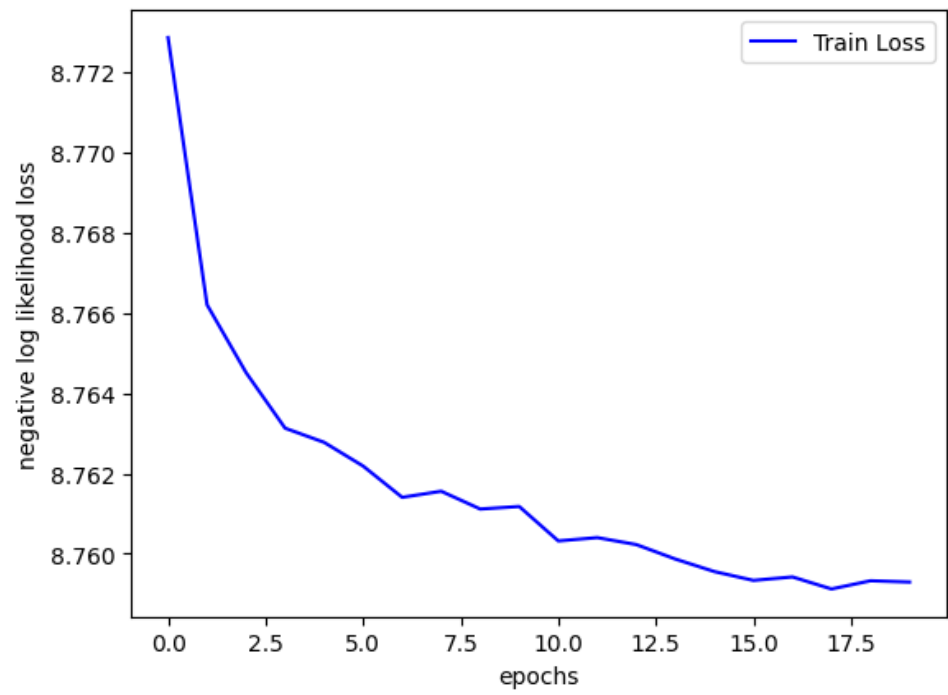
شبکه ی عصبی

بر اساس ساختار ارائه شده در مقاله ی Bengio² شبکه ی ایجاد شده شامل یک لایه embedding یک لایه hidden و یک لایه output و softmax می باشد. ورودی لایه embedding به اندازه ی طول vocab + ۱ می باشد (اضافه کردن یک به خاطر اضافه کردن توکن UNK به مجموعه کلمات می باشد). ساختار کلی مدل به شکل زیر می باشد که در آن d برابر 200 و d_h برابر 100 در نظر گرفته شده است.

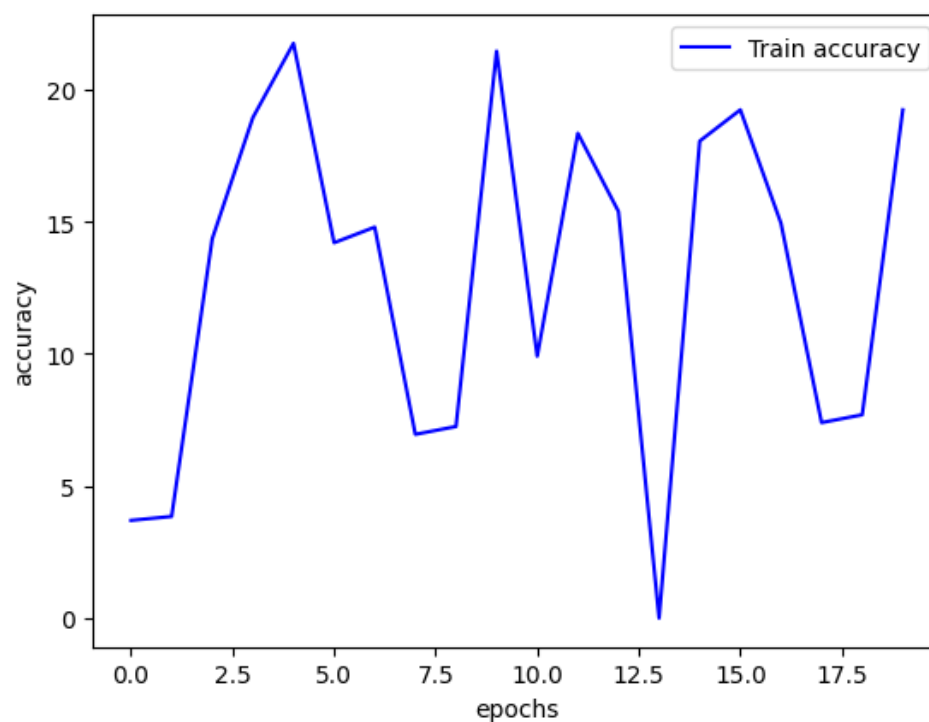
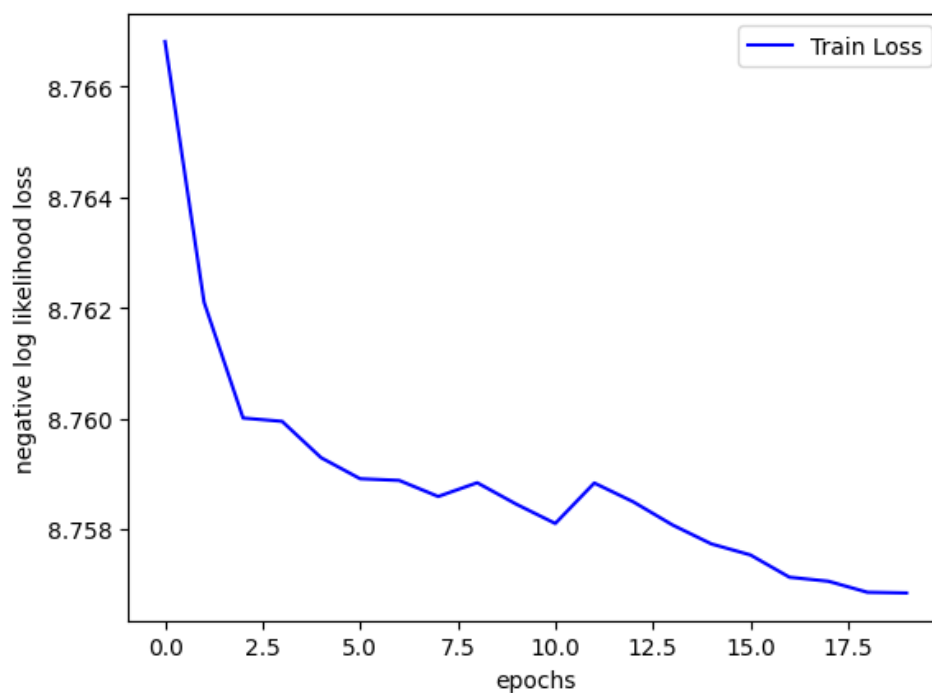


مدل trigram با ۲۰ اپیک آموزش دیده است. تغییرات loss و accuracy در طول آموزش به شرح زیر می باشد.

² <https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf>



مدل bigram با ۲۰ اپیاک آموزش دیده است. تغییرات loss و accuracy در طول آموزش به شرح زیر می باشد.



نتایج

نحوه ی محاسبه ی perplexity, براساس توضیحات این [مقاله](#) , این [لینک](#) میتوان با استفاده رابطه ی زیر اندازه ی perplexity برحسب cross entropy را به دست آورد.

$$\text{perplexity} = \exp(\text{cross-entropy loss})$$

	perplexity on validation data
trigram model	6530.5562
bigram model	8431.9

نتایج پیش بینی کلمات فایل *mask* و *incomplete*
 کد این بخش در فایل NLP_HW3_2.ipynb می باشد.

پیش بینی تکمیل فایل mask توسط مدل trigram

```
[ 'فعال', 'از', 'نظر' ]
<UNK>
[ 'در', 'سال', '۱۸۹۹' ]
<UNK>
[ 'کسب', 'یک', 'عنوان' ]
<UNK>
[ 'بزرگترین', 'کلیسای', '' ]
<UNK>
[ 'خان', 'پس', 'از' ]
<UNK>
[ 'بر', 'دشمنان', 'خود' ]
<UNK>
[ 'عمومی', 'ترین', '' ]
<UNK>
[ 'که', 'طی', 'مراحل' ]
<UNK>
[ 'پیش', 'از', 'همه' ]
<UNK>
```

```
[ 'محمدعلی', 'شاه', 'با' ]
<UNK>
[ 'خدمات', 'مالی', 'و' ]
<UNK>
[ 'موسی', 'تورات', 'را' ]
<UNK>
[ 'یک', 'لغت', 'نامه' ]
<UNK>
[ 'در', 'ابتدا', 'شعبه' ]
<UNK>
```

از آنجایی که نتایج perplexity مدل bigram پایین تر از مدل trigram می باشد نتایج این مدل هم مشابه trigram می باشد.

بررسی کلی نتایج

نتایج مدل های برحسب معیار perplexity به شرح زیر می باشد.

perplexity معیاری است که برای ارزیابی عملکرد مدل های زبانی استفاده می شود و اندازه گیری می کند که چگونه یک مدل زبان می تواند دنباله ای از کلمات را پیش بینی کند. perplexity کمتر نشان می دهد که مدل زبان در پیش بینی کلمه بعدی در یک دنباله بهتر است.

model	perplexity
uni_gram	1998
trigram+kneser-ney	6530.133
Feed forward language model for trigram	6530.55
Feed forward language model for bigram	8431
laplace +bigram	195797
trigram+laplace	383253

انتظار می رود یک مدل trigram از یک مدل unigram بهتر عمل کند. اما در اینجا برعکس رخ داده و مدل unigram کمترین perplexity را دارد. این می تواند در صورتی رخ دهد که مدل ترigram بیش از حد به داده های آموزشی تناسب داشته باشد و جرم احتمالی زیادی را به سه گرام های نادری که در داده های آموزشی دیده است اما به احتمال زیاد در داده های آزمون ظاهر نمی شوند، اختصاص دهد. این می تواند منجر به سردرگمی بیشتر در داده های آزمایشی برای مدل trigram در مقایسه با مدل unigram ساده شود.

مقایسه ی unigram و feedforward

به طور کلی انتظار می رود که یک مدل زبان بنژیو پیشخور بهتر از یک مدل یونیگرام ساده با هموارسازی لاپلاس عمل کند زیرا مدل بنژیو می تواند اطلاعات متنی کلمات قبلی را در دنباله ضبط کند، در حالی که مدل یونیگرام هر کلمه را به طور مستقل بررسی می کند. در این حالت احتمالاً به خاطر تعداد epoch های کم مدل پیشرو و ارزیابی مدل پیشرو مدل unigram عملکرد بهتری داشته است.

مقایسه ی مدل های trigram

هموارسازی kneser-ney و مدل شبکه ی عصبی میتوانند الگوها و وابستگی های پیچیده تری را نسبت به هموارسازی لاپلاس پیدا کنند و به نتایج بهتری میرسند.