

تمرین سری دو

درس PGM

فرهاد دلیرانی

۹۶۱۳۱۱۲۵

[dalirani@aut.ac.ir](mailto:dalirani@aut.ac.ir)

[dalirani.1373@gmail.com](mailto:dalirani.1373@gmail.com)

## ۱ فهرست

۱	۲ ابزارهای استفاده شده
۲	۳ سوال ۱- اندازه پارامتر توزیع های دیریکله متقارن
۲	۳.۱ دیتاست یک- تغییر آلفا
۲	۳,۱,۱ مدل یک
۶	۳,۱,۲ مدل دو
۱۰	۳,۱,۳ مدل سه
۱۴	۳,۲ دیتاست یک - تغییر بتا
۱۴	۳,۲,۱ مدل چهار
۱۷	۳,۲,۲ مدل پنج
۲۲	۳,۳ دیتاست دو- تغییر آلفا
۲۲	۳,۳,۱ مدل شش
۲۵	۳,۳,۲ مدل هفت
۲۹	۳.۴ دیتاست دو- تغییر بتا
۲۹	۳,۴,۱ مدل هشت
۳۳	۳,۵ نتیجه گیری
۳۳	۳,۵,۱ نتیجه گیری تغییر آلفا
۳۵	۳,۵,۲ نتیجه گیری تغییر بتا
۳۹	۴ سوال ۲- تعداد عنوان ها
۳۹	۴,۱ دیتاست یک - تغییر تعداد عنوان ها
۳۹	۴,۱,۱ مدل نه
۴۲	۴,۱,۲ مدل ده
۴۶	۴,۱,۳ مدل یازده
۵۱	۴,۱,۴ مدل دوازده
۵۵	۴,۱,۵ مدل سیزده
۵۹	۴,۱,۶ مدل چهارده
۶۳	۴.۲ دیتاست دو- تغییر تعداد عنوان ها

۶۳.....	۴,۲,۱	مدل پانزده
۶۷.....	۴,۲,۲	مدل شانزده
۷۱.....	۴,۲,۳	مدل هفده
۷۴.....	۴,۲,۴	مدل هجده
۷۸.....	۴,۲,۵	مدل نوزدهم
۸۲.....	۴,۲,۶	مدل بیستم
۸۶.....	۴,۲,۷	مدل بیست و یکم
۹۳.....	۴,۳	نتیجه‌گیری تغییر تعداد عنوان‌ها
۹۷.....	۵	سوال ۳-نمونه برداری
۹۷.....	۵,۱,۱	مدل بیست و دو
۱۰۱.....	۵,۱,۲	مدل بیست و سه
۱۰۴.....	۵,۱,۳	مدل بیست و چهار
۱۰۹.....	۵,۲	نتیجه گیری - روش‌های نمونه برداری
۱۱۱.....	۶	سوال ۴-خوشبندی اسناد
۱۱۱.....	۶,۱	مدل بیست و پنج
۱۱۵.....	۶,2	خوشبندی با استفاده از ترتیب مدل بیست و پنج
۱۱۷.....	۷	بخش‌های مختلف کد

## فهرست جدول‌ها

۲	جدول ۱ پارامترهای مدل یک
۳	جدول ۲ خروجی مدل یک
۶	جدول ۳ پارامترهای مدل دو
۷	جدول ۴ خروجی مدل دو
۱۰	جدول ۵ پارامترهای مدل سه
۱۱	جدول ۶ خروجی مدل سه
۱۴	جدول ۷ پارامترهای مدل چهار
۱۵	جدول ۸ پارامترهای مدل چهار
۱۸	جدول ۹ پارامترهای مدل پنج
۱۸	جدول ۱۰ خروجی مدل پنج
۲۲	جدول ۱۱ پارامترهای مدل شش
۲۳	جدول ۱۲ خروجی مدل شش
۲۵	جدول ۱۳ پارامترهای مدل هفت
۲۶	جدول ۱۴ خروجی مدل هفت
۲۹	جدول ۱۵ پارامترهای مدل هشت
۳۰	جدول ۱۶ خروجی مدل هشت
۳۴	جدول ۱۷ مقایسه‌ی تناهای مدل‌های دیتاست یک
۳۴	جدول ۱۸ مقایسه‌ی تناهای مدل‌های دیتاست دو
۳۶	جدول ۱۹ مقایسه‌ی فی‌های مدل‌های دیتاست یک
۳۷	جدول ۲۰ مقایسه‌ی فی‌های مدل‌های دیتاست دو
۳۹	جدول ۲۱ پارامترهای مدل نه
۴۰	جدول ۲۲ خروجی مدل نه
۴۲	جدول ۲۳ پارامترهای مدل ده
۴۳	جدول ۲۴ خروجی مدل ده
۴۶	جدول ۲۵ پارامترهای مدل یازده
۴۷	جدول ۲۶ خروجی مدل یازده
۵۱	جدول ۲۷ پارامترهای مدل دوازده

جداول ۲۸ خروجی مدل دوازده.....	۵۲.
جداول ۲۹ پارامترهای مدل سیزده.....	۵۵.
جداول ۳۰ خروجی مدل سیزده.....	۵۶.
جداول ۳۱ پارامترهای مدل چهارده.....	۵۹.
جداول ۳۲ خروجی مدل چهارده.....	۶۰.
جداول ۳۳ پارامترهای مدل پانزده .....	۶۳.
جداول ۳۴ خروجی مدل شانزده.....	۶۴.
جداول ۳۵ پارامترهای مدل شانزده .....	۶۷.
جداول ۳۶ خروجی مدل شانزده.....	۶۸.
جداول ۳۷ پارامترهای مدل هفدهم .....	۷۱.
جداول ۳۸ خروجی مدل هفدهم .....	۷۲.
جداول ۳۹ پارامترهای مدل هجدهم .....	۷۴.
جداول ۴۰ خروجی مدل هجدهم .....	۷۵.
جداول ۴۱ پارامترهای مدل نوزدهم .....	۷۸.
جداول ۴۲ خروجی مدل نوزدهم .....	۷۹.
جداول ۴۳ پارامترهای مدل بیست .....	۸۲.
جداول ۴۴ خروجی مدل بیست .....	۸۳.
جداول ۴۵ پارامترهای مدل بیست و یکم .....	۸۶.
جداول ۴۶ خروجی مدل بیست و یکم .....	۸۷.
جدول ۴۷ مقایسه زمان اجرا و پرپلیسکی مدل های دیتاست یک با تغییر تعداد عنوانها .....	۹۳.
جدول ۴۸ مقایسه زمان اجرا و پرپلیسکی مدل های دیتاست یک با تغییر تعداد عنوانها .....	۹۴.
جدول ۴۹ پارامترهای مدل بیست و دو .....	۹۷.
جدول ۵۰ خروجی مدل بیست و دو .....	۹۸.
جدول ۵۱ پارامترهای مدل بیست و سه .....	۱۰۱.
جدول ۵۲ خروجی مدل سیزده .....	۱۰۲.
جدول ۵۳ پارامترهای مدل بیست و چهار .....	۱۰۴.
جدول ۵۴ خروجی مدل بیست و چهار .....	۱۰۵.
جدول ۵۵ مقایسه پرپلیسکی در روش های مختلف نمونه برداری.....	۱۰۹.

۱۰۹.....	جدول ۵۶ مقایسه‌ی عنوان‌ها در روش‌های مختلف نمونه برداری
۱۱۱.....	جدول ۵۷ پارامترهای مدل بیست و پنج
۱۱۲.....	جدول ۵۸ خروجی مدل بیست و پنج
۱۱۶.....	جدول ۵۹ خروجی کلاسترینگ
۱۱۷.....	جدول ۶۰ بخش‌های مختلف کد

## 2 ابزارهای استفاده شده

زبان برنامه نویسی: زبان برنامه نویسی پایتون ۳.۶، (آنکوندا ۳)

محیط توسعه: Pycharm

سیستم عامل: Windows

### ۳ سوال ۱- اندازه پارامتر توزیع های دیریکله متقاض

برای بررسی تاثیر تغییر پارامترهای  $\alpha$  و  $\beta$ ، برای هر دو دیتاست مدل هایی با مقدارهای متفاوت آلفا و بتا آموزش می دهیم. سپس نتایج هر کدام از مدل ها را ارائه می دهیم و در ادامه نتیجه گیری می کنیم.

#### ۳.۱ دیتاست یک- تغییر آلفا

##### ۳.۱.۱ مدل یک

پارامترهای مختلف مدل:

جدول ۱ پارامترهای مدل یک

۱	$\alpha$
۱	$\beta$
۲۵	(تعداد کلمه ها در دیکشنری) $W$
۱۰	تعداد عنوان ها $T$
۱۰۰	حداکثر تعداد ایپاک <b>Max Epoch</b>
دیتاست یک	دیتاست

خروجی های حاصل از اجرای برنامه:

کلیهی خروجی های حاصل از اجرای کد برای مدل یک در پوشی **model-1-dataset1** موجود است، خروجی ها به شکل زیر اند:

:**Figure\_1.png** تصویر تعدادی از داکیومنت- تصویر ها

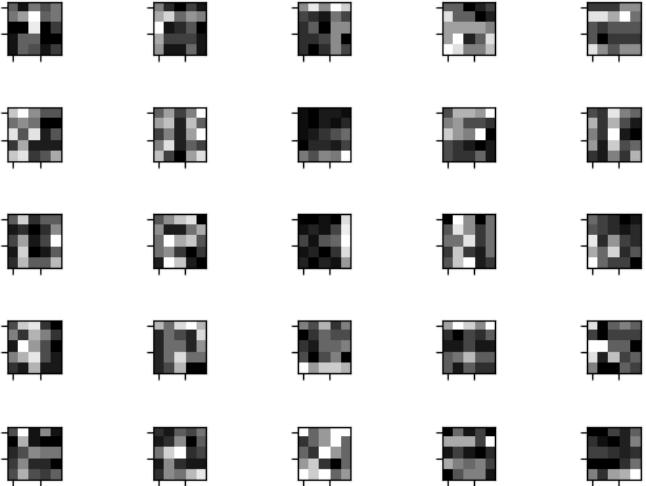
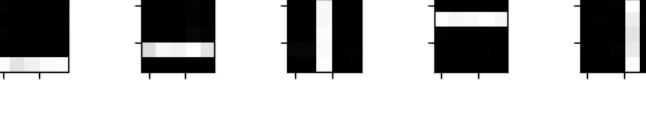
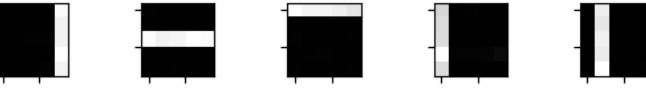
:**Figure\_2.png** تصویر عنوان های به دست آمده (Topics)

:**Figure\_3.png** تصویر Perplexity در حین آموزش

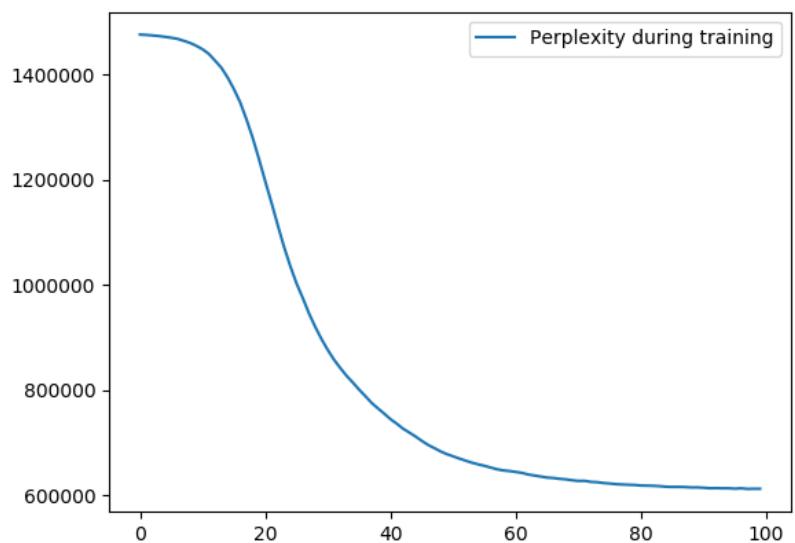
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، Perplexity تعداد ایپاک‌ها، زمان انجام کلیه ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۲ خروجی مدل یک

<p>some randomly selected samples</p> 	<p>تصویر تعدادی از داکیومنت‌ها</p>
 	<p>تصویر عنوان‌های به دست آمده Topics</p>

Perplexity  
در ایپاک‌های مختلف



زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۴۶۷,۰۳۳۲۵۸۴۳۸۱۱۰۳۵ s

زمان میانگین برای  
انجام یک ایپاک

۴,۶۷۰۰۳۳۲۵۸۴۳۸۱۱۰۴ s

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

98 epochs

زمان لازم برای رسیدن  
mixing به حالت

$98 * 4.6703 = 457.6894$  s

```

model-dataset1.json ✘ | 
1   "total_time": 467.03325843811035,
2   "each_epoch_time": 4.670332584381104,
3   "W": 25,
4   "T": 10,
5   "alpha": 1,
6   "beta": 1,
7   "dataset": 1,
8   "phi": [
9     [
10       [
11         0.0030324471848781967,
12         0.0003537855049024563,
13         0.00020216314565854644,
14         0.0002527039320731831,
15         0.0008086525826341858,
16         0.0007075710098049126,
17         0.0004043262913170929,
18         0.0015162235924390983,
19         0.0003032447184878197,
20         0.0009602749418780956,
21         0.0025775801071464674,
22         0.0005054078641463662,
23         0.0004043262913170929,
24         0.00015162235924390984,
25         0.0008591933690488224,
26         0.0011118973011220055,
27         0.0002527039320731831,
28         0.0003537855049024563,
29         0.0003032447184878197,
30         0.00015162235924390984,
31         0.20297179824118064,
32         0.18285656524815527,
33         0.19301526331749722,
34         0.20089962599818054,
35         0.20504397048418074
36       ],
37     ],
38   ]

```

فایل  
model-dataset1.json  
در پوششی  
Model-1-dataset-1  
که شامل ،  $\theta$ ،  $\Phi$  و  
سایر خروجی‌ها است

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف را به کاہش‌است و عنوان‌های به دست آمده مانند عنوان‌هایی است که نمونه‌ها با آن‌ها تولید شده‌اند.

### ۳.۱.۲ مدل دو

پارامترهای مختلف مدل:

جدول ۳ پارامترهای مدل دو

۰,۱	$\alpha$
۱	$\beta$
۲۵	W(تعداد کلمه‌ها در دیکشنری)
۱۰	T تعداد عنوان‌ها
۱۰۰	Max Epoch حداکثر تعداد ایپاک
دیتاست یک	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-2-dataset1** موجود است، خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت-تصویر ها **Figure\_1.png**

تصویر عنوان‌های به دست آمده (Topics) **Figure\_2.png**

تصویر Perplexity در حین آموزش **Figure\_3.png**

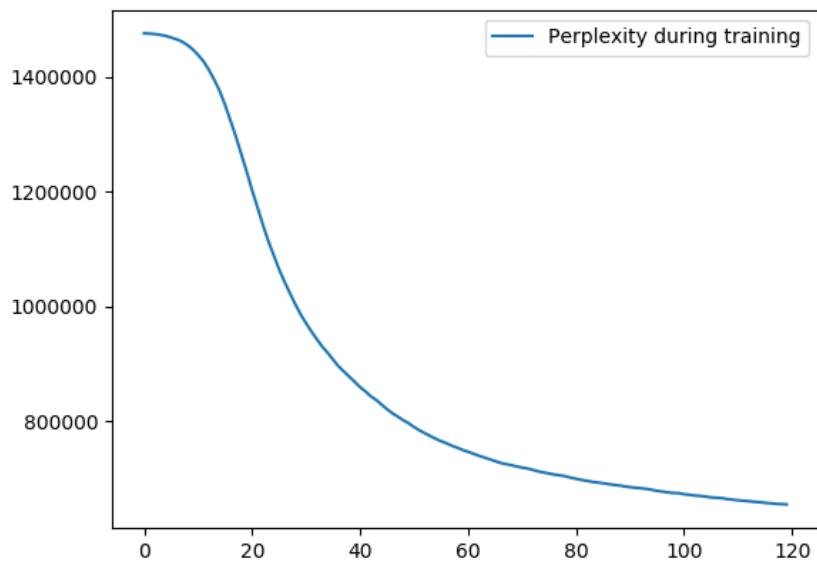
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۴: خروجی مدل دو

تصویر تعدادی از دکیومنتمها	some randomly selected samples					تصویر عنوانهای به دست آمده Topics
topics						

## Perplexity در ایپاک‌های مختلف



۵۳۹,۹۸۸۳۰۵۳۳۰۲۷۶۵ s	زمان کل اجرای حداکثر ایپاک‌های مجاز
۴,۴۹۹۹۰۲۵۴۴۴۱۸۹۷۱ s	زمان میانگین برای انجام یک ایپاک
۱۲۰	ایپاک‌های لازم برای رسیدن به حالت Mixing
۵۳۹,۹۸۸۳۰۵۳۳۰۲۷۶۵ s	زمان لازم برای رسیدن mixing به حالت

```

model-dataset1.json
1  {
2      "total_time": 539.9883053302765,
3      "each_epoch_time": 4.499902544418
4      "W": 25,
5      "T": 10,
6      "alpha": 0.1,
7      "beta": 1,
8      "dataset": 1,
9      "phi": [
10         [
11             0.0002490411914130597,
12             0.14817950889077053,
13             0.004931015589978583,
14             0.0017432883398914181,
15             0.006275838023609105,
16             0.004582357922000299,
17             0.14324849330079195,
18             0.002590028390695821,
19             0.0018927130547392538,
20             0.0003486576679782836,
21             0.006574687453304777,
22             0.15739403297305374,
23             0.004432933207152463,
24             0.006425262738456941,
25             0.012800717238631269,
26             0.004084275539174179,
27             0.15769288240274942,
28             0.003536384918065448,
29             0.009413757035413658,
30             0.006126413308761269,
31             0.030333217114110675,
32             0.19544752702096926,
33             0.02714548986402351,
34             0.026348558051501717,
35             0.03820291876276336
36         ],
37     ],
38 ]

```

فایل  
model-dataset1.json  
در پوششی  
Model-2-dataset-1  
که شامل  $\Phi$ ،  $\theta$  و  
سایر خروجی‌ها است

همین‌طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش‌است و عنوان‌های به دست آمده مانند عنوان‌هایی است که نمونه‌ها با آن‌ها تولید شده‌اند.

### ۳.۱.۳ مدل سه

پارامترهای مختلف مدل:

جدول ۵ پارامترهای مدل سه

جدول ۵ پارامترهای مدل سه	
۱۰	$\alpha$
۱	$\beta$
۲۵	(تعداد کلمه‌ها در دیکشنری) $W$
۱۰	تعداد عنوان‌ها $T$
۱۰۰	حداکثر تعداد ایپاک Max Epoch
دیتاست یک	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-3-dataset1** موجود است، خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت‌تصویر ها **Figure\_1.png**

تصویر عنوان‌های به دست آمده (Topics) **Figure\_2.png**

تصویر Perplexity در حین آموزش **Figure\_3.png**

یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان Perplexity در هر ایپاک است.

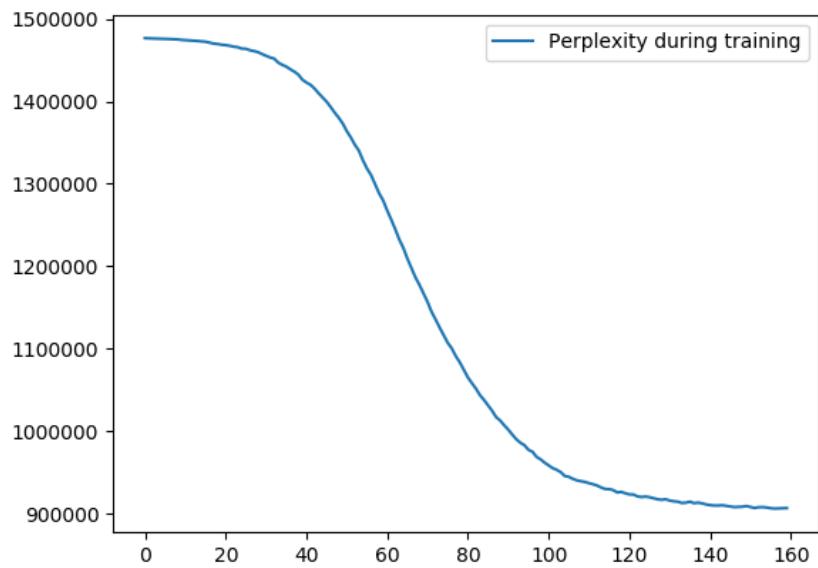
در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

## جدول ٦ خروجی مدل سه

تصویر تعدادی از دکیومنط‌ها	some randomly selected samples					تصویر عنوان‌های به دست آمده Topics

## Perplexity

در ایپاک‌های مختلف



۷۵۹,۳۲۲۰۴۷۴۷۲۰۰۰۱ s

زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۴,۷۴۵۷۶۲۷۹۶۷۰۰۰۱ s

زمان میانگین برای  
انجام یک ایپاک

۱۶۰

ایپاک‌های لازم برای  
رسیدن به حالت  
Mixing

۷۵۹,۳۲۲۰۴۷۴۷۲۰۰۰۱ s

زمان لازم برای رسیدن  
mixing به حالت

```

model-dataset1.json
1  {
2      "total_time": 759.3220474720001,
3      "each_epoch_time": 4.745762796700001,
4      "W": 25,
5      "T": 10,
6      "alpha": 10,
7      "beta": 1,
8      "dataset": 1,
9      "phi": [
10         [
11             0.00034129692832764505,
12             0.0005850804485616773,
13             0.0001462701121404193,
14             9.751340809361287e-05,
15             0.12028278888347148,
16             0.00034129692832764505,
17             0.0011701608971233545,
18             0.00034129692832764505,
19             0.0004388103364212579,
20             0.03437347635299854,
21             0.0001462701121404193,
22             0.0017064846416382253,
23             0.0002925402242808386,
24             0.0006338371526084836,
25             0.3917113603120429,
26             0.0012676743052169673,
27             0.00024378352023403217,
28             0.0003900536323744515,
29             0.0001462701121404193,
30             0.22754753778644563,
31             0.0005850804485616773,
32             0.0004388103364212579,
33             0.0002925402242808386,
34             0.0002925402242808386,
35             0.21618722574353974
36         ],
37         [
38             0.0002899251026818072,

```

فایل  
model-dataset1.json  
در پوششی  
Model-3-dataset-1  
که شامل  $\Phi$  و  $\Theta$  و  
سایر خروجی‌ها است

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف را به کاهش‌است و عنوان‌های به دست آمده مانند عنوان‌هایی است که نمونه‌ها با آن‌ها تولید شده‌اند.

## ۳.۲ دیتاست یک - تغییر بتا

### ۳.۲.۱ مدل چهار

پارامترهای مختلف مدل:

جدول ۷ پارامترهای مدل چهار

۱	$\alpha$
.۱	$\beta$
۲۵	W(تعداد کلمه‌ها در دیکشنری)
۱۰	T تعداد عنوان‌ها
۱۰۰	Max Epoch حداکثر تعداد ایپاک
دیتاست یک	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-4-dataset1** موجود است، خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت-تصویر ها :**Figure\_1.png**

تصویر عنوان‌های به دست آمده (Topics) :**Figure\_2.png**

تصویر Perplexity در حین آموزش :**Figure\_3.png**

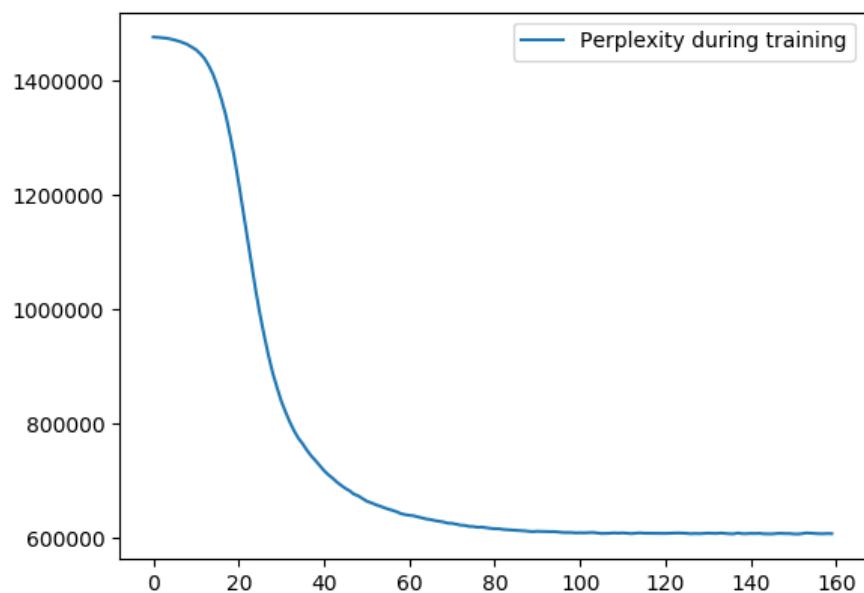
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۸ پارامترهای مدل چهار

تصویر تعدادی از داکیومنت‌ها	some randomly selected samples						تصویر عنوان‌های به دست آمده Topics
topics							

Perplexity  
در ایپاک‌های مختلف



۷۷۳,۱۸۴۱۴۷۳۵۷۹۴۰۷ s

زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۴,۸۳۲۴۰.۹۲۰.۹۸۷۱۲۹ s

زمان میانگین برای  
انجام یک ایپاک

120 epochs

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

$120 * 4.8324 = 579.888$  s

زمان لازم برای رسیدن  
به حالت mixing

فایل

model-dataset1.json

در پوششی

Model-4-dataset-1

که شامل ،  $\Phi$  ،  $\theta$  و

سایر خروجی‌ها است

```
1  {
2      "total_time": 773.1841473579407,
3      "each_epoch_time": 4.832400920987129,
4      "W": 25,
5      "T": 10,
6      "alpha": 1,
7      "beta": 0.1,
8      "dataset": 1,
9      "phi": [
10         [
11             0.20056161785353596,
12             0.19849249944577185,
13             0.20622706111289013,
14             0.20469985466906424,
15             0.18868881937088947,
16             4.9264723994383824e-06,
17             0.0002512500923713575,
18             0.00015272064438258986,
19             4.9264723994383824e-06,
20             4.9264723994383824e-06,
21             4.9264723994383824e-06,
22             4.9264723994383824e-06,
23             0.00010345592038820603,
24             4.9264723994383824e-06,
25             0.00010345592038820603,
26             5.419119639382221e-05,
27             0.00010345592038820603,
28             4.9264723994383824e-06,
29             0.00010345592038820603,
30             5.419119639382221e-05,
31             4.9264723994383824e-06,
32             0.00020198536837697366,
33             4.9264723994383824e-06,
34             4.9264723994383824e-06,
35             0.00015272064438258986
36         ],
37         [
38             0.19542630862517016,
```

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف را به کاهش‌است و عنوان‌های

به دست آمده مانند عنوان‌هایی است که نمونه‌ها با آن‌ها تولید شده‌اند.

## ۳.۲.۲ مدل پنج

پارامترهای مختلف مدل:

جدول ۹ پارامترهای مدل پنج

۱	$\alpha$
۱۰	$\beta$
۲۵	W(تعداد کلمه‌ها در دیکشنری)
۱۰	T تعداد عنوان‌ها
۱۰۰	Max Epoch حداکثر تعداد ایپاک
دیتاست یک	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوشه‌ی **model-5-dataset1** موجود است، خروجی‌ها به شکل زیر اند:

:تصویر تعدادی از داکیومنت-تصویر ها **Figure\_1.png**

:تصویر عنوان‌های به دست آمده (Topics) **Figure\_2.png**

:تصویر Perplexity در حین آموزش **Figure\_3.png**

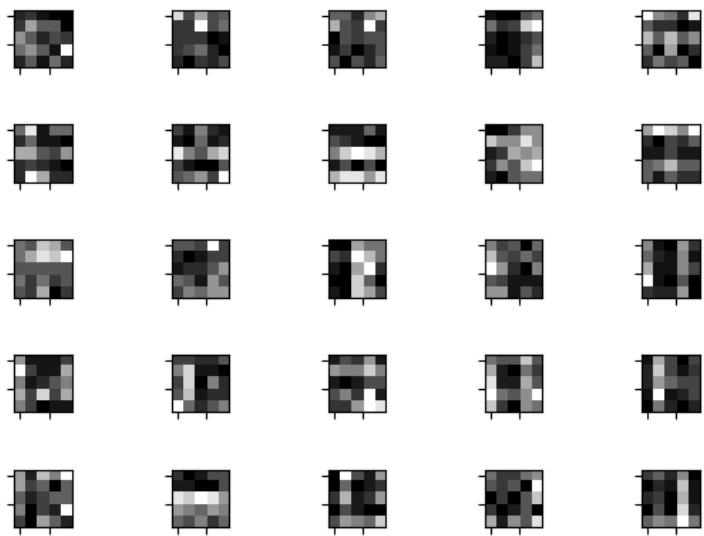
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++) پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\theta$  ،  $\Phi$  ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۱۰ خروجی مدل پنج

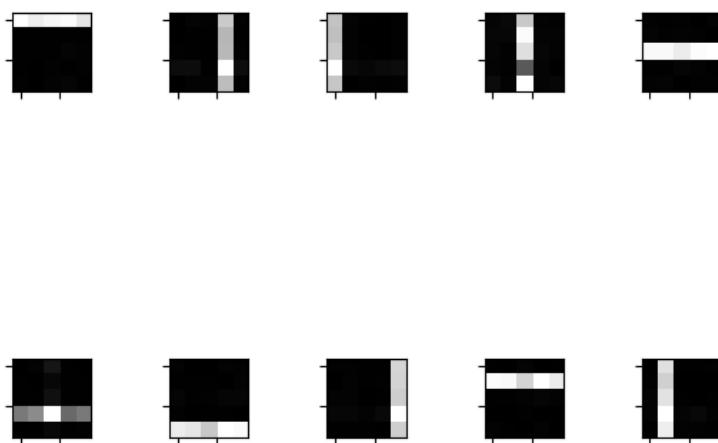
تصویر تعدادی از  
دکیومنټها

some randomly selected samples

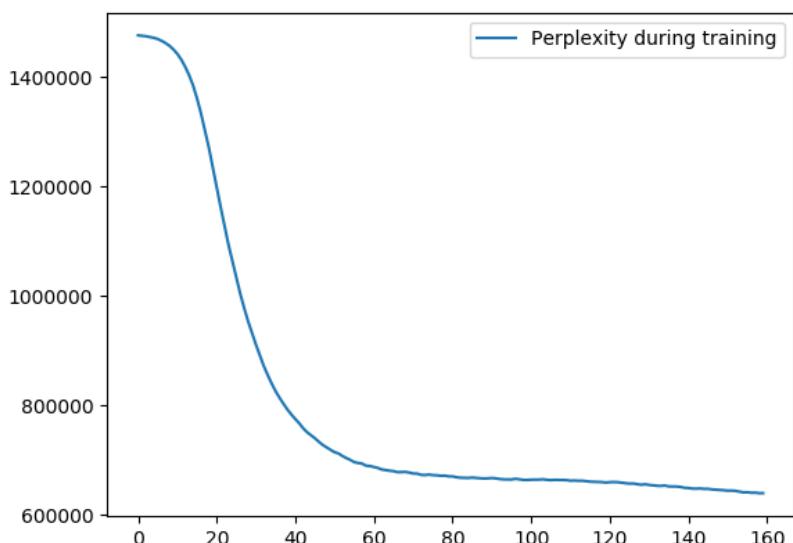


تصویر عنوان‌های به  
دست آمده  
Topics

topics



Perplexity  
در ایپاک‌های مختلف



۷۹۱,۷۷۴۲۴۰.۴۹۳۷۷۴۴ s

زمان کل اجرای حداکثر  
ایپاک‌های مجاز

4.94858900308609 s

زمان میانگین برای  
انجام یک ایپاک

160 epochs

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

۷۹۱,۷۷۴۲۴۰.۴۹۳۷۷۴۴ s

زمان لازم برای رسیدن  
mixing به حالت

```

model-dataset1.json
1   {
2     "total_time": 791.7742404937744,
3     "each_epoch_time": 4.94858900308609,
4     "W": 25,
5     "T": 10,
6     "alpha": 1,
7     "beta": 10,
8     "dataset": 1,
9     "phi": [
10       [
11         0.19034550355305072,
12         0.1785346728742955,
13         0.18392550845381034,
14         0.18701298701298702,
15         0.17196765498652292,
16         0.0039206076941926,
17         0.004165645675079637,
18         0.0039206076941926,
19         0.004018622886547415,
20         0.0037735849056603774,
21         0.004263660867434452,
22         0.003969615290370008,
23         0.00372457730948297,
24         0.006616025483950012,
25         0.004949767213918157,
26         0.00372457730948297,
27         0.0033815241362411172,
28         0.004606714040676305,
29         0.00411663807890223,
30         0.00470472923303112,
31         0.0044596912521440825,
32         0.0029894633668218575,
33         0.005537858368047047,
34         0.007302131830433717,
35         0.004067630482724822
36       ],
37     ],
38   ]

```

فایل  
model-dataset1.json  
در پوشه‌ی  
Model-5-dataset-1  
که شامل  $\Phi$  و  $\theta$  و  
سایر خروجی‌ها است

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف را به کاهش‌است و عنوان‌های به دست آمده مانند عنوان‌هایی است که نمونه‌ها با آن‌ها تولید شده‌اند.

## ۳.۳ دیتاست دو-تغییر آلفا

### ۳.۳.۱ مدل شش

پارامترهای مختلف مدل:

جدول ۱۱ پارامترهای مدل شش

۱	$\alpha$
۱	$\beta$
۱۰۴۷۳	W(تعداد کلمه‌ها در دیکشنری)
۲۰	T تعداد عنوان‌ها
۱۲۰	Max Epoch حداکثر تعداد ایپاک
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-6-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش :**Figure\_1.png**

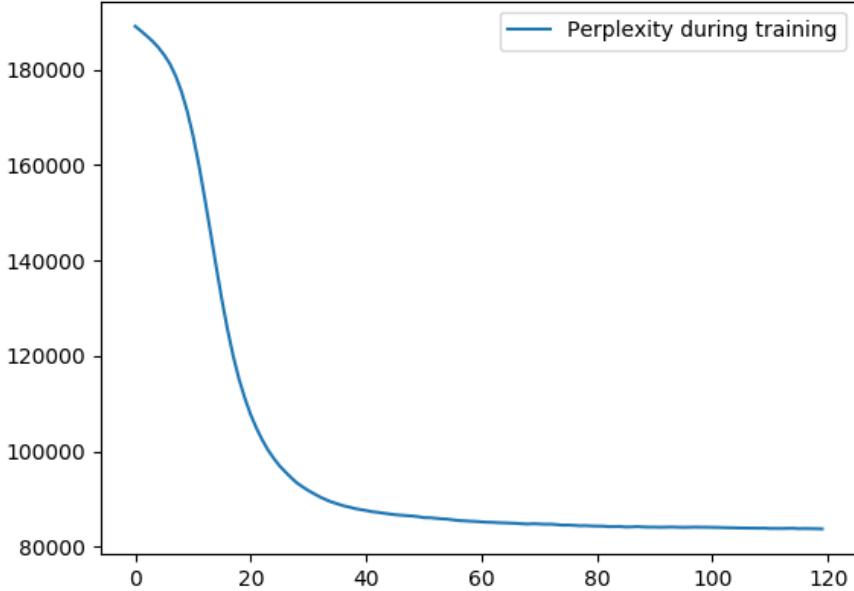
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$ ،  $\theta$ ، میزان Perplexity در هر ایپاک است.

در فایل **model-dataset2.json**، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل **topic-topics.txt**، کلمه‌ها رو از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را

در ابتدا قرار می‌دهیم، به این ترتیب عنوان‌ها در فایل `topics.txt` قابل مشاهده‌اند. (برای نمایش بهتر از یک خوب مانند `notepad++` استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۱۲ خروجی مدل شش

Perplexity در ایپاک‌های مختلف	
زمان کل اجرای حداقل ایپاک‌های مجاز	۲۶۲۱,۲۶۱۶۵۱۵۱۵۹۶۰۷ s
زمان میانگین برای انجام یک ایپاک	21.84384709596634 s
ایپاک‌های لازم برای	100 Epochs

رسیدم به  
حالت  
**Mixing**  
 $100 * 21.8438 = 457.6894 \text{ s}$

زمان لازم  
برای رسیدن  
به حالت  
mixing

فایل  
model-  
dataset1.json  
در پوششی  
Model-6-  
dataset-2  
که شامل ،  $\theta$  و سایر  $\Phi$   
خروجی‌ها  
است

```

1  {
2      "total_time": 2621.2616515159607,
3      "each_epoch_time": 21.84384709596634,
4      "W": 10473,
5      "T": 20,
6      "alpha": 1,
7      "beta": 1,
8      "dataset": 2,
9      "phi": [
10         [
11             7.01311452416018e-05,
12             7.01311452416018e-05,
13             7.01311452416018e-05,
14             0.00021039343572480537,
15             7.01311452416018e-05,
16             0.000350655726208009,
17             0.00021039343572480537,
18             7.01311452416018e-05,
19             7.01311452416018e-05,
20             0.00021039343572480537,
21             7.01311452416018e-05,
22             0.000350655726208009,
23             7.01311452416018e-05,
24             7.01311452416018e-05,
25             0.0001402622904832036,
26             7.01311452416018e-05,
27             0.0001402622904832036,
28             7.01311452416018e-05,
29             7.01311452416018e-05,
30             0.0001402622904832036,
31             7.01311452416018e-05,
32             7.01311452416018e-05,
33             7.01311452416018e-05,
34             0.000350655726208009,
35             7.01311452416018e-05,
36             0.00042078687144961075,
37             0.00021039343572480537,
38             0.00021039343572480537,
39             0.0001402622904832036,
40             7.01311452416018e-05,
41             0.0001402622904832036,
42             7.01311452416018e-05,
43             0.0001402622904832036,
44             0.0001402622904832036
        ]
    ]
}

```

فایل

topics.txt

در پوشه‌ی

Model-6-

dataset-2

که شامل

عنوان‌ها به

صورت کلمه

است

```

topics.txt
1 ['north', 'walsh', 'gesell', 'norts', 'irancontra', 'dec', 'aug', 'arms', 'poindexter', 'nov', 'intellig
2 ['percent', 'year', 'million', 'billion', 'last', 'report', 'sales', 'rate', 'new', 'years', 'increase',
3 ['i', 'years', 'people', 'like', 'dont', 'get', 'new', 'just', 'think', 'year', 'first', 'mrs', 'going',
4 ['dollar', 'cents', 'oil', 'late', 'yen', 'prices', 'lower', 'gold', 'cent', 'higher', 'market', 'trading
5 ['police', 'people', 'two', 'killed', 'city', 'man', 'three', 'death', 'shot', 'told', 'army', 'night', '
6 ['stock', 'market', 'index', 'points', 'exchange', 'million', 'trading', 'stocks', 'rose', 'shares', 'iss
7 ['party', 'government', 'political', 'soviet', 'people', 'national', 'president', 'opposition', 'communis
8 ['trade', 'farmers', 'japan', 'farm', 'states', 'japanese', 'agreement', 'agriculture', 'environmental',
9 ['court', 'case', 'attorney', 'judge', 'federal', 'trial', 'state', 'charges', 'law', 'prison', 'years',
10 ['soviet', 'united', 'states', 'south', 'president', 'foreign', 'war', 'union', 'talks', 'countries', 'wo
11 ['space', 'plant', 'water', 'shuttle', 'environmental', 'launch', 'nuclear', 'mission', 'test', 'nasa', '
12 ['fire', 'people', 'miles', 'area', 'two', 'officials', 'water', 'damage', 'city', 'state', 'homes', 'nor
13 ['school', 'students', 'university', 'news', 'student', 'schools', 'cbs', 'network', 'nbc', 'education',
14 ['air', 'force', 'military', 'iraq', 'kuwait', 'iraqi', 'plane', 'saudi', 'flight', 'navy', 'gulf', 'airc
15 ['company', 'million', 'new', 'corp', 'inc', 'workers', 'co', 'president', 'business', 'offer', 'employee
16 ['health', 'children', 'hospital', 'medical', 'aids', 'people', 'disease', 'drug', 'care', 'dr', 'heart',
17 ['bush', 'president', 'house', 'i', 'dukakis', 'campaign', 'new', 'bill', 'senate', 'committee', 'republi
18 ['new', 'southern', 'fair', 'northern', 'high', 'degrees', 'snow', 'city', 'temperatures', 'rain', 'inche
19 ['government', 'military', 'aid', 'united', 'rebels', 'states', 'president', 'panama', 'human', 'rights'],
20 ['west', 'east', 'german', 'germany', 'berlin', 'germania', 'germans', 'unification', 'city', 'manville']
21

```

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

### ۳.۳.۲ مدل هفت

پارامترهای مختلف مدل:

جدول ۱۳ پارامترهای مدل هفت

$\alpha$	۰,۱
$\beta$	۱
W (تعداد کلمه‌ها در دیکشنری)	۱۰۴۷۳
T (تعداد عنوان‌ها)	۲۰
Max Epoch (حداکثر تعداد ایپاک)	۱۲۰
دیتاست	دیتاست دو

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-7-dataset2** موجود است،  
خروجی‌ها به شکل زیر اند:

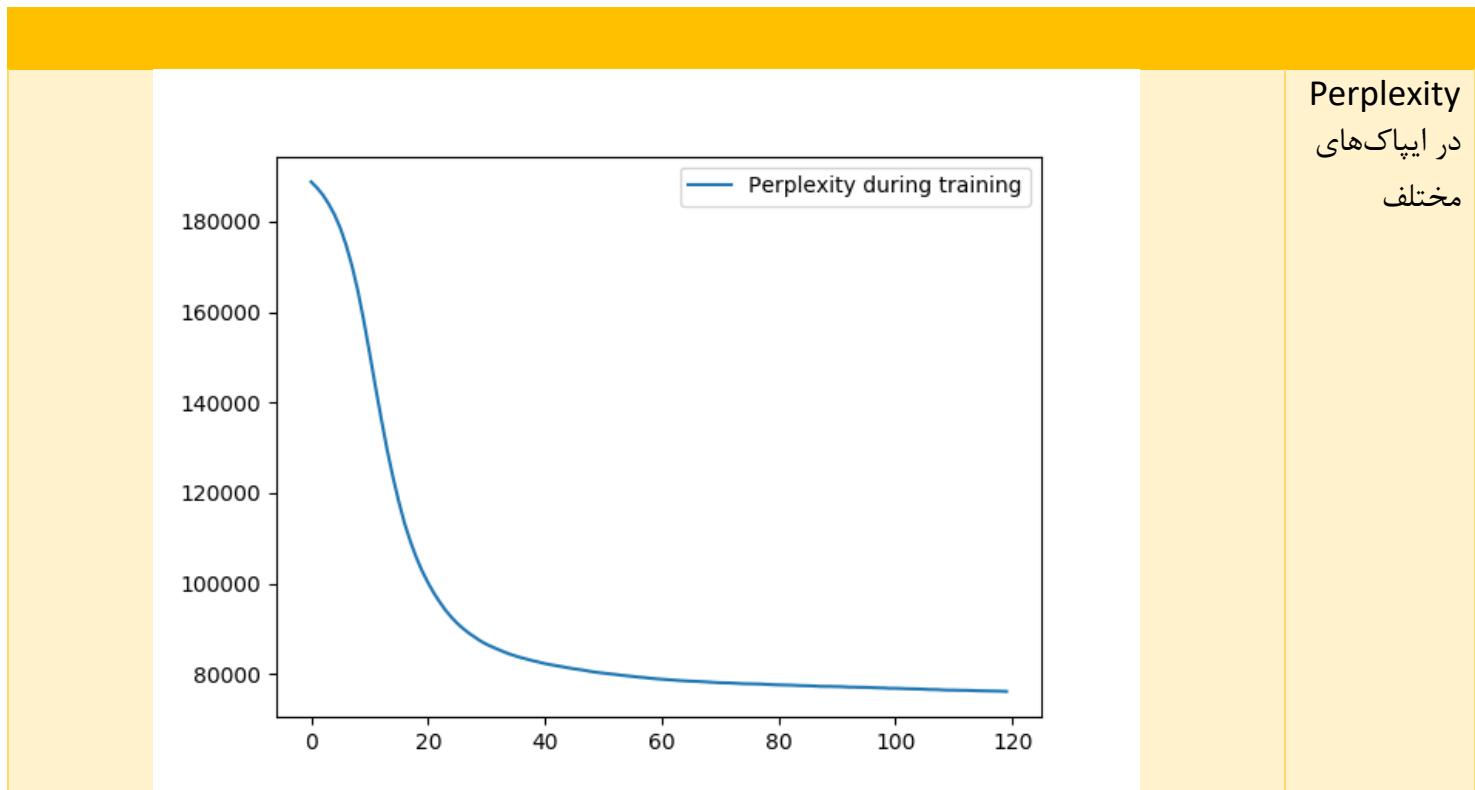
نمودار Perplexity در حین آموزش: **Figure\_1.png**

: یک فایل json که با انواع text-editor قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، **Perplexity** حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان در هر ایپاک است.

: در فایل **model-dataset2.json**، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل **topic- topics.txt**، کلمه‌ها را از دیکشنری بر می‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتدا قرار می‌دهیم. به این ترتیب عنوان‌ها در فایل **topics.txt** قابل مشاهده اند. (برای نمایش بهتر از یک **text-editor** خوب مانند **notepad++** استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۱: خروجی مدل هفت



۲۵۱۸,۹۸۱۸۰۳۴۱۷۲۰۶ s	زمان کل اجرای حداکثر ایپاکهای مجاز
20.991515028476716 s	زمان میانگین برای انجام یک ایپاک
120 Epochs	ایپاکهای لازم برای رسیدم به حالت Mixing
۲۵۱۸,۹۸۱۸۰۳۴۱۷۲۰۶ s	زمان لازم برای رسیدن به حالت mixing

```

model-dataset2.json
1   {
2     "total_time": 2518.981803417206,
3     "each_epoch_time": 20.991515028476716,
4     "W": 10473,
5     "T": 20,
6     "alpha": 0.1,
7     "beta": 1,
8     "dataset": 2,
9     "phi": [
10       [
11         0.003123295820732735,
12         0.001288979227603986,
13         7.436418620792226e-05,
14         0.0067175648207823116,
15         0.0006692776758713004,
16         0.007362054434584304,
17         0.0003470328689703039,
18         0.00014872837241584452,
19         0.0015616479103663676,
20         0.00019830449655445937,
21         0.0008427941103564523,
22         0.0007932179862178375,
23         0.019037231669228098,
24         0.0016608001586435972,
25         4.957612413861484e-05,
26         0.003643845124188191,
27         2.478806206930742e-05,
28         0.00039660899310891873,
29         2.478806206930742e-05,
30         0.0015368598482970602,
31         0.003916513806950572,
32         4.957612413861484e-05,
33         0.0002057409151752516,
34         0.000470973179316841,
35         0.0023796539586535123,
36         0.0006940657379406078,
37         0.002577958455207972,
38         0.00014872837241584452,
39         0.0010906747310495265,
40         0.0003718209310396113,
41         4.957612413861484e-05,

```

فایل  
model-dataset1.json  
در پوششی  
Model-7-dataset-2  
که شامل ،  $\theta$   
 $\Phi$  و سایر  
خروجی ها  
است

```

topics.txt
1  ["police", "pp", "two", "des", "people", "18", "killed", "96", "three", "style", "mant", "type", "officials", "shot", "night", "found", "authorities", "i", "fire",
2  ["dollar", "cents", "yen", "late", "lower", "cent", "higher", "gold", "futures", "bid", "prices", "london", "trading", "israel", "israeli", "jewish", "united", "palestinian", "peace", "arab", "minister", "states", "palestinians", "meeting", "company", "workers", "new", "corp", "president", "million", "union", "air", "contract", "inc", "employees", "co", "bus", "south", "iraq", "united", "kuwait", "africa", "iraqi", "states", "war", "president", "saudi", "gulf", "iran", "i", "bu", "children", "health", "aids", "hospital", "medical", "drug", "disease", "care", "patients", "doctors", "heart", "report", "i", "years", "people", "new", "year", "mrs", "dont", "just", "think", "like", "first", "says", "im", "women", "get", "military", "government", "troops", "army", "force", "officials", "air", "united", "rebels", "forces", "soldiers", "nav", "bush", "dukakis", "president", "campaign", "i", "new", "jackson", "democratic", "presidential", "republican", "vice", "government", "party", "people", "political", "police", "national", "president", "opposition", "minister", "leader", "blockbuster", "saw", "chain", "chinese", "cano", "side", "hu", "creque", "avoid", "video", "tree", "erols", "y", "n", "environmental", "water", "plant", "state", "years", "new", "time", "two", "people", "department", "waste", "space", "s", "fire", "southern", "miles", "northern", "rain", "people", "coast", "fair", "central", "weather", "inches", "water", "r", "stock", "market", "index", "trading", "million", "exchange", "oil", "new", "york", "points", "shares", "stocks", "price", "soviet", "gorbachev", "union", "president", "united", "states", "trade", "moscow", "world", "summit", "germany", "soviet", "west", "east", "german", "germany", "germans", "germans", "berlin", "network", "unification", "spain", "radio", "robe", "court", "case", "attorney", "judge", "trial", "federal", "i", "state", "charges", "prison", "years", "two", "law", "dr", "percent", "million", "year", "billion", "last", "new", "sales", "rate", "report", "prices", "government", "years", "ir", "house", "senate", "bill", "committee", "i", "congress", "bush", "budget", "rep", "sen", "president", "members", "vote", "i", "school", "city", "years", "new", "people", "first", "two", "like", "ago", "john", "town", "get", "york", "just", "21"]

```

فایل  
topics.txt  
در پوششی  
Model-6-dataset-2  
که شامل  
عنوان ها به  
صورت کلمه  
است

همین طور که در خروجی های بالا دیده می شود، perplexity در ایپاک های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان ها، به هم مرتبط اند.

## ۳.۴ دیتاست دو-تغییر بتا

### ۳.۴.۱ مدل هشت

با تغییر  $\beta$  مدل شش، مدل هشت را ایجاد می‌کنیم که فقط در بتا با مدل شش تفاوت دارد.

پارامترهای مختلف مدل:

جدول ۱۵ پارامترهای مدل هشت

۱	$\alpha$
۱۰	$\beta$
۱۰۴۷۳	W(تعداد کلمه‌ها در دیکشنری)
۲۰	T تعداد عنوان‌ها
۱۲۰	Max Epoch حداقل تعداد ایپاک
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-8-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش :**Figure\_1.png**

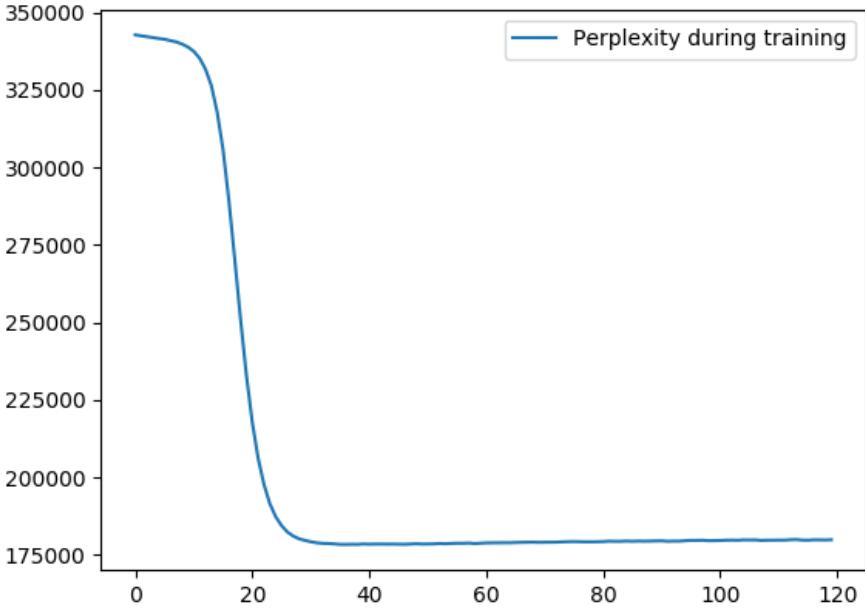
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، حداقل تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\theta$ ،  $\Phi$ ، میزان Perplexity در هر ایپاک است.

در فایل **model-dataset2.json**، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل **topic-topics.txt**، کلمه‌ها رو از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را

در ابتدا قرار می‌دهیم، به این ترتیب عنوان‌ها در فایل `topics.txt` قابل مشاهده‌اند. (برای نمایش بهتر از یک خوب مانند `notepad++` استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

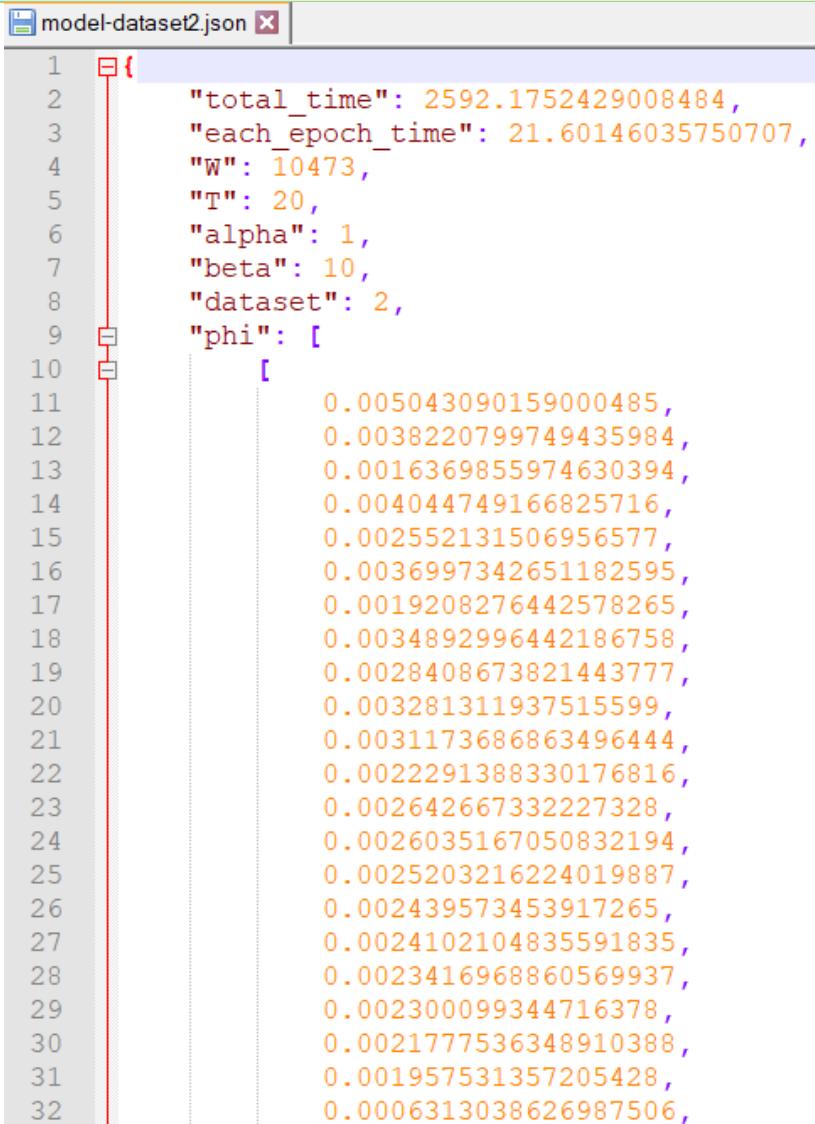
جدول ۱۷ خروجی مدل هشت

Perplexity در ایپاک‌های مختلف	
زمان کل اجرای حداکثر ایپاک‌های مجاز	۲۵۹۲,۱۷۵۲۴۲۹۰۰۸۴۸۴ s
زمان میانگین برای انجام یک ایپاک	21.60146035750707 s
ایپاک‌های لازم برای	37 Epochs

رسیدم به  
حالت  
Mixing

$$37 * 21.6014 = 799.2518 \text{ s}$$

زمان لازم  
برای رسیدن  
به حالت  
mixing



```

1  {
2      "total_time": 2592.1752429008484,
3      "each_epoch_time": 21.60146035750707,
4      "W": 10473,
5      "T": 20,
6      "alpha": 1,
7      "beta": 10,
8      "dataset": 2,
9      "phi": [
10          [
11              0.005043090159000485,
12              0.0038220799749435984,
13              0.0016369855974630394,
14              0.004044749166825716,
15              0.002552131506956577,
16              0.0036997342651182595,
17              0.0019208276442578265,
18              0.0034892996442186758,
19              0.0028408673821443777,
20              0.003281311937515599,
21              0.0031173686863496444,
22              0.0022291388330176816,
23              0.002642667332227328,
24              0.0026035167050832194,
25              0.0025203216224019887,
26              0.002439573453917265,
27              0.0024102104835591835,
28              0.0023416968860569937,
29              0.002300099344716378,
30              0.0021777536348910388,
31              0.001957531357205428,
32              0.0006313038626987506,

```

فایل  
model-  
dataset1.json  
در پوششی  
Model-8-  
dataset-2  
که شامل ،  $\theta$   
و سایر  $\Phi$   
خروجی‌ها  
است

فایل

topics.txt

در پوششی

Model-8-  
dataset-2

که شامل

عنوان‌ها به

صورت کلمه

است

```
topics.txt
1 ['i', 'people', 'new', 'two', 'president', 'government', 'years', 'last', 'police', 'state', 'year', 'states', 'of
2 ['temperatures', 'stage', 'four', 'white', 'turkish', 'selling', 'roman', 'performance', 'opinion', 'lawmakers', '!
3 ['shot', 'wagner', 'trying', 'tell', 'suits', 'south', 'record', 'read', 'opera', 'members', 'magazine', 'just', '!
4 ['attend', 'tigers', 'shuttle', 'july', 'art', 'wrote', 'warren', 'ticket', 'telecharge', 'school', 'purchase', 'p
5 ['wednesday', 'museum', 'inc', 'temperatures', 'roy', 'police', 'old', 'nj', 'musicians', 'make', 'house', 'discrini
6 ['small', 'signals', 'seized', 'plates', 'million', 'makes', 'interview', 'influence', 'hearing', 'featuring', 'de
7 ['scheduled', 'veto', 'terror', 'signed', 'show', 'san', 'reservation', 'range', 'organization', 'museum', 'march'
8 ['miles', 'martin', 'authority', 'wrong', 'travels', 'store', 'station', 'pilot', 'paying', 'modern', 'just', 'hel
9 ['players', 'records', 'numbers', 'actor', 'winning', 'venus', 'shes', 'researchers', 'represents', 'poland', 'peri
10 ['band', 'world', 'tuesday', 'texas', 'television', 'species', 'remains', 'real', 'placed', 'listed', 'highs', 'gr
11 ['report', 'improved', 'turned', 'television', 'spanish', 'socalled', 'proposal', 'planets', 'picture', 'mountain'
12 ['percent', 'million', 'billion', 'market', 'year', 'new', 'prices', 'stock', 'company', 'dollar', 'rose', 'oil',
13 ['investigators', 'edward', 'movie', 'store', 'stage', 'raised', 'natural', 'journal', 'institute', 'grant', 'gott
14 ['sold', 'memorial', 'couple', 'weeks', 'wednesday', 'school', 'route', 'peter', 'opera', 'mothers', 'male', 'lloy
15 ['financial', 'venus', 'deep', 'chief', 'turkey', 'sports', 'spokesman', 'spacecraft', 'pledged', 'performed', 'pa
16 ['san', 'roberts', 'prosecutor', 'electric', 'yugoslavia', 'years', 'stay', 'recently', 'paper', 'original', 'orde
17 ['m', 'cdy', 'clr', 'play', 'death', 'thought', 'tested', 'states', 'sent', 'really', 'noriega', 'music', 'missing
18 ['yards', 'fourth', 'sweeping', 'specialist', 'seat', 'r', 'petition', 'option', 'network', 'insurance', 'hudson',
19 ['care', 'bcspehealth', 'b', 'items', 'denied', 'computer', 'summer', 'slightly', 'record', 'reason', 'range', 'mi
20 ['wine', 'washington', 'twin', 'street', 'stations', 'rate', 'pm', 'offered', 'neck', 'music', 'jury', 'immigratio
```

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

## ۳.۵ نتیجه‌گیری

### ۳.۵.۱ نتیجه‌گیری تغییر آلفا

در بخش‌های ۱-۳ و ۳-۳، مدل‌های ۱، ۲ و ۳ را برای دیتاست شماره‌ی یک و مدل‌های ۶ و ۷ را برای دیتاست دو، به ازای  $\alpha$ ‌های مختلف ایجاد کردیم، که پارامترهای هر کدام از مدل‌ها برای آموزش در جدول‌های ۱۱، ۱۳، ۵ و ۱۳ قابل مشاهده است. خروجی‌های هر کدام از آن‌ها در جدول‌های ۴، ۶، ۲، ۱۲ و ۱۴ قرار دارد.

آلفا پارامتری است که تتا را تنظیم می‌کند و بر روی آن تاثیر دارد:

$$\theta \sim Dirichlet(\alpha)$$

تتا اهمیت هر عنوان در یک سند را نشان می‌دهد.

با بررسی مدل‌های ۱، ۲ و ۳ که بر روی دیتاست یک آموزش داده شده‌اند و دقت در تтай آن‌ها که بخشی از آن‌ها در جدول ۱۷ آورده شده‌اند، متوجه می‌شویم هر چه آلفا زیاده شده است، عنوان‌های بیشتری تمایل دارند در سند حضور داشته باشند و از آنجایی که بیشتر عنوان‌ها متمایل در شرکت در سند هستند، احتمال حضور هر کدام عدد کوچکی و نزدیک به هم است، در آلفا برابر با ۰،۱۰، احتمال حضور همه‌ی عنوان‌ها تقریباً برابر با ۰،۱ است. در حالی که هر چه آلفا کم شده است، تعداد کمتری عنوان تمایل در شرکت در سند را دارند که آن عنوان‌ها هر کدام از احتمال بالایی برای حضور در سند برخورداراند و سایر عنوان‌ها از احتمال پایینی نسبت به آن‌ها برای حضور در سند برخوردارند، به طور مثال در آلفا برابر با ۰،۱، همانطور که با رنگ آبی مشخص شده است، یکی از عنوان‌ها احتمال حضور ۴۵ درصد در متن را دارد در حالی که بیشتر عنوان‌های دیگر احتمالی نزدیک به یک درصد دارند.

به طور خلاصه هر چه آلفا بیشتر می‌شود، یک سند تمایل دارد از ترکیب عنوان‌های بیشتری تشکیل شود.

جدول ۱۷ مقایسه‌ی تناهای مدل‌های دیتاست یک

$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$
<pre>"theta": [   [     0.0009900990099009901,     0.1594059405940594,     0.24851485148514854,     0.060396039603960394,     0.0009900990099009901,     0.010891089108910892,     0.0009900990099009901,     0.32772277227722774,     0.1891089108910891,     0.0009900990099009901   ],   [     0.0801980198019802,     0.0009900990099009901,     0.0009900990099009901,     0.020792079207920793,     0.0900990099009901,     0.1198019801980198,     0.1198019801980198,     0.4564356435645,     0.0900990099009901,     0.020792079207920793   ],   [     0.1495049504950495,     0.1099009900990099   ] ]</pre>	<pre>"theta": [   [     0.00909090909090909,     0.12727272727272726,     0.1,     0.17272727272727273,     0.218181818181817,     0.02727272727272727,     0.03636363636363636,     0.2545454545454545,     0.045454545454545456,     0.00909090909090909   ],   [     0.07272727272727272,     0.018181818181818,     0.118181818181818,     0.1,     0.081818181818182,     0.13636363636363635,     0.07272727272727272,     0.2727272727272727,     0.018181818181818,     0.10909090909090909   ],   [     0.081818181818182,     0.218181818181817,     0.17272727272727273   ] ]</pre>	<pre>"theta": [   [     0.065,     0.065,     0.095,     0.135,     0.12,     0.13,     0.06,     0.105,     0.085,     0.14   ],   [     0.11,     0.1,     0.095,     0.145,     0.11,     0.1,     0.095,     0.055,     0.08,     0.11   ],   [     0.09,     0.13,     0.09   ] ]</pre>

با بررسی مدل‌های ۷ و ۶ که بر روی دیتاست دو آموزش داده شده‌اند که پارامترهای آموزش و خروجی‌های آن‌ها در جدول‌های ۱۳، ۱۴، ۱۵ و ۱۶ قرار دارد و با دقت در ترتیب آن‌ها که بخشی از آن‌ها در جدول ۱۸ آورده شده‌اند، به همان نتیجه‌ی حاصل از تغییر آلفا در مدل‌های مربوط به دیتاست یک می‌رسیم که در بالا آن را بیان کردیم.

جدول ۱۸ مقایسه‌ی تناهای مدل‌های دیتاست دو

$\alpha = 0.1$	$\alpha = 1$
----------------	--------------

### ٣.٥.٢ نتیجه‌گیری تغییر بتا

در بخش‌های ۲-۳ و ۴-۵، مدل‌های ۱، ۴ و ۵ را برای دیتاست شماره‌ی یک و مدل‌های ۶ و ۸ را برای دیتاست دو، به ازای  $\beta$ ‌های مختلف ایجاد کردیم، که پارامترهای هر کدام از مدل‌ها برای آموزش در جدول‌های ۱، ۷، ۹، ۱۰، ۱۲، ۱۶ و ۱۵ قابل مشاهده است. خروجی‌های هر کدام در از آن مدل‌ها در جدول‌های ۲، ۸، ۱۰، ۱۲، ۱۶ قرار دارد.

بایارمتری است که  $\phi$  را تنظیم می‌کند و بروی آن تاثیر دارد:

$$\phi^{(j)} \sim Dirichlet(\beta)$$

$\phi$  اهمیت هر کلمه در عنوان را نشان می‌دهد.

با بررسی مدل‌های ۱، ۴، ۵ که بر روی دیتاست یک آموزش داده شده‌اند و دقت در  $\phi$  آن‌ها که بخشی از آن‌ها در جدول ۱۹ آورده شده‌اند، متوجه می‌شویم هر چه بتا کمتر شده است، کلمات بیشتری از یک عنوان مقدار بسیار کوچکی پیدا کرده‌اند و تعدادی کمی از کلمات احتمال زیادی دارند، به عبارت دیگر عنوان‌ها بیشتر متمایل هستند که از ترکیب کلمات کمتری تشکیل شده باشند. هر چه بتا بیشتر می‌شود احتمال کلمات مختلف به هم نزدیک می‌شود، به عبارتی هر عنوان متمایل است کلمات بیشتری در خود جا دهد و از کلمات مشخصی تشکیل نشده باشد.

به طور خلاصه هر چه بتا بیشتر می‌شود، احتمال اینکه یک عنوان از ترکیب کلمه‌های بیشتر تشکیل شده باشد بیشتر می‌شود.

جدول ۱۹ مقایسه‌ی فی‌های مدل‌های دیتاست یک

$\beta = 0.1$

$\beta = 1$

$\beta = 10$

dataset ۱	dataset ۲	dataset ۳
<pre> "phi": [   [     0.20056161785353596,     0.19849249944577185,     0.20622706111289013,     0.20469985466906424,     0.18868881937088947,     4.9264723994383824e-06,     0.0002512500923713575,     0.00015272064438258986,     4.9264723994383824e-06,     4.9264723994383824e-06,     4.9264723994383824e-06,     4.9264723994383824e-06,     0.00010345592038820603,     4.9264723994383824e-06,     0.00010345592038820603,     5.419119639382221e-05,     0.00010345592038820603,     4.9264723994383824e-06,     0.00010345592038820603,     5.419119639382221e-05,     4.9264723994383824e-06,     0.00020198536837697366,     4.9264723994383824e-06,     4.9264723994383824e-06,     0.00015272064438258986   ],   [     0.19542630862517016,     4.949882440292043e-06,     4.949882440292043e-06   ] ] </pre>	<pre> "phi": [   [     0.0030324471848781967,     0.0003537855049024563,     0.00020216314565854644,     0.0002527039320731831,     0.0008086525826341858,     0.0007075710098049126,     0.0004043262913170929,     0.0015162235924390983,     0.0003032447184878197,     0.0009602749418780956,     0.0025775801071464674,     0.0005054078641463662,     0.0004043262913170929,     0.00015162235924390984,     0.0008591933690488224,     0.0011118973011220055,     0.0002527039320731831,     0.0003537855049024563,     0.0003032447184878197,     0.00015162235924390984,     0.20297179824118064,     0.18285656524815527,     0.19301526331749722,     0.20089962599818054,     0.20504397048418074   ],   [     9.949753743594846e-05,     0.0016417093676931497   ] ] </pre>	<pre> "phi": [   [     0.19034550355305072,     0.1785346728742955,     0.18392550845381034,     0.18701298701298702,     0.17196765498652292,     0.0039206076941926,     0.004165645675079637,     0.0039206076941926,     0.004018622886547415,     0.0037735849056603774,     0.004263660867434452,     0.003969615290370008,     0.00372457730948297,     0.006616025483950012,     0.004949767213918157,     0.00372457730948297,     0.0033815241362411172,     0.004606714040676305,     0.00411663807890223,     0.00470472923303112,     0.0044596912521440825,     0.0029894633668218575,     0.005537858368047047,     0.007302131830433717,     0.004067630482724822   ],   [     0.003109729009329187,     0.006367540352435954,     0.004689273902956711   ] ] </pre>

با بررسی مدل‌های ۸ و ۶ که بر روی دیتاست دو آموزش داده شده‌اند و فقط در بتا با یک دیگر متفاوت‌اند و با دقت در فی آن‌ها که بخشی از آن‌ها در جدول ۲۰ آورده شده‌اند، به همان نتیجه‌ی حاصل از تغییر بتا در مدل‌های مربوط به دیتاست یک می‌رسیم که در بالا آن را بیان کردیم.

جدول ۲۰ مقایسه‌ی فی‌های مدل‌های دیتاست دو

$$\beta = 1$$

$$\beta = 10$$

```
"phi": [  
  7.01311452416018e-05,  
  7.01311452416018e-05,  
  7.01311452416018e-05,  
  0.00021039343572480537,  
  7.01311452416018e-05,  
  0.000350655726208009,  
  0.00021039343572480537,  
  7.01311452416018e-05,  
  7.01311452416018e-05,  
  0.00021039343572480537,  
  7.01311452416018e-05,  
  0.000350655726208009,  
  7.01311452416018e-05,  
  7.01311452416018e-05,  
  0.0001402622904832036,  
  7.01311452416018e-05,  
  0.0001402622904832036,  
  7.01311452416018e-05
```

```
"phi": [  
  0.005043090159000485,  
  0.0038220799749435984,  
  0.0016369855974630394,  
  0.004044749166825716,  
  0.002552131506956577,  
  0.0036997342651182595,  
  0.0019208276442578265,  
  0.0034892996442186758,  
  0.0028408673821443777,  
  0.003281311937515599,  
  0.0031173686863496444,  
  0.0022291388330176816,  
  0.002642667332227328,  
  0.0026035167050832194,  
  0.0025203216224019887,  
  0.002439573453917265,  
  0.0024102104835591835,  
  0.0023416968860569937,  
  0.002300099344716378,
```

## ۴ سوال ۲-تعداد عنوان‌ها

برای بررسی اثر تعداد عنوان‌ها، مدل‌های مختلفی را بر روی دیتاست یک و دو آموزش می‌دهیم. مدل‌های آموزش داده شده با دیتاست یک همه‌ی پارامترهایشان یکسان است و فقط در تعداد عنوان متفاوت‌اند، مدل‌هایی هم که در این بخش بر روی مجموعه‌ی دو آموزش می‌دهیم هم فقط در تعداد عنوان‌ها متفاوت‌اند و در سایر پارامترها یکسان‌اند.

### ۴.۱ دیتاست یک – تغییر تعداد عنوان‌ها

در این قسمت مدل‌های مختلفی را با دیتاست یک ایجاد می‌کنیم که فقط در تعداد عنوان با یک دیگر متفاوت‌اند.

#### ۴.۱.۱ مدل نه

پارامترهای مختلف مدل:

جدول ۲۱ پارامترهای مدل نه

۰,۵	$\alpha$
۰,۵	$\beta$
۲۵	W (تعداد کلمه‌ها در دیکشنری)
۲	T تعداد عنوان‌ها
۱۰۰	Max Epoch حداقل تعداد ایپاک
دیتاست یک	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-9-dataset1** موجود است، خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت-تصویر ها :Figure\_1.png

تصویر عنوان های به دست آمده (Topics) :Figure\_2.png

تصویر Perplexity در حین آموزش :Figure\_3.png

یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، Perplexity حداکثر تعداد ایپاک‌ها، زمان انجام کلیه ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\theta$  ،  $\Phi$  ، میزان در هر ایپاک است.

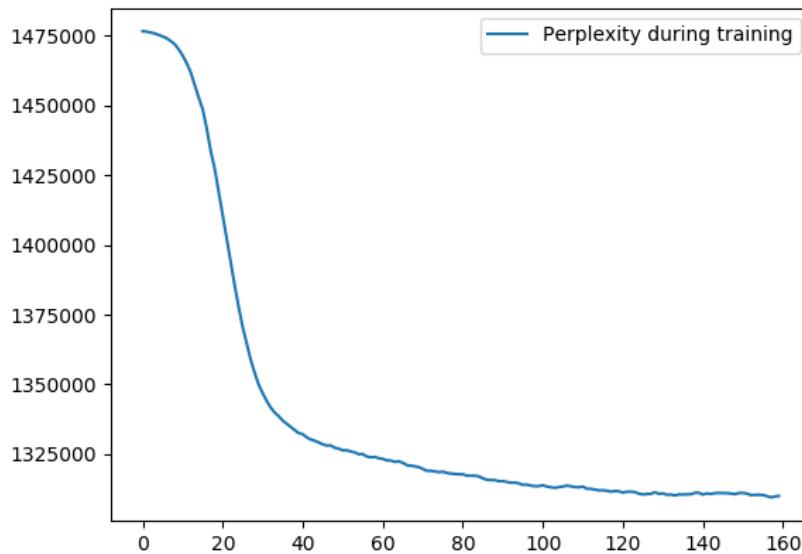
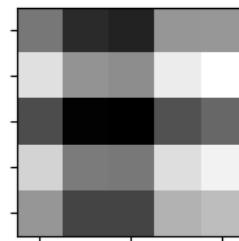
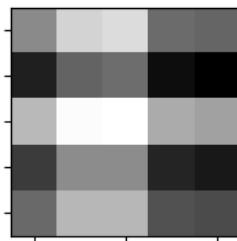
در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۲۲ خروجی مدل نه

تصویر تعدادی از داکیومنت‌ها	some randomly selected samples

topics

تصویر عنوان‌های به  
دست آمده  
Topics



Perplexity  
در ایپاک‌های مختلف

۶۷۶,۵۳۲۴۷۹۷۶۳۰۳۱ S

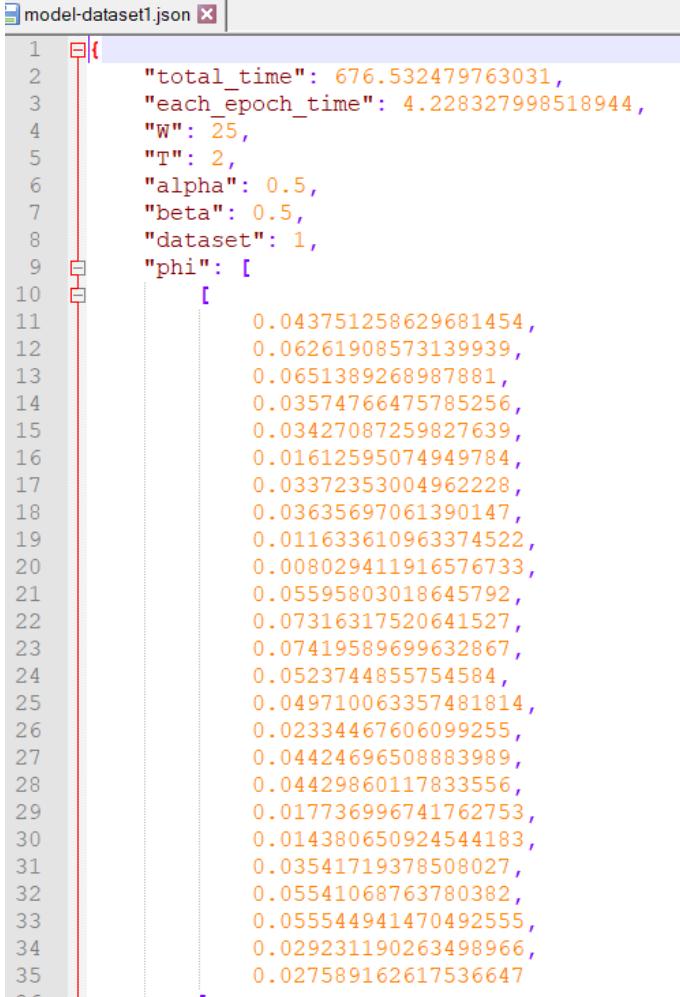
زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۴,۲۲۸۳۲۷۹۹۸۵۱۸۹۴۴ S

زمان میانگین برای  
انجام یک ایپاک

160

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

<p style="text-align: center;">676.532479763031 s</p>	زمان لازم برای رسیدن به حالت mixing
 <pre> 1  { 2   "total_time": 676.532479763031, 3   "each_epoch_time": 4.228327998518944, 4   "W": 25, 5   "T": 2, 6   "alpha": 0.5, 7   "beta": 0.5, 8   "dataset": 1, 9   "phi": [ 10    [ 11      0.043751258629681454, 12      0.06261908573139939, 13      0.0651389268987881, 14      0.03574766475785256, 15      0.03427087259827639, 16      0.01612595074949784, 17      0.03372353004962228, 18      0.03635697061390147, 19      0.011633610963374522, 20      0.008029411916576733, 21      0.05595803018645792, 22      0.07316317520641527, 23      0.07419589699632867, 24      0.0523744855754584, 25      0.049710063357481814, 26      0.02334467606099255, 27      0.04424696508883989, 28      0.04429860117833556, 29      0.017736996741762753, 30      0.014380650924544183, 31      0.03541719378508027, 32      0.05541068763780382, 33      0.055544941470492555, 34      0.029231190263498966, 35      0.027589162617536647 36         ]       ]     }   </pre>	

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است.

#### ۴.۱.۲ مدل ده

پارامترهای مختلف مدل:

جدول ۲۳ پارامترهای مدل ده

•, ۵	$\alpha$

$\beta$	•,٥
W(تعداد کلمه‌ها در دیکشنری)	٢٥
T تعداد عنوان‌ها	٥
Max Epoch حداکثر تعداد ایپاک	١٠٠
دیتاست دیتاست یک	

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-10-dataset1** موجود است،

خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت-تصویر ها :**Figure\_1.png**

تصویر عنوان‌های به دست آمده (Topics) :**Figure\_2.png**

تصویر Perplexity در حین آموزش :**Figure\_3.png**

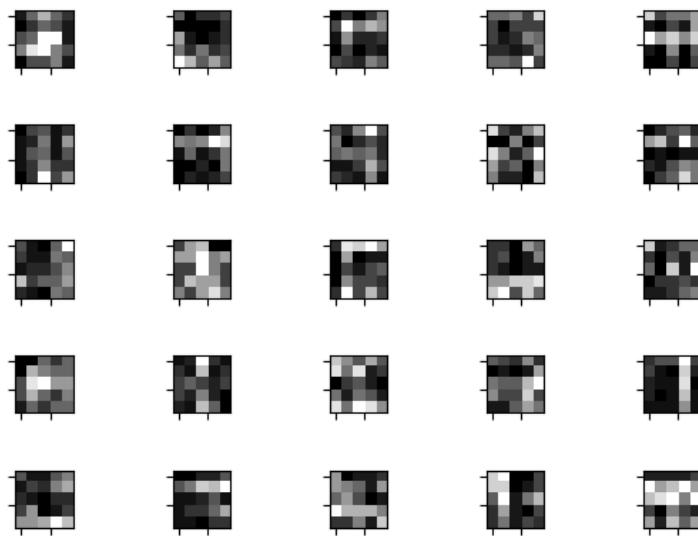
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\theta$  ،  $\Phi$  ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

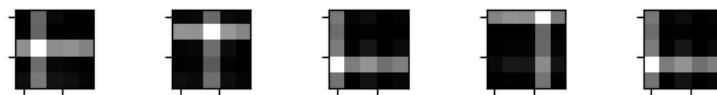
جدول ۲۴ خروجی مدل ده

تصویر تعدادی از  
دکیومنټها

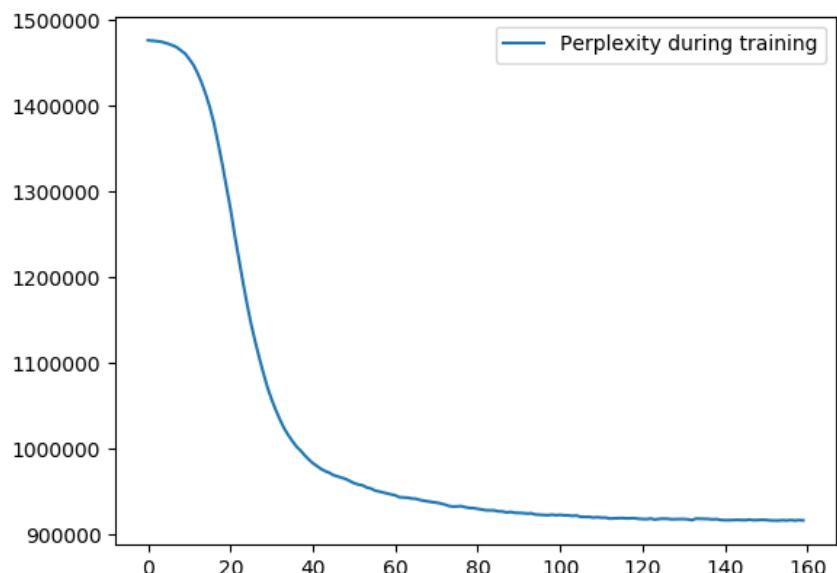
some randomly selected samples



تصویر عنوان‌های به  
دست آمده  
Topics



Perplexity  
در ایپاک‌های مختلف



651,7941944599152 s	زمان کل اجرای حداکثر ایپاک‌های مجاز
۴,۰۷۳۷۱۳۷۱۵۳۷۴۴۷ s	زمان میانگین برای انجام یک ایپاک
160	ایپاک‌های لازم برای رسیدن به حالت Mixing
651.7941944599152 s	زمان لازم برای رسیدن به حالت mixing

فایل  
model-dataset1.json  
در پوششی  
Model-10-dataset-1  
که شامل،  $\Phi$ ،  $\theta$  و  
سایر خروجی‌ها است

```

model-dataset1.json | 
1  {
2      "total_time": 651.7941944599152,
3      "each_epoch_time": 4.07371371537447,
4      "W": 25,
5      "T": 5,
6      "alpha": 0.5,
7      "beta": 0.5,
8      "dataset": 1,
9      "phi": [
10         [
11             0.005722260990928122,
12             0.007194062044027152,
13             0.07725686734758612,
14             0.007980714331028357,
15             0.005747636871153968,
16             0.10740341305589038,
17             0.10593161200279134,
18             0.18655078348030196,
19             0.11120979508976718,
20             0.09722768508532639,
21             0.003768318213538032,
22             0.0057730127513798135,
23             0.07936306540633128,
24             0.009554018905030768,
25             0.0009515955084692001,
26             0.0012814819514051893,
27             0.0028040347649559093,
28             0.0662183594493434,
29             0.0025756518429233015,
30             0.00039332614350060267,
31             0.010822812916323034,
32             0.008691238977352026,
33             0.08129163230349552,
34             0.010569054114064582,
35             0.0037175664530863413
36         ],
37         [
38             0.0031014510805256273,
39             0.0052438188951859,

```

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف را به کاهش است.

### ۴.۱.۳ مدل یازده

پارامترهای مختلف مدل:

جدول ۲۵ پارامترهای مدل یازده

$\alpha$	$\beta$
۰.۵	۰.۵

$\beta$	۰,۵
W (تعداد کلمه‌ها در دیکشنری)	۲۵
T تعداد عنوان‌ها	۷
Max Epoch حداکثر تعداد ایپاک	۱۰۰
دیتابست	دیتابست یک

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-11-dataset1** موجود است،

خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت-تصویر ها :**Figure\_1.png**

تصویر عنوان‌های به دست آمده (Topics) :**Figure\_2.png**

تصویر Perplexity در حین آموزش :**Figure\_3.png**

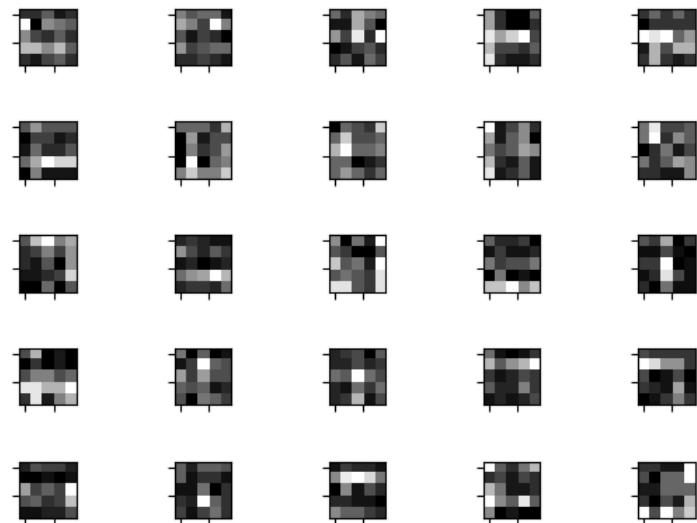
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\theta$ ،  $\Phi$ ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۲۶ خروجی مدل پازدوه

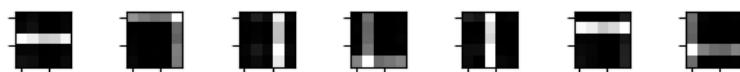
تصویر تعدادی از  
دکیومنت‌ها

some randomly selected samples

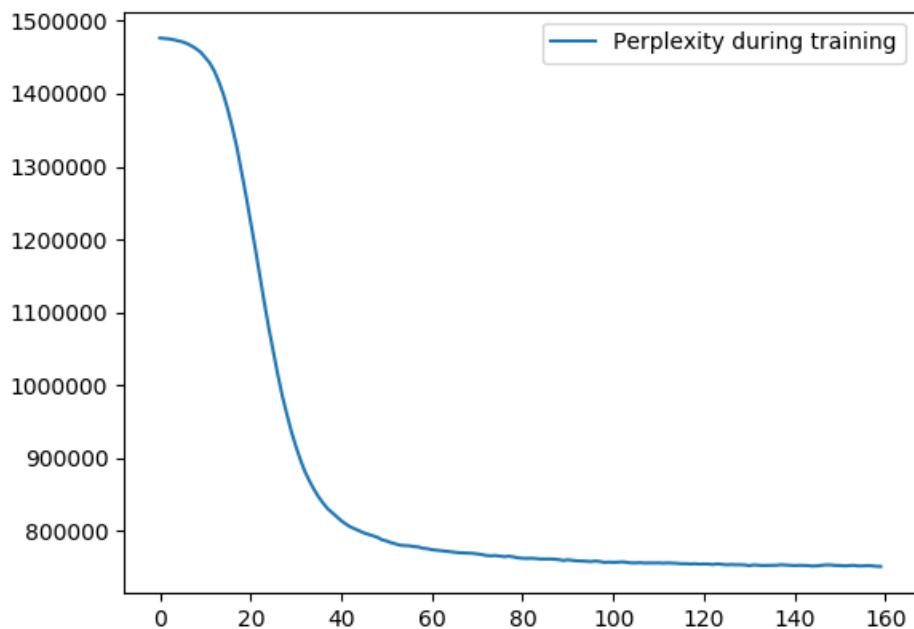


تصویر عنوان‌های به  
دست آمده  
Topics

topics



Perplexity  
در ایپاک‌های مختلف



زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۷۲۵,۴۰۸۶۳۳۲۳۲۱۱۶۷ s

زمان میانگین برای  
انجام یک ایپاک

۴,۵۳۳۸۰۳۹۵۷۷۰۰۷۲۹ s

ایپاک‌های لازم برای  
رسیدن به حالت  
Mixing

160

زمان لازم برای رسیدن  
به حالت mixing

۷۲۵,۴۰۸۶۳۳۲۳۲۱۱۶۷ s

```

model-dataset1.json | فایل
1   {
2     "total_time": 725.4086332321167,
3     "each_epoch_time": 4.533803957700729,
4     "W": 25,
5     "T": 7,
6     "alpha": 0.5,
7     "beta": 0.5,
8     "dataset": 1,
9     "phi": [
10       [
11         0.0023169330815209985,
12         0.007690660228578104,
13         0.00025311033663674774,
14         5.841007768540332e-05,
15         0.004302875722824711,
16         0.005548957380113315,
17         0.013142267479215748,
18         1.947002589513444e-05,
19         9.73501294756722e-05,
20         0.005471077276532778,
21         0.2025466793870836,
22         0.1888787212086992,
23         0.16169856505909153,
24         0.16551469013453787,
25         0.19390198788964388,
26         0.004224995619244174,
27         0.0050037966550495515,
28         0.0006425108545394365,
29         0.0010708514242323943,
30         0.0025895134440528807,
31         0.013531667997118435,
32         0.01092268452717042,
33         0.0013434317867642763,
34         0.0009929713206518565,
35         0.008235820953641868
36       ],
37       [
38         0.11884613656447048,
39         0.10204112765138467

```

فایل  
model-dataset1.json  
در پوششی  
Model-11-dataset-1  
که شامل  $\Phi$ ،  $\theta$  و  
سایر خروجی‌ها است

همین‌طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است.

#### ۴.۱.۴ مدل دوازده

پارامترهای مختلف مدل:

جدول ۲۷ پارامترهای مدل دوازده

۰,۵	$\alpha$
۰,۵	$\beta$
۲۵	W(تعداد کلمه‌ها در دیکشنری)
۱۰	T تعداد عنوان‌ها
۱۰۰	Max Epoch حداکثر تعداد ایپاک
دیتاست یک	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی موجود است، **model-12-dataset1** می‌باشد.

خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت-تصویر ها **Figure\_1.png**

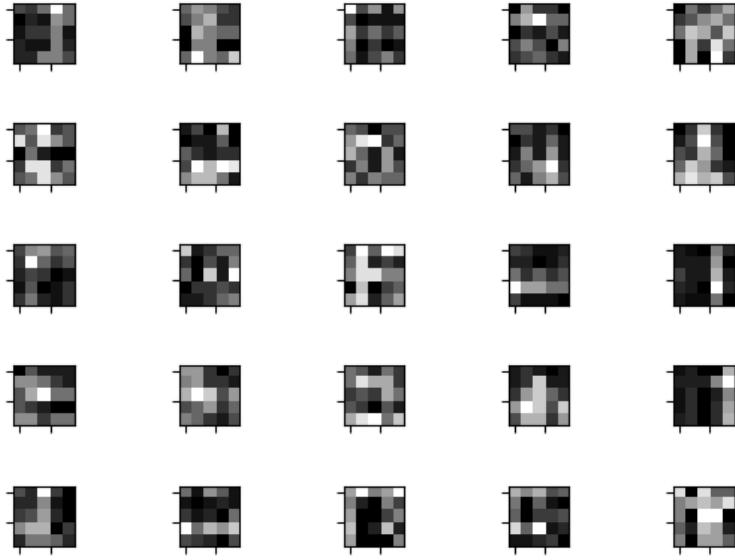
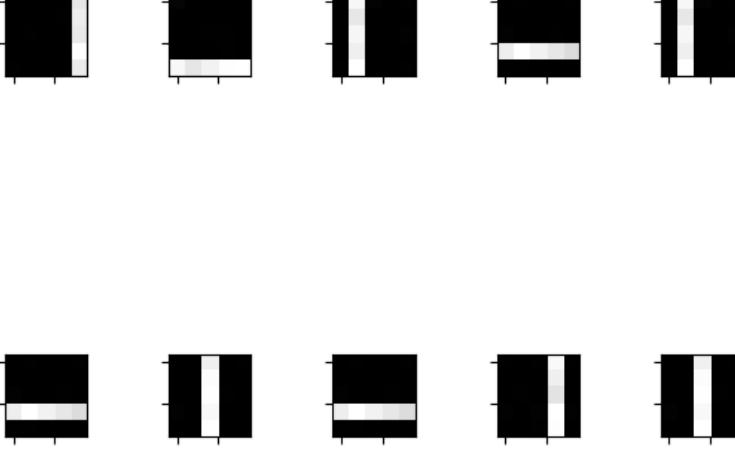
تصویر عنوان‌های به دست آمده (Topics) **Figure\_2.png**

تصویر Perplexity در حین آموزش **Figure\_3.png**

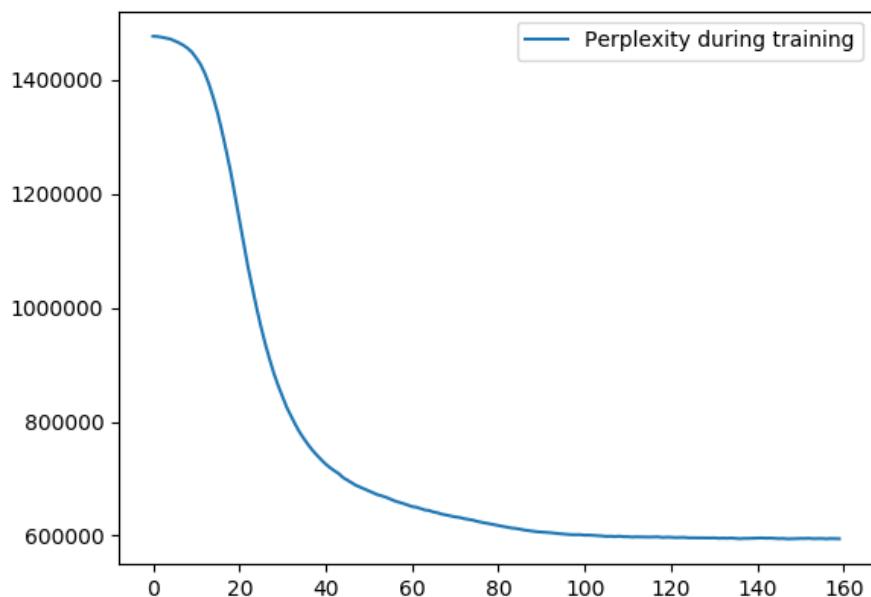
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۲۱ خروجی مدل دوازده

some randomly selected samples		تصویر تعدادی از دکیومنټها
		تصویر عنوانهای به دست آمده Topics

Perplexity  
در ایپاک‌های مختلف



۷۳۰,۷۵۸۸۳۵۷۹۲۵۴۱۵ S

زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۴,۵۶۷۲۴۲۷۲۳۷۰۳۳۸۵ S

زمان میانگین برای  
انجام یک ایپاک

120

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

$120 * 4.567 = 548.04 S$

زمان لازم برای رسیدن  
mixing به حالت

```

model-dataset1.json x
1  {
2      "total_time": 730.7588357925415,
3      "each_epoch_time": 4.567242723703385,
4      "W": 25,
5      "T": 10,
6      "alpha": 0.5,
7      "beta": 0.5,
8      "dataset": 1,
9      "phi": [
10         [
11             0.00017330593449035677,
12             0.0014112054665643336,
13             0.000321853878339234,
14             0.00017330593449035677,
15             0.0008665296724517838,
16             0.0001237899532073977,
17             2.4757990641479537e-05,
18             0.00017330593449035677,
19             0.0002723378970562749,
20             0.00022282191577331585,
21             0.0020053972419598427,
22             0.0004704018221881112,
23             0.00022282191577331585,
24             0.0009160456537347429,
25             0.0002723378970562749,
26             0.1985343269540244,
27             0.21279492956351662,
28             0.20190141368126563,
29             0.19259240920006931,
30             0.18323388873759006,
31             0.0014607214478472928,
32             0.0012626575227154564,
33             0.0004704018221881112,
34             2.4757990641479537e-05,
35             7.427397192443861e-05
36         ],
37     ],
38     [
39         0.19522803529265564,
40         0.19681869019510376

```

فایل  
model-dataset1.json  
در پوشیده  
Model-12-dataset-1  
که شامل  $\Phi$  و  $\theta$  و  
سایر خروجی‌ها است

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است.

## ۴.۱.۵ مدل سیزده

پارامترهای مختلف مدل:

جدول ۲۹ پارامترهای مدل سیزده

۰,۵	$\alpha$
۰,۵	$\beta$
۲۵	W(تعداد کلمه‌ها در دیکشنری)
۱۵	T تعداد عنوان‌ها
۱۰۰	Max Epoch حداکثر تعداد ایپاک
دیتاست یک	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی موجود است، **model-13-dataset1** می‌باشد.

خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت-تصویر ها: **Figure\_1.png**

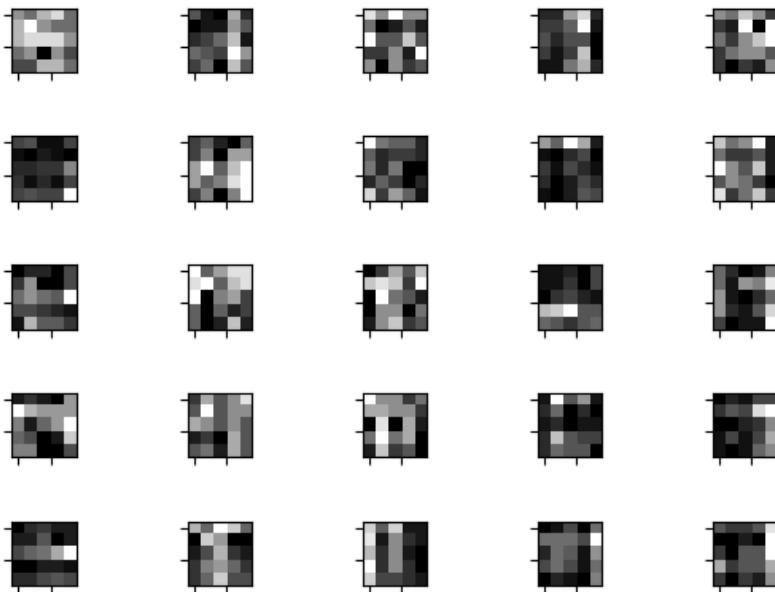
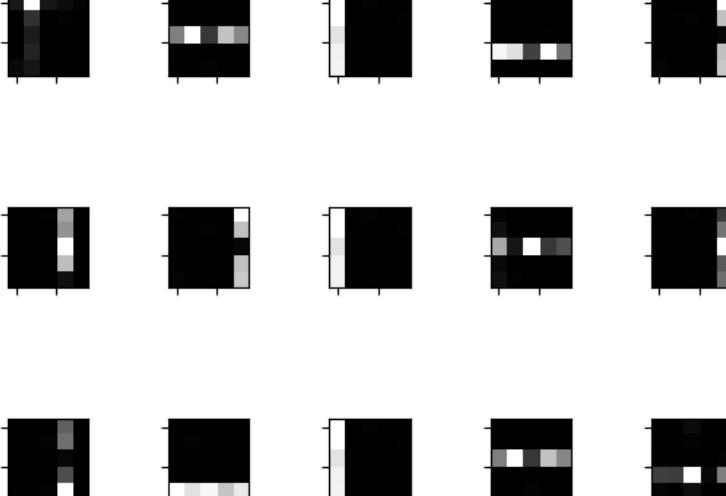
تصویر عنوان‌های به دست آمده (Topics): **Figure\_2.png**

تصویر Perplexity در حین آموزش: **Figure\_3.png**

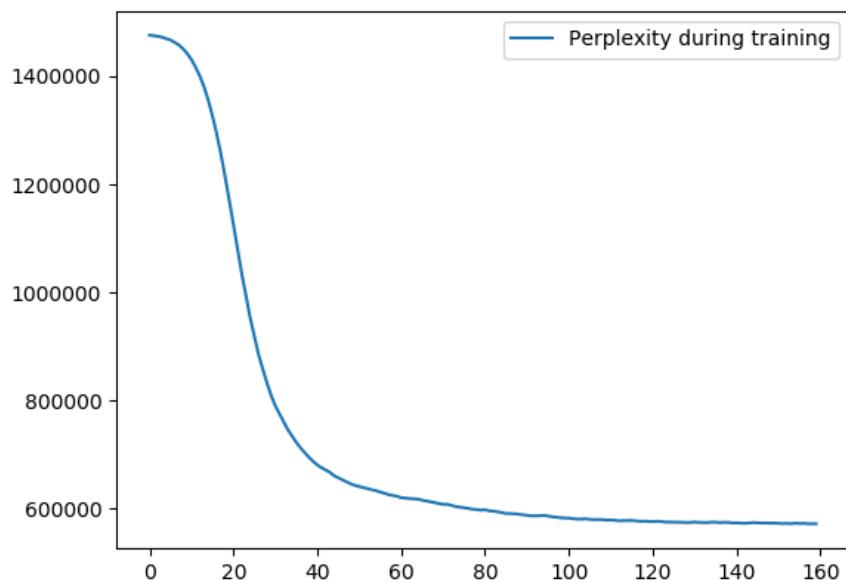
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۳۰ خروجی مدل سیزده

some randomly selected samples	تصویر تعدادی از داکیومنت‌ها
	تصویر عنوان‌های به دست آمده Topics
	

Perplexity  
در ایپاک‌های مختلف



زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۸۰۵,۴۵۸۰۵۳۵۸۸۸۸۶۷۲ s

زمان میانگین برای  
انجام یک ایپاک

۵,۰۳۴۱۱۲۸۳۴۹۳۰۴۲ s

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

160

زمان لازم برای رسیدن  
mixing به حالت

۸۰۵,۴۵۸۰۵۳۵۸۸۸۸۶۷۲ s

فایل

model-dataset1.json

در پوشه‌ی

Model-13-dataset-1

که شامل ،  $\Phi$ ،  $\theta$  و

سایر خروجی‌ها است

```
model-dataset1.json x |  
1  {  
2      "total_time": 805.4580535888672,  
3      "each_epoch_time": 5.03411283493042  
4      "W": 25,  
5      "T": 15,  
6      "alpha": 0.5,  
7      "beta": 0.5,  
8      "dataset": 1,  
9      "phi": [  
10         [  
11             0.00026077130356677195,  
12             0.00014487294642598442,  
13             0.0006664155535595282,  
14             0.00014487294642598442,  
15             0.00043461883927795324,  
16             8.692376785559064e-05,  
17             0.002057195839248979,  
18             0.00020282212499637817,  
19             0.00014487294642598442,  
20             0.00037666966070755946,  
21             0.00043461883927795324,  
22             2.8974589285196883e-05,  
23             0.000492568017848347,  
24             2.8974589285196883e-05,  
25             0.000724364732129922,  
26             0.00026077130356677195,  
27             0.00026077130356677195,  
28             0.0006084663749891346,  
29             0.0008402630892707096,  
30             8.692376785559064e-05,  
31             0.2184394286210993,  
32             0.1934633326572596,  
33             0.21107988294265928,  
34             0.16877698258627183,  
35             0.1999536406571437  
36         ],  
37         [  
38             0.0058490477750222265,  
39             0.0007018857330026672,  
40             0.0007954704974030228,  
41             0.0004211314398016003,
```

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است.

## ۴.۱.۶ مدل چهارده

پارامترهای مختلف مدل:

جدول ۳۱ پارامترهای مدل چهارده

۰,۵	$\alpha$
۰,۵	$\beta$
۲۵	W(تعداد کلمه‌ها در دیکشنری)
۲۰	T تعداد عنوان‌ها
۱۰۰	Max Epoch حداکثر تعداد ایپاک
دیتاست یک	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی موجود است، **model-14-dataset1** می‌باشد.

خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت‌ تصویر ها : **Figure\_1.png**

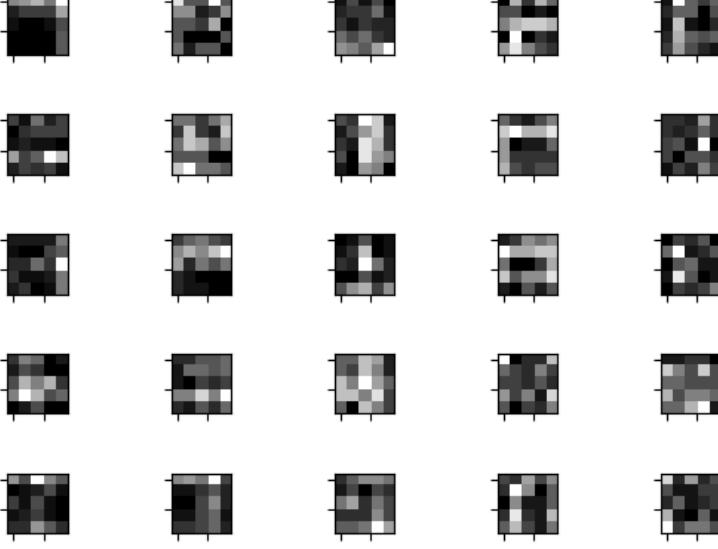
تصویر عنوان‌های به دست آمده (Topics) : **Figure\_2.png**

تصویر Perplexity در حین آموزش : **Figure\_3.png**

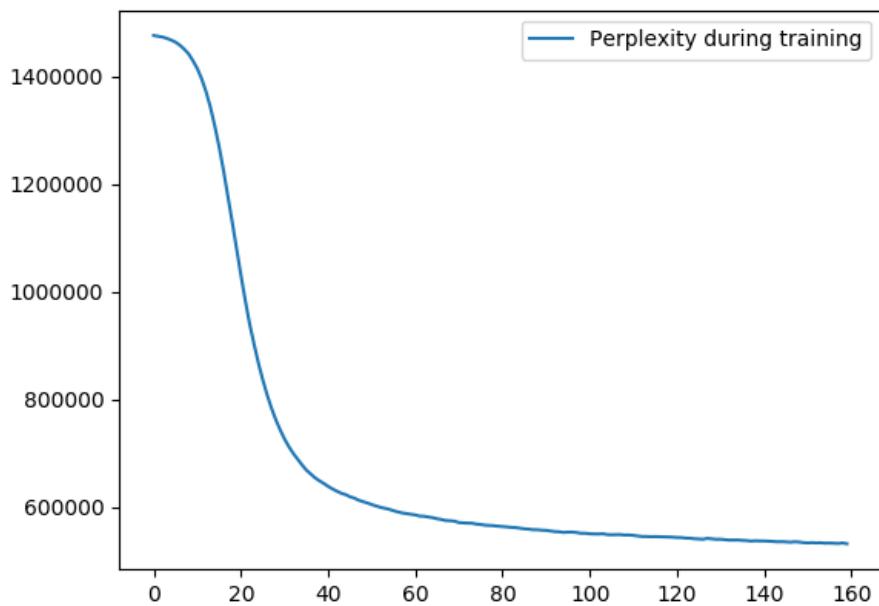
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۳۲ خروجی مدل چهارده

some randomly selected samples	تصویر تعدادی از داکیومنات‌ها
	تصویر عنوان‌های به دست آمده Topics

Perplexity  
در ایپاک‌های مختلف



زمان کل اجرای حداکثر  
ایپاک‌های مجاز

زمان میانگین برای  
انجام یک ایپاک

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

زمان لازم برای رسیدن  
به حالت mixing

160

۹۰۰,۲۸۸۲۷۷۱۴۹۲۰۰۴۵

۵,۶۲۶۸۰۱۷۳۲۱۸۲۵۰۳۵

۹۰۰,۲۸۸۲۷۷۱۴۹۲۰۰۴۵

فایل

model-dataset1.json

در پوشه‌ی

Model-14-dataset-1

که شامل ،  $\Phi$ ،  $\theta$  و

سایر خروجی‌ها است

```
model-dataset1.json | C:\Users\Home\Dropbox\codes\PGM\PGM_S18_P2_Report_96131125\|  
2     "total_time": 900.2882771492004,  
3     "each_epoch_time": 5.626801732182503,  
4     "W": 25,  
5     "T": 20,  
6     "alpha": 0.5,  
7     "beta": 0.5,  
8     "dataset": 1,  
9     "phi": [  
10        [  
11            0.00017300040366760857,  
12            0.0006343348134478981,  
13            0.0016723372354535493,  
14            0.0022490052476789113,  
15            0.0009803356207831152,  
16            0.0006343348134478981,  
17            0.0014416700305634046,  
18            0.00017300040366760857,  
19            0.0004036676085577533,  
20            0.0007496684158929704,  
21            0.0009803356207831152,  
22            0.0013263364281183322,  
23            0.003056340464794418,  
24            0.0013263364281183322,  
25            0.0012110028256732599,  
26            0.144224669857563,  
27            0.012513695865290352,  
28            0.062337812121561614,  
29            0.6861772677469581,  
30            0.07087249870249697,  
31            0.003402341272129635,  
32            0.00028833400611268095,  
33            0.0009803356207831152,  
34            0.001903004440343694,  
35            0.00028833400611268095  
36        ],  
37        [  
38            5.300259712725924e-05,  
39            0.001113054539672444,
```

همین‌طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است.

## ۴.۲ دیتاست دو-تغییر تعداد عنوان‌ها

در این قسمت مدل‌های مختلفی را با دیتاست دو ایجاد می‌کنیم که فقط در تعداد عنوان‌ها با یک دیگر متفاوت‌اند.

### ۴.۲.۱ مدل پانزده

پارامترهای مختلف مدل:

جدول ۳۳ پارامترهای مدل پانزده

۱	$\alpha$
۱	$\beta$
۱۰۴۷۳	W (تعداد کلمه‌ها در دیکشنری)
۲	T (تعداد عنوان‌ها)
۱۲۰	Max Epoch (حداکثر تعداد ایپاک)
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-15-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش **Figure\_1.png**

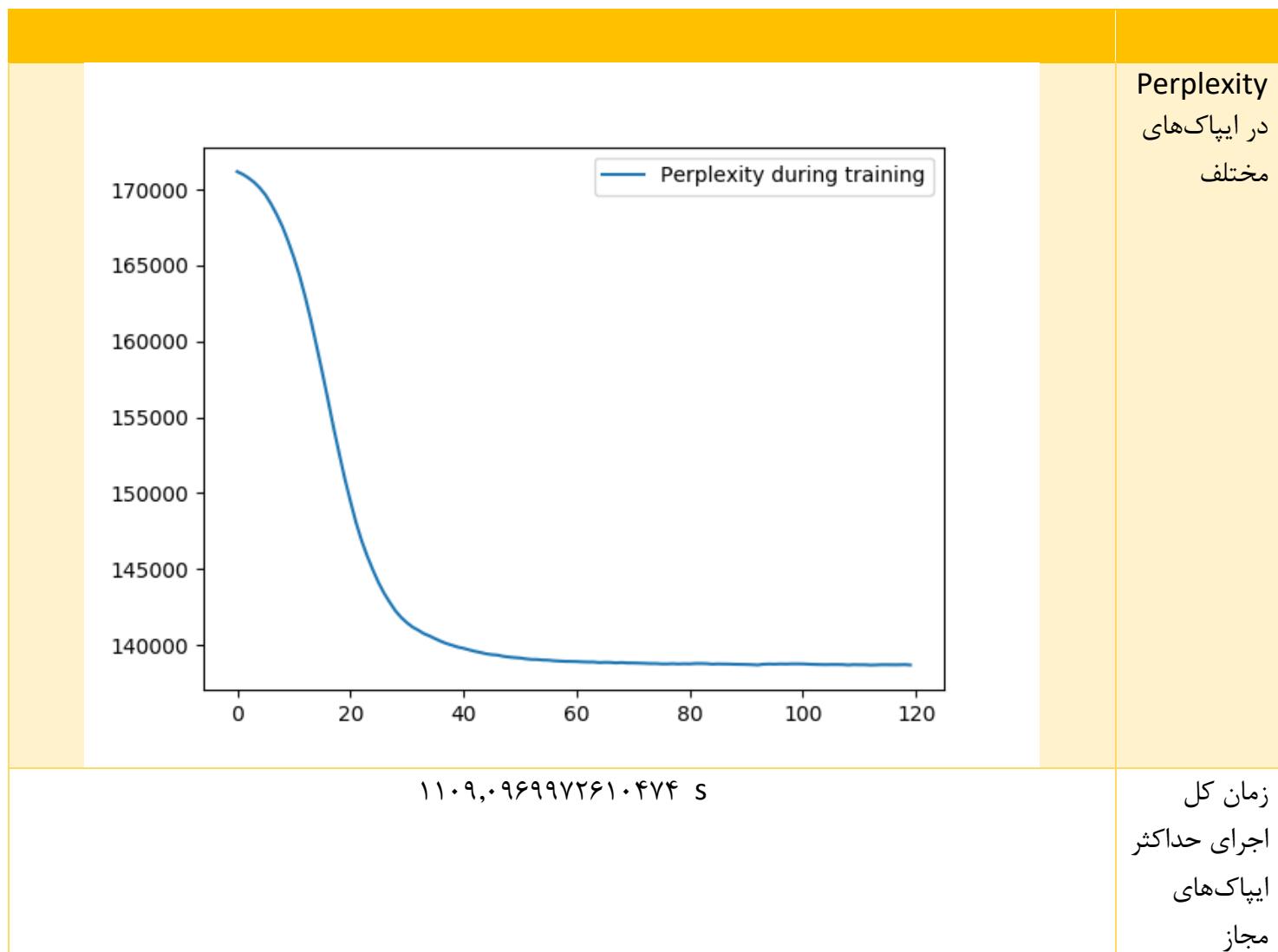
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\theta$ ،  $\Phi$ ، میزان Perplexity در هر ایپاک است.

در فایل `model-dataset2.json`, مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل `topicc-.topics.txt`

کلمه‌ها را از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتدا قرار می‌دهیم. به این ترتیب عنوان‌ها در فایل `topics.txt` قابل مشاهده‌اند. (برای نمایش بهتر از یک `text-editor` خوب مانند `notepad++` استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۳۴ خروجی مدل شانزده



$9.242474977175394 \text{ s}$  $100 \text{ Epochs}$	زمان میانگین برای انجام یک ایپاک
$80 * 9.2424 = 739.392 \text{ s}$	زمان لازم برای رسیدن به حالت <b>Mixing</b>
<pre>model-dataset2.json ✘  </pre> <pre> 1  { 2    "total_time": 1109.0969972610474, 3    "each_epoch_time": 9.242474977175394, 4    "W": 10473, 5    "T": 2, 6    "alpha": 1, 7    "beta": 1, 8    "dataset": 2, 9    "phi": [ 10      [ 11        0.006223930645293203, 12        0.0027889097898094383, 13        2.0446552711212888e-05, 14        0.005381532673591233, 15        0.0014885090373762983, 16        0.00483765437147297, 17        0.0009487200458002781, 18        0.0021836918295575366, 19        0.002584444262697309, 20        0.003439110166026008, 21        0.003353234644638914, 22        0.002191870450642022, 23        0.004432812627790954, 24        0.0028134456530628936, 25        0.0016970638750306698, 26        0.0029524822114991413, 27        0.002032387339494561, 28        0.001832011122924675, 29        7.36075897603664e-05, 30        0.0023186390774515417, 31        0.002302281835282571, 32        8.178621084485156e-06, 33        0.002032387339494561, 34        0.0017502249120798234, 35        0.002191870450642022, 36        0.0014189907581581744, 37        0.001594831114746055, 38        0.0008015048662795453, </pre>	فایل <b>model-dataset1.json</b> در پوشه‌ی <b>Model-15-dataset-2</b> که شامل ، $\theta$ و سایر $\Phi$ خروجی‌ها است

<pre>topics.txt x 1 ['i', 'people', 'two', 'police', 'government', 'years', 'officials', 'state', 'new', 'last', 'city', 'time', 'three', 'cc 2 ['percent', 'new', 'million', 'year', 'president', 'bush', 'billion', 'last', 'states', 'market', 'company', 'government' 3</pre>	<p>فایل <b>topics.txt</b> در پوشه‌ی Model-15- dataset-2 که شامل عنوان‌ها به صورت کلمه است</p>
۱۳۸۶۷۲,۵۵۳۰ ۱۲۳۰ ۵۳۲	<p><b>Perplexity</b> نهایی</p>

همین‌طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

## ۴.۲.۲ مدل شانزده

پارامترهای مختلف مدل:

جدول ۳۵ پارامترهای مدل شانزده

۱	$\alpha$
۱	$\beta$
۱۰۴۷۳	W(تعداد کلمه‌ها در دیکشنری)
۱۰	T تعداد عنوان‌ها
۱۲۰	Max Epoch حداکثر تعداد ایپاک
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-16-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش :**Figure\_1.png**

یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، Perplexity، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\theta$ ،  $\Phi$ ، میزان خوبی ایپاک است.

در فایل **model-dataset2.json**، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل **topic-  
topics.txt**، کلمه‌ها رو از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتدا قرار می‌دهیم، به این ترتیب عنوان‌ها در فایل **topics.txt** قابل مشاهده اند. (برای نمایش بهتر از یک خوب مانند notepad++ text-editor استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۳۶ خروجی مدل شانزده

	<table border="1"> <thead> <tr> <th>Epoch</th> <th>Perplexity</th> </tr> </thead> <tbody> <tr><td>0</td><td>~175,000</td></tr> <tr><td>20</td><td>~105,000</td></tr> <tr><td>40</td><td>~95,000</td></tr> <tr><td>60</td><td>~88,000</td></tr> <tr><td>80</td><td>~85,000</td></tr> <tr><td>100</td><td>~82,000</td></tr> </tbody> </table>	Epoch	Perplexity	0	~175,000	20	~105,000	40	~95,000	60	~88,000	80	~85,000	100	~82,000	Perplexity در ایپاک‌های مختلف
Epoch	Perplexity															
0	~175,000															
20	~105,000															
40	~95,000															
60	~88,000															
80	~85,000															
100	~82,000															
1548,4073979854584 s		زمان کل اجرای حداکثر ایپاک‌های مجاز														
15.484073979854584 s		زمان میانگین برای انجام یک ایپاک														
100 Epochs		ایپاک‌های لازم برای رسیدم به حالت Mixing														

```

model-dataset2.json x |
1   {
2     "total_time": 1548.4073979854584,
3     "each_epoch_time": 15.484073979854584,
4     "W": 10473,
5     "T": 10,
6     "alpha": 1,
7     "beta": 1,
8     "dataset": 2,
9     "phi": [
10       [
11         0.003553299492385787,
12         0.0028426395939086294,
13         0.002050761421319797,
14         0.0019898477157360406,
15         0.0040609137055837565,
16         0.0011573604060913705,
17         0.0033096446700507614,
18         0.0026802030456852793,
19         0.003187817258883249,
20         0.0011776649746192893,
21         0.0020913705583756343,
22         0.002517766497461929,
23         2.030456852791878e-05,
24         0.004812182741116751,
25         0.0024568527918781727,
26         0.0005685279187817259,
27         2.030456852791878e-05,
28         2.030456852791878e-05,
29         0.002071065989847716,
30         0.0013401015228426396,
31         0.0008527918781725888,
32         0.00030456852791878173,
33         0.0006091370558375635,
34         0.001847715736040609,
35         0.001197969543147208,
36         0.0010964467005076142,

```

است	صورت کلمه
۸۹۱۲۶,۶۲۲۰۳۴۷۰۶۵۲	Perplexity نهایی

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

## ۴.۲.۳ مدل هفده

پارامترهای مختلف مدل:

جدول ۳۷ پارامترهای مدل هفدهم

۱	$\alpha$
۱	$\beta$
۱۰۴۷۳	W(تعداد کلمه‌ها در دیکشنری)
۲۰	T تعداد عنوان‌ها
۱۲۰	Max Epoch حداکثر تعداد ایپاک
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-17-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش **Figure\_1.png**

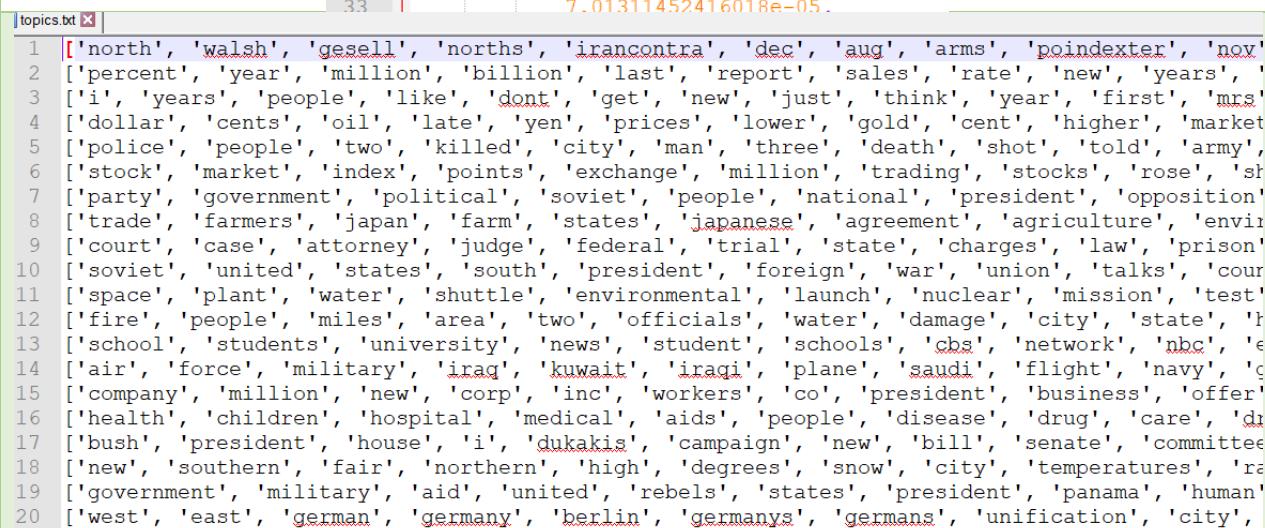
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$ ،  $\theta$ ، میزان Perplexity در هر ایپاک است.

در فایل **model-dataset2.json**، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل **topic-**topics.txt، کلمه‌ها را از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتداء قرار می‌دهیم. به این ترتیب عنوان‌ها در فایل topics.txt قابل مشاهده اند. (برای نمایش بهتر از یک text-editor خوب مانند notepad++ استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۳۱ خروجی مدل هفتدهم

<p>Perplexity during training</p> <table border="1"> <thead> <tr> <th>Epoch</th> <th>Perplexity</th> </tr> </thead> <tbody> <tr><td>0</td><td>~185,000</td></tr> <tr><td>20</td><td>~105,000</td></tr> <tr><td>40</td><td>~88,000</td></tr> <tr><td>60</td><td>~85,000</td></tr> <tr><td>80</td><td>~84,000</td></tr> <tr><td>100</td><td>~83,000</td></tr> <tr><td>120</td><td>~82,000</td></tr> </tbody> </table>	Epoch	Perplexity	0	~185,000	20	~105,000	40	~88,000	60	~85,000	80	~84,000	100	~83,000	120	~82,000	Perplexity در ایپاک‌های مخالف
Epoch	Perplexity																
0	~185,000																
20	~105,000																
40	~88,000																
60	~85,000																
80	~84,000																
100	~83,000																
120	~82,000																
۲۶۲۱,۲۶۱۶۵۱۵۱۸۹۶۰۷ s	زمان کل اجرای حداکثر ایپاک‌های مجاز																
21.84384709596634 s	زمان میانگین برای انجام یک ایپاک																
120 Epochs	ایپاک‌های لازم برای رسیدم به حالت Mixing																

<b>۲۶۲۱,۲۶۱۶۵۱۵۱۵۹۶۰۷۵</b>	<b>زمان لازم برای رسیدن به حالت mixing</b>
 <pre> model-dataset2.json 1   { 2     "total_time": 2621.2616515159607, 3     "each_epoch_time": 21.84384709596634, 4     "W": 10473, 5     "T": 20, 6     "alpha": 1, 7     "beta": 1, 8     "dataset": 2, 9     "phi": [ 10       [ 11         7.01311452416018e-05, 12         7.01311452416018e-05, 13         7.01311452416018e-05, 14         0.00021039343572480537, 15         7.01311452416018e-05, 16         0.000350655726208009, 17         0.00021039343572480537, 18         7.01311452416018e-05, 19         7.01311452416018e-05, 20         0.00021039343572480537, 21         7.01311452416018e-05, 22         0.000350655726208009, 23         7.01311452416018e-05, 24         7.01311452416018e-05, 25         0.0001402622904832036, 26         7.01311452416018e-05, 27         0.0001402622904832036, 28         7.01311452416018e-05, 29         7.01311452416018e-05, 30         0.0001402622904832036, 31         7.01311452416018e-05, 32         7.01311452416018e-05, 33         7.01311452416018e-05. </pre>	<b>فایل model- dataset1.json در پوششی Model-17- dataset-2 که شامل ، <math>\theta</math> ، <math>\Phi</math> و سایر خروجی‌ها است</b>
 <pre> topics.txt 1  ['north', 'walsh', 'gesell', 'norths', 'irancontra', 'dec', 'aug', 'arms', 'poindexter', 'nov', 2  ['percent', 'year', 'million', 'billion', 'last', 'report', 'sales', 'rate', 'new', 'years', 'i', 3  ['i', 'years', 'people', 'like', 'dont', 'get', 'new', 'just', 'think', 'year', 'first', 'mrs', 4  ['dollar', 'cents', 'oil', 'late', 'yen', 'prices', 'lower', 'gold', 'cent', 'higher', 'market', 5  ['police', 'people', 'two', 'killed', 'city', 'man', 'three', 'death', 'shot', 'told', 'army', 6  ['stock', 'market', 'index', 'points', 'exchange', 'million', 'trading', 'stocks', 'rose', 'si', 7  ['party', 'government', 'political', 'soviet', 'people', 'national', 'president', 'opposition', 8  ['trade', 'farmers', 'japan', 'farm', 'states', 'japanese', 'agreement', 'agriculture', 'envir 9  ['court', 'case', 'attorney', 'judge', 'federal', 'trial', 'state', 'charges', 'law', 'prison', 10 ['soviet', 'united', 'states', 'south', 'president', 'foreign', 'war', 'union', 'talks', 'cour 11 ['space', 'plant', 'water', 'shuttle', 'environmental', 'launch', 'nuclear', 'mission', 'test', 12 ['fire', 'people', 'miles', 'area', 'two', 'officials', 'water', 'damage', 'city', 'state', 'i 13 ['school', 'students', 'university', 'news', 'student', 'schools', 'cbs', 'network', 'nbc', 'e 14 ['air', 'force', 'military', 'iraq', 'kuwait', 'iraqi', 'plane', 'saudi', 'flight', 'navy', 'c 15 ['company', 'million', 'new', 'corp', 'inc', 'workers', 'co', 'president', 'business', 'offer', 16 ['health', 'children', 'hospital', 'medical', 'aids', 'people', 'disease', 'drug', 'care', 'dr 17 ['bush', 'president', 'house', 'i', 'dukakis', 'campaign', 'new', 'bill', 'senate', 'committee 18 ['new', 'southern', 'fair', 'northern', 'high', 'degrees', 'snow', 'city', 'temperatures', 'ra 19 ['government', 'military', 'aid', 'united', 'rebels', 'states', 'president', 'panama', 'human' 20 ['west', 'east', 'german', 'germany', 'berlin', 'germans', 'germans', 'unification', 'city', </pre>	<b>فایل topics.txt در پوششی Model-17- dataset-2 که شامل عنوان‌ها به صورت کلمه است</b>
<b>۸۳۷۶۲,۸۶۷۰۶۱۸۰۳۵۸</b>	<b>Perplexity نهایی</b>

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

#### ۴.۲.۴ مدل هجده

پارامترهای مختلف مدل:

جدول ۳۹ پارامترهای مدل هجدهم

۱	$\alpha$
۱	$\beta$
۱۰۴۷۳	W (تعداد کلمه‌ها در دیکشنری)
۳۰	T (تعداد عنوان‌ها)
۱۲۰	Max Epoch (حداکثر تعداد ایپاک)
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیهی خروجی‌های حاصل از اجرای کد برای مدل یک در پوشی **model-18-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش: **Figure\_1.png**

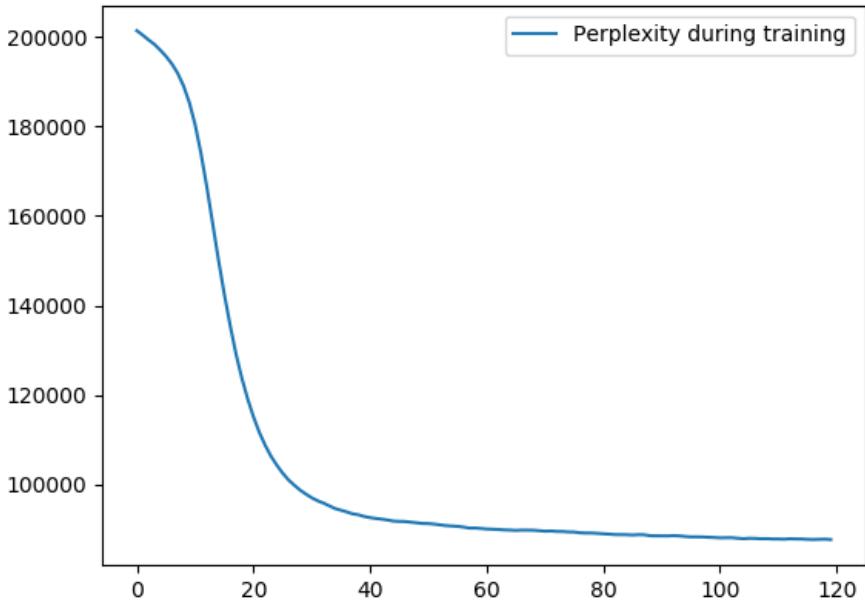
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++) پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، Perplexity حداکثر تعداد ایپاک‌ها، زمان انجام کلیهی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$ ،  $\theta$ ، میزان در هر ایپاک است.

در فایل `model-dataset2.json`، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل `topic-topics.txt`:

کلمه‌ها را از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتدا قرار می‌دهیم. به این ترتیب عنوان‌ها در فایل `topics.txt` قابل مشاهده‌اند. (برای نمایش بهتر از یک `text-editor` خوب مانند `notepad++` استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۰۰ خروجی مدل هجدهم

	Perplexity در ایپاک‌های مختلف
۳۴۵۹,۲۷۷۱۲۵۸۳۵۴۱۸۷ s	زمان کل اجرای حداکثر ایپاک‌های مجاز
28.827309381961822 s	زمان میانگین برای انجام یک ایپاک

## 120 Epochs

ایپاکهای

لازم برای

رسیدم به

حالت

Mixing

۳۴۵۹.۲۷۷۱۲۵۸۳۵۴۱۸۷ s

زمان لازم

برای رسیدن

به حالت

mixing

| model-dataset2.json x |

```

1   {
2     "total_time": 3459.2771258354187,
3     "each_epoch_time": 28.827309381961822,
4     "W": 10473,
5     "T": 30,
6     "alpha": 1,
7     "beta": 1,
8     "dataset": 2,
9     "phi": [
10       [
11         6.196938712276135e-05,
12         0.0006196938712276135,
13         6.196938712276135e-05,
14         0.0002478775484910454,
15         6.196938712276135e-05,
16         6.196938712276135e-05,
17         0.0002478775484910454,
18         0.00018590816136828406,
19         6.196938712276135e-05,
20         6.196938712276135e-05,
21         6.196938712276135e-05,
22         0.0003718163227365681,
23         6.196938712276135e-05,
24         0.0002478775484910454,
25         6.196938712276135e-05,
26         6.196938712276135e-05,
27         0.00018590816136828406,
28         6.196938712276135e-05,
29         6.196938712276135e-05,
30         6.196938712276135e-05,
31         6.196938712276135e-05,
32         6.196938712276135e-05,
33         0.00030984693561380677,
```

فایل  
model-  
dataset1.json

در پوشه‌ی

Model-18-  
dataset-2

که شامل ،  $\theta$   
 $\Phi$  و سایر

خروجی‌ها  
است

فایل

topics.txt

در پوششی

Model-18-  
dataset-2

که شامل

عنوان‌ها به

صورت کلمه

است

```

1  ['israel', 'israeli', 'jewish', 'palestinian', 'arab', 'palestinians', 'plo', 'occupy',
2  ['dollar', 'late', 'yen', 'london', 'gold', 'bid', 'tokyo', 'ounce', 'british', 'do'],
3  ['school', 'students', 'university', 'student', 'schools', 'education', 'board', 'co'],
4  ['court', 'case', 'attorney', 'judge', 'trial', 'state', 'federal', 'charges', 'law'],
5  ['film', 'cbs', 'nbc', 'news', 'abc', 'show', 'movie', 'tv', 'network', 'series', 'o'],
6  ['cents', 'west', 'east', 'futures', 'german', 'cent', 'lower', 'higher', 'tons', 'o'],
7  ['smoking', 'meese', 'keating', 'trust', 'lincoln', 'thrift', 'regulators', 'deconc'],
8  ['party', 'soviet', 'government', 'political', 'gorbachev', 'communist', 'leader'],
9  ['united', 'states', 'soviet', 'president', 'government', 'foreign', 'trade', 'talk'],
10 ['health', 'aids', 'medical', 'drug', 'study', 'research', 'patients', 'disease', 'o'],
11 ['monastery', 'island', 'cereal', 'patmos', 'islands', 'st', 'bruce', 'boxes', 'rare'],
12 ['police', 'people', 'two', 'killed', 'government', 'city', 'army', 'three', 'offic'],
13 ['members', 'homeless', 'club', 'rural', 'pacs', 'contributions', 'andreas', 'clubs'],
14 ['air', 'defense', 'program', 'plane', 'tax', 'service', 'aircraft', 'report', 'flic'],
15 ['iraq', 'kuwait', 'iraqi', 'military', 'saudi', 'iran', 'gulf', 'war', 'force', 't'],
16 ['percent', 'year', 'million', 'market', 'billion', 'prices', 'new', 'last', 'stock'],
17 ['computer', 'computers', 'apple', 'hildreth', 'software', 'programs', 'number', 'bu'],
18 ['new', 'award', 'won', 'music', 'awards', 'best', 'york', 'year', 'band', 'art', 't'],
19 ['art', 'works', 'paintings', 'theft', 'thieves', 'stolen', 'museum', 'today', 'thre'],
20 ['bush', 'dukakis', 'campaign', 'president', 'democratic', 'republican', 'new', 'jac'],
21 ['company', 'million', 'new', 'inc', 'corp', 'billion', 'co', 'business', 'bank', 'o'],
22 ['workers', 'plant', 'union', 'contract', 'environmental', 'company', 'strike', 'lab'],
23 ['new', 'city', 'york', 'inheritance', 'games', 'william', 'john', 'real', 'calif'],
24 ['mrs', 'children', 'ms', 'hospital', 'family', 'yearold', 'mother', 'wife', 'child'],
25 ['house', 'bill', 'senate', 'committee', 'congress', 'rep', 'budget', 'sen', 'legis'],
26 ['editor', 'ap', 'monet', 'news', 'elephants', 'managing', 'gotti', 'daily', 'named'],
27 ['south', 'africa', 'black', 'african', 'mandela', 'de', 'apartheid', 'blacks', 'and'],
28 ['i', 'people', 'years', 'dont', 'time', 'like', 'think', 'get', 'just', 'two', 'go'],
29 ['nauvoo', 'church', 'toussaint', 'restored', 'mormon', 'management', 'haiti', 'grav'],
30 ['fire', 'miles', 'water', 'area', 'southern', 'coast', 'state', 'north', 'official']

```

۸۷۶۸۹,۵۰۶۵۵۰۰۷۵۶۶

Perplexity

نهایی

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

## ۴.۲.۵ مدل نوزدهم

پارامترهای مختلف مدل:

جدول ۱: پارامترهای مدل نوزدهم

۱	$\alpha$
۱	$\beta$
۱۰۴۷۳	W (تعداد کلمه‌ها در دیکشنری)
۵۰	T (تعداد عنوان‌ها)
۱۲۰	Max Epoch (حداکثر تعداد ایپاک)
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-19-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش: **Figure\_1.png**

یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، Perplexity حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$ ،  $\theta$ ، میزان در هر ایپاک است.

در فایل **model-dataset2.json**، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل **topic-topics.txt**، کلمه‌ها را از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتداء قرار می‌دهیم. به این ترتیب عنوان‌ها در فایل **topics.txt** قابل مشاهده اند. (برای نمایش بهتر از یک text-editor خوب مانند notepad++ استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۲ خروجی مدل نوزدهم

<p>Perplexity during training</p> <table border="1"> <thead> <tr> <th>Epoch</th> <th>Perplexity</th> </tr> </thead> <tbody> <tr><td>0</td><td>~225,000</td></tr> <tr><td>10</td><td>~180,000</td></tr> <tr><td>20</td><td>~140,000</td></tr> <tr><td>30</td><td>~115,000</td></tr> <tr><td>40</td><td>~105,000</td></tr> <tr><td>60</td><td>~100,000</td></tr> <tr><td>80</td><td>~98,000</td></tr> <tr><td>100</td><td>~97,000</td></tr> <tr><td>120</td><td>~96,000</td></tr> </tbody> </table>	Epoch	Perplexity	0	~225,000	10	~180,000	20	~140,000	30	~115,000	40	~105,000	60	~100,000	80	~98,000	100	~97,000	120	~96,000	Perplexity در ایپاک‌های مختلف
Epoch	Perplexity																				
0	~225,000																				
10	~180,000																				
20	~140,000																				
30	~115,000																				
40	~105,000																				
60	~100,000																				
80	~98,000																				
100	~97,000																				
120	~96,000																				
۴۹۷۲,۲۲۹۵۳۷۰۱۰۱۹۳ s	زمان کل اجرای حداکثر ایپاک‌های مجاز																				
41.435246141751605 s	زمان میانگین برای انجام یک ایپاک																				
120 Epochs	ایپاک‌های لازم برای رسیدم به حالت Mixing																				

۴۹۷۲,۲۲۹۵۳۷۰۱۰۱۹۳۵

زمان لازم  
برای رسیدن  
به حالت  
mixing

فایل	model-dataset1.json	در پوششی	Model-19-dataset-2	که شامل ، $\theta$ و سایر $\Phi$	خروجی‌ها	است
model-dataset2.json	{ "total_time": 4972.229537010193, "each_epoch_time": 41.435246141751605, "W": 10473, "T": 50, "alpha": 1, "beta": 1, "dataset": 2, "phi": [ [ 0.000147693034796479, 0.008625273232114374, 0.001181544278371832, 0.0001181544278371832, 0.0032197081585632423, 0.0021563183080285935, 0.012819755420334377, 0.004342175223016482, 0.0037809416907898623, 0.0001181544278371832, 0.0012996987062090153, 0.003396939800319017, 2.95386069592958e-05, 0.0003249246765522538, 0.000147693034796479, 0.0011520056714125362, 2.95386069592958e-05, 0.0010929284574939446, 2.95386069592958e-05, 0.000886158208778874, 0.002333549949784368, ... ] ] }					

فایل

topics.txt

در پوششی

Model-19-  
dataset-2

که شامل

عنوان‌ها به

صورت کلمه

است

```

1 ['company', 'million', 'new', 'inc', 'corp', 'co', 'billion', 'bank', 'business', '']
2 ['police', 'people', 'two', 'killed', 'officials', 'city', 'three', 'fire', 'miles']
3 ['smoking', 'cigarettes', 'tobacco', 'cigarette', 'smoke', 'reynolds', 'trains', 'm']
4 ['fatal', 'academy', 'moonstruck', 'nominees', 'nominated', 'hurt', 'broadcast', 'e']
5 ['recent', 'program', 'degree', 'turn', 'john', 'france', 'version', 'spotted', 're']
6 ['united', 'military', 'states', 'war', 'government', 'iraq', 'troops', 'president']
7 ['eastern', 'pilots', 'orion', 'easterns', 'airline', 'travel', 'chechchi']
8 ['removed', 'la', 'wednesday', 'va', 'tour', 'surprised', 'significant', 'request', ]
9 ['groups', 'th', 'year', 'works', 'walters', 'united', 'treatment', 'traveling', 't']
10 ['dresses', 'dress', 'look', 'bridal', 'wedding', 'style', 'recently', 'bridesmaids']
11 ['wynberg', 'today', 'rest', 'remaining', 'mark', 'king', 'high', 'condition', 'won'
12 ['week', 'signal', 'saturday', 'motion', 'million', 'washington', 'sunday', 'states']
13 ['list', 'inmates', 'armory', 'voice', 'reports', 'pressure', 'permit', 'people', '']
14 ['general', 'winning', 'watched', 'placing', 'official', 'missiles', 'means', 'jimm
15 ['like', 'give', 'gave', 'threat', 'third', 'terry', 'taking', 'seven', 'sen', 'say
16 ['jackpot', 'prize', 'lottery', 'drawing', 'outside', 'numbers', 'wednesday', 'keep
17 ['waste', 'new', 'work', 'city', 'john', 'inheritance', 'william', 'california', 'b
18 ['available', 'park', 'game', 'disney', 'number', 'kasparov', 'karpov', 'creek', 'v
19 ['give', 'wage', 'thought', 'simon', 'showing', 'round', 'right', 'richard', 'retir
20 ['air', 'plane', 'flight', 'aircraft', 'navy', 'airlines', 'airport', 'ship', 'plan
21 ['market', 'dollar', 'stock', 'late', 'trading', 'rose', 'index', 'yen', 'prices',
22 ['trust', 'claims', 'manville', 'asbestos', 'victims', 'payments', 'weinstein', 'se
23 ['workers', 'union', 'contract', 'strike', 'labor', 'plant', 'chrysler', 'unions',
24 ['space', 'shuttle', 'launch', 'nasa', 'mission', 'earth', 'venus', 'two', 'spacecr
25 ['wine', 'care', 'b', 'bcspehealth', 'written', 'produce', 'wines', 'ny', 'british'
26 ['bush', 'president', 'house', 'dukakis', 'campaign', 'senate', 'bill', 'i', 'new',
27 ['yearold', 'total', 'present', 'expected', 'time', 'texas', 'site', 'second', 'res
28 ['court', 'case', 'attorney', 'judge', 'federal', 'trial', 'state', 'charges', 'law
29 ['water', 'environmental', 'plant', 'state', 'species', 'river', 'epa', 'department
30 ['israel', 'israeli', 'jewish', 'arab', 'palestinian', 'peace', 'palestinians', 'ba
31 ['weapons', 'department', 'team', 'nuclear', 'energy', 'tiger', 'work', 'plants', '
32 ['soviet', 'party', 'government', 'union', 'president', 'gorbachev', 'minister', 'u
33 ['editor', 'symphony', 'ap', 'orchestra', 'musicians', 'managing', 'francisco', 'fl
34 ['state', 'paper', 'memorial', 'firms', 'backing', 'trying', 'september', 'role', '
35 ['southern', 'high', 'northern', 'rain', 'fair', 'central', 'coast', 'degrees', 'in
36 ['art', 'museum', 'sothebys', 'monet', 'works', 'vase', 'short', 'paintings', 'suit
37 ['south', 'students', 'africa', 'government', 'black', 'african', 'rights', 'presid
38 ['news', 'network', 'cbs', 'nbc', 'abc', 'rating', 'television', 'tv', 'week', 'rat
39 ['i', 'people', 'years', 'dont', 'like', 'new', 'time', 'just', 'get', 'two', 'thin
40 ['abortion', 'souter', 'medication', 'abortions', 'award', 'perrys', 'miners', 'ida
41 ['health', 'children', 'aids', 'medical', 'hospital', 'disease', 'patients', 'drug'
42 ['states', 'work', 'newspaper', 'execution', 'delegation', 'try', 'test', 'technica
43 ['cents', 'oil', 'futures', 'lower', 'cent', 'higher', 'prices', 'farmers', 'tons',
44 ['lincoln', 'keating', 'deconcini', 'senators', 'regulators', 'thrift', 'meeting',
45 ['director', 'theres', 'station', 'service', 'row', 'raised', 'prepared', 'plastic'
46 ['form', 'stolen', 'states', 'range', 'capitol', 'banned', 'work', 'think', 'subcom
47 ['tip', 'sept', 'co', 'centers', 'carrying', 'back', 'arts', 'americans', 'writing'
48 ['percent', 'year', 'million', 'billion', 'last', 'report', 'new', 'sales', 'govern
49 ['wednesday', 'miller', 'correct', 'winter', 'show', 'season', 'relief', 'relations
50 ['west', 'east', 'german', 'germany', 'berlin', 'germanys', 'germans', 'unification
51

```

۹۹۰۴۰,۶۳۷۳۱۱۰۵۲۷۲

Perplexity  
نهایی

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

## ۴.۲.۶ مدل بیستم

پارامترهای مختلف مدل:

جدول ۴ پارامترهای مدل بیستم

جدول ۴ پارامترهای مدل بیستم	
۱	$\alpha$
۱	$\beta$
۱۰۴۷۳	W(تعداد کلمه‌ها در دیکشنری)
۷۰	T تعداد عنوان‌ها
۱۲۰	Max Epoch حداکثر تعداد ایپاک
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-20-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش **Figure\_1.png**

یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان Perplexity در هر ایپاک است.

در فایل **model-dataset2.json**، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل **topic-**topics.txt، کلمه‌ها را از دیکشنری بر می‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتدا قرار می‌دهیم. به این ترتیب عنوان‌ها در فایل topics.txt قابل مشاهده اند. (برای نمایش بهتر از یک text-editor خوب مانند notepad++ استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۴ خروجی مدل بیستم

	<p>Perplexity during training</p> <table border="1"> <thead> <tr> <th>Epoch</th> <th>Perplexity</th> </tr> </thead> <tbody> <tr><td>0</td><td>~245,000</td></tr> <tr><td>10</td><td>~235,000</td></tr> <tr><td>20</td><td>~180,000</td></tr> <tr><td>30</td><td>~145,000</td></tr> <tr><td>40</td><td>~125,000</td></tr> <tr><td>50</td><td>~118,000</td></tr> <tr><td>60</td><td>~115,000</td></tr> <tr><td>70</td><td>~114,000</td></tr> <tr><td>80</td><td>~113,000</td></tr> <tr><td>90</td><td>~112,000</td></tr> <tr><td>100</td><td>~111,000</td></tr> <tr><td>110</td><td>~110,000</td></tr> <tr><td>120</td><td>~110,000</td></tr> </tbody> </table>	Epoch	Perplexity	0	~245,000	10	~235,000	20	~180,000	30	~145,000	40	~125,000	50	~118,000	60	~115,000	70	~114,000	80	~113,000	90	~112,000	100	~111,000	110	~110,000	120	~110,000	Perplexity در ایپاک‌های مختلف
Epoch	Perplexity																													
0	~245,000																													
10	~235,000																													
20	~180,000																													
30	~145,000																													
40	~125,000																													
50	~118,000																													
60	~115,000																													
70	~114,000																													
80	~113,000																													
90	~112,000																													
100	~111,000																													
110	~110,000																													
120	~110,000																													
6566,578573226929 s	زمان کل اجرای حداکثر ایپاک‌های مجاز																													
54.721488110224406 s	زمان میانگین برای انجام یک ایپاک																													
120 Epochs	ایپاک‌های لازم برای رسیدم به حالت Mixing																													

۶۵۶۶.۵۷۸۵۷۳۲۲۶۹۲۹ s

زمان لازم  
برای رسیدن  
به حالت  
mixing

```
model-dataset2.json ✘ |
```

```

1   {
2     "total_time": 6566.578573226929,
3     "each_epoch_time": 54.721488110224406,
4     "W": 10473,
5     "T": 70,
6     "alpha": 1,
7     "beta": 1,
8     "dataset": 2,
9     "phi": [
10       [
11         8.205464839583163e-05,
12         8.205464839583163e-05,
13         8.205464839583163e-05,
14         8.205464839583163e-05,
15         0.0002461639451874949,
16         8.205464839583163e-05,
17         0.00016410929679166325,
18         0.00016410929679166325,
19         8.205464839583163e-05,
20         0.0002461639451874949,
21         0.00016410929679166325,
22         0.00016410929679166325,
23         0.00016410929679166325,
24         8.205464839583163e-05,
25         8.205464839583163e-05,
26         8.205464839583163e-05,
27         8.205464839583163e-05,
28         8.205464839583163e-05,
29         8.205464839583163e-05,
30         0.00016410929679166325,
31         0.00016410929679166325,
32         8.205464839583163e-05,
33         0.00016410929679166325,
34         0.00016410929679166325,
35         8.205464839583163e-05,
36         8.205464839583163e-05,
```

فایل  
model-  
dataset1.json  
در پوششی  
Model-20-  
dataset-2  
که شامل ،  $\theta$  ،  
 $\Phi$  و سایر  
خروجی‌ها  
است

فایل

topics.txt

در پوششی

Model-20-  
dataset-2

که شامل

عنوان‌ها به

صورت کلمه

است

```
1 ['added', 'september', 'extended', 'widely', 'transport', 'transfer', 'tickets', 'single',  
2 ['guilty', 'sense', 'trump', 'share', 'says', 'roger', 'restraint', 'remained', 'oct', 'mil  
3 ['court', 'case', 'attorney', 'judge', 'trial', 'state', 'federal', 'charges', 'prison', 'l  
4 ['disney', 'environmental', 'mca', 'yosemite', 'park', 'paper', 'television', 'recycling',  
5 ['money', 'family', 'experts', 'emergency', 'bob', 'weeks', 'wants', 'vice', 'special', 'sc  
6 ['work', 'wife', 'stable', 'services', 'saying', 'protested', 'preceded', 'possibility', 'n  
7 ['front', 'telephone', 'refugees', 'morning', 'four', 'eye', 'won', 'wing', 'voted', 'tired  
8 ['police', 'people', 'two', 'killed', 'officials', 'fire', 'miles', 'city', 'three', 'autho  
9 ['worlds', 'top', 'range', 'raise', 'outcome', 'unspecified', 'title', 'things', 'term', 't  
10 ['network', 'cbs', 'nbc', 'news', 'abc', 'rating', 'week', 'television', 'cable', 'ratings'  
11 ['government', 'party', 'political', 'soviet', 'president', 'communist', 'opposition', 'pec  
12 ['space', 'launch', 'shuttle', 'nasa', 'soviet', 'mission', 'mars', 'test', 'telescope', 'a  
13 ['cents', 'oil', 'futures', 'cent', 'lower', 'higher', 'corn', 'prices', 'tons', 'bushel',  
14 ['company', 'new', 'million', 'bank', 'late', 'inc', 'york', 'dollar', 'corp', 'stock', 'cc  
15 ['school', 'students', 'student', 'schools', 'education', 'university', 'board', 'college']  
16 ['house', 'bill', 'committee', 'congress', 'senate', 'budget', 'federal', 'rep', 'members',  
17 ['health', 'children', 'aids', 'medical', 'hospital', 'drug', 'disease', 'patients', 'dr',  
18 ['won', 'los', 'july', 'attributed', 'amid', 'wide', 'voted', 'visit', 'time', 'texas', 'sa  
19 ['take', 'martin', 'dec', 'wilson', 'standards', 'shortly', 'returned', 'responded', 'range  
20 ['reasons', 'working', 'saturday', 'board', 'asked', 'threatening', 'tensions', 'special',  
21 ['year', 'rule', 'dog', 'company', 'added', 'thousands', 'saying', 'robert', 'race', 'prais  
22 ['reported', 'tuesday', 'side', 'open', 'meeting', 'last', 'international', 'hotel', 'won',  
23 ['went', 'public', 'purchased', 'profits', 'columbia', 'yearold', 'virtually', 'valley', 't  
24 ['opened', 'return', 'quality', 'positive', 'poor', 'live', 'gov', 'friday', 'fed', 'events  
25 ['military', 'united', 'war', 'iraq', 'states', 'troops', 'force', 'president', 'american',  
26 ['market', 'stock', 'index', 'trading', 'rose', 'points', 'exchange', 'prices', 'stocks', '  
27 ['washington', 'standard', 'sources', 'little', 'friday', 'earlier', 'considered', 'atlanti  
28 ['charge', 'prime', 'impact', 'forced', 'university', 'true', 'terms', 'states', 'sept', 's  
29 ['tuesday', 'widely', 'thompson', 'taste', 'take', 'summer', 'state', 'star', 'space', 'prc  
30 ['leader', 'time', 'brought', 'age', 'two', 'trying', 'studied', 'states', 'sons', 'schedul  
31 ['percent', 'year', 'million', 'billion', 'last']  
32 ['influence', 'development', 'work', 'subject', 'separate', 'rules', 'population', 'lt', 'k  
33 ['issues', 'issued', 'cigarettes', 'anonymity', 'seen', 'place', 'pictures', 'museveni', 'n  
34 ['art', 'thieves', 'works', 'theft', 'stolen', 'experts', 'worth', 'paintings', 'gardner',  
35 ['north', 'korean', 'korea', 'south', 'games', 'roh', 'koreas']  
36 ['members', 'lead', 'commission', 'applications', 'western', 'survey', 'supply', 'spokeswom  
37 ['disaster', 'county', 'worth', 'washington', 'university', 'trump', 'true', 'transit', 'tc  
38 ['thomas', 'program', 'present', 'pledged', 'p', 'note', 'free', 'take', 'station', 'spent'  
39 ['soviet', 'west', 'east', 'german', 'germany', 'union', 'israel', 'united', 'europe', 'isr  
40 ['fair', 'southern', 'rain', 'new', 'high', 'northern', 'inches', 'coast', 'snow', 'degrees  
41 ['room', 'question', 'years', 'wood', 'spokesman', 'seat', 'saturday', 'review', 'possibly'  
42 ['united', 'states', 'south', 'trade', 'africa', 'japan', 'countries', 'african', 'japanese', 'agreement  
43 ['north', 'meese', 'reagan', 'walsh', 'norts', 'irancontra', 'gesell', 'poindexter', 'arms', 'testimony  
44 ['small', 'necessary', 'write', 'southern', 'ronald', 'oct', 'obscene', 'matters', 'management', 'hundre  
45 ['dozen', 'offered', 'difficult', 'won', 'third', 'theres', 'suffered', 'session', 'second', 'permission  
46 ['result', 'count', 'century', 'today', 'times', 'takes', 'ruling', 'reason', 'memories', 'march', 'incr  
47 ['williams', 'led', 'written', 'speech', 'second', 'records', 'provides', 'order', 'asked', 'weve', 'wes  
48 ['officer', 'view', 'six', 'place', 'david', 'week', 'urged', 'tuesday', 'testament', 'subject', 'strong  
49 ['wont', 'supervision', 'get', 'denied', 'victory', 'trip', 'think', 'testified', 'resulted', 'plane', '  
50 ['washington', 'last', 'blue', 'train', 'technical', 'st', 'soon', 'showed', 'says', 'responsibility', '  
51 ['bush', 'president', 'dukakis', 'campaign', 'i', 'democratic', 'republican', 'jackson', 'reagan', 'pres  
52 ['vargas', 'fujimori', 'llosa', 'perus', 'sunday', 'second', 'put', 'person', 'light', 'expected', 'cont  
53 ['dresses', 'dress', 'bridal', 'style', 'leonard', 'bridesmaids', 'wedding', 'fabrics', 'couture', 'cost  
54 ['m', 'cdy', 'clr', 'rn']  
55 ['ap', 'editor', 'thompson', 'managing', 'joined', 'bureau', 'soon', 'radio', 'payments', 'named', 'mear  
56 ['workers', 'air', 'union', 'plant', 'contract', 'company', 'airlines', 'strike', 'employees', 'safety',  
57 ['water', 'species', 'inheritance', 'city', 'wildlife', 'fish', 'lake', 'endangered', 'william', 'turkey  
58 ['venus', 'earth', 'spacecraft', 'magellan', 'radar', 'pictures', 'mission', 'galileo', 'planet', 'conta  
59 ['b', 'care', 'bcspehealth']  
60 ['three', 'two', 'seven', 'ran', 'good', 'came', 'valdez', 'told', 'th', 'tennessee', 'island', 'hours',  
61 ['western', 'suffered', 'removed', 'monday', 'head', 'growing', 'turkey', 'todays', 'thought', 'made', '  
62 ['week', 'reach', 'plenty', 'organized', 'old', 'field', 'family', 'academy', 'women', 'troops', 'teenag  
63 ['i', 'years', 'people', 'new', 'like', 'dont', 'time', 'just', 'get']  
64 ['zubal', 'started', 'returned', 'judgment', 'warfare', 'traditional', 'session', 'see', 'searched', 'kn  
65 ['creek', 'yellow', 'tannery', 'water', 'sewage', 'wilson', 'residents', 'national', 'concerned', 'chrom  
66 ['estate', 'property', 'ruby', 'assets', 'federal', 'value', 'tax', 'life', 'deduction', 'mayer', 'amoun  
67 ['provides', 'national', 'issue', 'group', 'attempt', 'yearold', 'war', 'unit', 'twothirds', 'turner', '  
68 ['week', 'range', 'mrs', 'believed', 'trouble', 'thought', 'stations', 'six', 'seven', 'press', 'opportu  
69 ['spokesman', 'funds', 'world', 'stable', 'race', 'president', 'placed', 'known', 'july', 'executive', '  
70 ['tuesday', 'money', 'mail', 'identify', 'hand', 'groups', 'blamed', 'william', 'two', 'telling', 'stage  
71 ]
```

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

#### ۴.۲.۷ مدل بیست و یکم

پارامترهای مختلف مدل:

جدول ۴ پارامترهای مدل بیست و یکم

۱	$\alpha$
۱	$\beta$
۱۰۴۷۳	W (تعداد کلمه‌ها در دیکشنری)
۱۰۰	T (تعداد عنوان‌ها)
۱۲۰	Max Epoch (حداکثر تعداد ایپاک)
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیهی خروجی‌های حاصل از اجرای کد برای مدل یک در پوشی **model-21-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش: **Figure\_1.png**

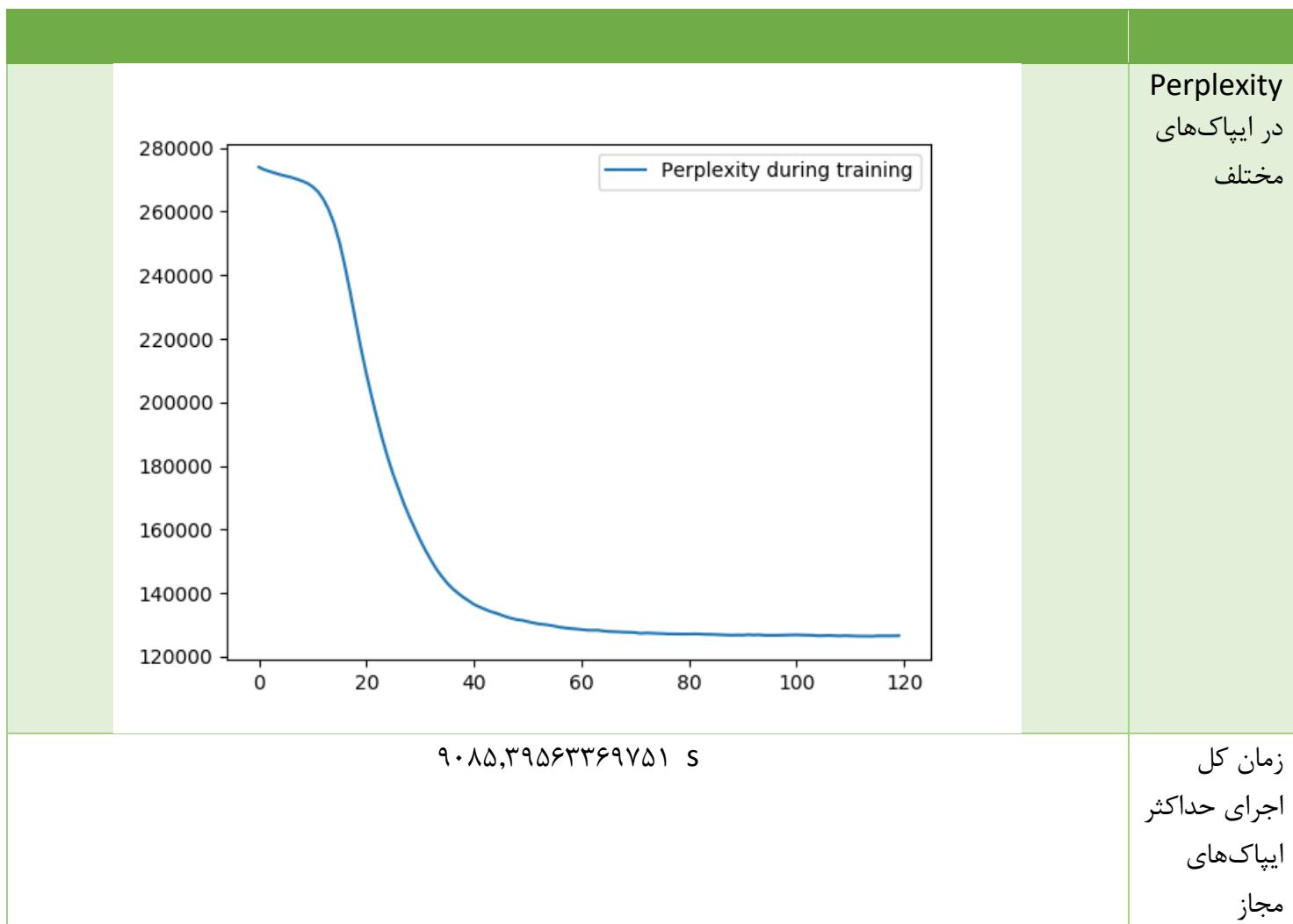
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++) پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، Perplexity حداکثر تعداد ایپاک‌ها، زمان انجام کلیهی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$ ،  $\theta$ ، میزان در هر ایپاک است.

در فایل `model-dataset2.json`، مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل `topic-.topics.txt`

کلمه‌ها را از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتدا قرار می‌دهیم. به این ترتیب عنوان‌ها در فایل `topics.txt` قابل مشاهده‌اند. (برای نمایش بهتر از یک خوب مانند `notepad++` استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۶۴ خروجی مدل بیست و یکم



75.71163028081259 s	زمان میانگین برای انجام یک ایپاک
120 Epochs	ایپاکهای لازم برای رسیدم به حالت <b>Mixing</b>
۹۰,۸۵,۳۹۵۶۳۳۶۹۷۵۱ s	زمان لازم برای رسیدن به حالت <b>mixing</b>

```

1   "total_time": 9085.39563369751,
2   "each_epoch_time": 75.71163028081259,
3   "W": 10473,
4   "T": 100,
5   "alpha": 1,
6   "beta": 1,
7   "dataset": 2,
8   "phi": [
9     [
10       8.401243384020835e-05,
11       8.401243384020835e-05,
12       8.401243384020835e-05,
13       0.0001680248676804167,
14       0.0001680248676804167,
15       8.401243384020835e-05,
16       8.401243384020835e-05,
17       8.401243384020835e-05,
18       8.401243384020835e-05,
19       8.401243384020835e-05,
20       8.401243384020835e-05,
21       0.0001680248676804167,
22       0.0001680248676804167,
23       8.401243384020835e-05,
24       8.401243384020835e-05,
25       8.401243384020835e-05,
26       0.0001680248676804167,
27       8.401243384020835e-05,
28       8.401243384020835e-05,
29       8.401243384020835e-05,
30       0.0002520373015206251,
31       0.0002520373015206251,
32       8.401243384020835e-05,
33       8.401243384020835e-05,

```

فایل  
model-dataset1.json  
در پوشه Model-21-dataset-2  
که شامل  $\theta$  و سایر  $\Phi$  خروجی‌ها است

فایل  
topics.txt  
در پوشه Model-21-dataset-2  
که شامل عنوان‌ها به صورت کلمه است

```
topics.txt
1 ['made', 'remains', 'manager', 'chicago', 'wednesday', 'ways', 'victims', 'su
2 ['southern', 'fair', 'rain', 'high', 'northern', 'central', 'texas', 'inches'
3 ['show', 'december', 'waste', 'motor', 'chief', 'britain', 'won', 'widespread
4 ['rear', 'made', 'worth', 'wife', 'white', 'wave', 'try', 'saturday', 'return
5 ['headed', 'face', 'charge', 'ability', 'week', 'threatened', 'teacher', 'sou
6 ['cited', 'involving', 'afternoon', 'working', 'tried', 'track', 'threat', 't
7 ['south', 'africa', 'african', 'black', 'north', 'korea', 'mandela', 'korean'
8 ['thursday', 'keep', 'scheduled', 'shouldnt', 'reflecting', 'recording', 'rec
9 ['students', 'school', 'student', 'schools', 'education', 'teachers', 'board'
10 ['defense', 'addition', 'early', 'win', 'union', 'talked', 'somebody', 'revea
11 ['miles', 'lines', 'system', 'smog', 'sea', 'says', 'prime', 'present', 'open
12 ['smoking', 'trains', 'stations', 'smokers', 'municipal', 'martin', 'malls',
13 ['west', 'east', 'german', 'israel', 'germany', 'israeli', 'jewish', 'palesti
14 ['transportation', 'taking', 'statement', 'speed', 'severe', 'seat', 'ryan',
15 ['employees', 'victims', 'two', 'successful', 'spokeswoman', 'severe', 'rober
16 ['public', 'known', 'engineering', 'concern', 'villages', 'seats', 'resources
17 ['department', 'saturday', 'made', 'written', 'world', 'warning', 'suburban',
18 ['today', 'soon', 'panel', 'exxon', 'commission', 'businesses', 'april', 'age
19 ['effect', 'side', 'small', 'radio', 'people', 'form', 'fled', 'business', 'y
20 ['security', 'prove', 'john', 'found', 'different', 'urban', 'son', 'prior',
21 ['m', 'cdy', 'clr', 'close', 'raised', 'rn', 'list', 'j', 'wearing', 'truly',
22 ['oil', 'cents', 'futures', 'cent', 'lower', 'higher', 'prices', 'crude', 'bu
23 ['united', 'states', 'president', 'military', 'soviet', 'american', 'foreign'
24 ['win', 'million', 'warnings', 'states', 'soft', 'north', 'lose', 'long', 'in
25 ['scheduled', 'apparently', 'william', 'uss', 'surprised', 'sitting', 'series
26 ['gain', 'urged', 'unable', 'starting', 'shell', 'receive', 'prevented', 'por
27 ['york', 'th', 'coalition', 'active', 'wanted', 'stations', 'shopping', 'sept
28 ['keating', 'lincoln', 'deconcini', 'senators', 'regulators', 'meeting', 'gra
29 ['supporters', 'safe', 'activity', 'yearold', 'wrote', 'wars', 'true', 'test'
30 ['food', 'art', 'issue', 'th', 'first', 'council', 'case', 'walking', 'system
31 ['dec', 'identified', 'discuss', 'consider', 'wear', 'time', 'taking', 'story
32 ['republic', 'jim', 'foreign', 'construction', 'california', 'worth', 'wednes
33 ['improve', 'continue', 'expect', 'determine', 'wednesday', 'services', 'sele
34 ['results', 'prime', 'ended', 'works', 'three', 'thing', 'romania', 'returned
35 ['soviet', 'party', 'government', 'political', 'gorbachev', 'union', 'preside
36 ['free', 'current', 'share', 'seven', 'groups', 'attempt', 'transport', 'take
37 ['normal', 'failed', 'expected', 'am', 'worked', 'william', 'violations', 'th
38 ['houses', 'four', 'creating', 'usual', 'ties', 'three', 'sunday', 'stations'
39 ['sunday', 'dispute', 'total', 'stabilize', 'seven', 'russell', 'reporters',
40 ['property', 'ago', 'weekend', 'turned', 'survey', 'state', 'september', 'pre
41 ['site', 'going', 'regulations', 'programs', 'organization', 'word', 'specifi
```

```

topics.txt
42 ['care', 'b', 'bcspehealth', 'issued', 'training', 'ny', 'found', 'am', 'sta
43 ['louis', 'based', 'angeles', 'taking', 'sides', 'service', 'sat', 'remains',
44 ['steps', 'received', 'previously', 'april', 'workers', 'treatment', 'tom',
45 ['time', 'service', 'give', 'get', 'wages', 'thursday', 'three', 'severe', 'i
46 ['immediate', 'learned', 'royal', 'problem', 'peter', 'paper', 'national', 'i
47 ['water', 'species', 'epa', 'wildlife', 'turkey', 'endangered', 'lake', 'plan
48 ['agency', 'years', 'expect', 'estimated', 'think', 'six', 'signs', 'sent',
49 ['new', 'present', 'worked', 'went', 'wednesday', 'sound', 'seven', 'rock',
50 ['wine', 'brought', 'written', 'running', 'red', 'produce', 'measure', 'going
51 ['letter', 'assured', 'war', 'rules', 'remain', 'recommended', 'ready', 'prev
52 ['rights', 'going', 'american', 'introduced', 'ill', 'help', 'governments',
53 ['health', 'medical', 'children', 'hospital', 'disease', 'drug', 'dr', 'patie
54 ['bush', 'dukakis', 'house', 'president', 'campaign', 'senate', 'bill', 'repu
55 ['serious', 'york', 'works', 'system', 'state', 'single', 'lines', 'washingto
56 ['director', 'airline', 'wont', 'told', 'threat', 'theres', 'stop', 'sold',
57 ['wife', 'west', 'systems', 'spend', 'sources', 'report', 'remain', 'regulat
58 ['treasury', 'thursday', 'stopped', 'show', 'selling', 'road', 'remarks', 're
59 ['space', 'shuttle', 'nasa', 'launch', 'mission', 'earth', 'venus', 'test',
60 ['effect', 'cut', 'served', 'produced', 'possible', 'picture', 'participation
61 ['unit', 'time', 'takes', 'saying', 'center', 'right', 'reports', 'red', 'pow
62 ['show', 'revealed', 'jan', 'four', 'wednesday', 'training', 'states', 'sons
63 ['company', 'million', 'new', 'inc', 'corp', 'president', 'co', 'workers', 'b
64 ['spokeswoman', 'heavy', 'wednesday', 'unknown', 'two', 'supply', 'public',
65 ['issued', 'working', 'summer', 'records', 'hasnt', 'el', 'basically', 'week
66 ['six', 'three', 'top', 'replace', 'rd', 'owned', 'month', 'met', 'location',
67 ['times', 'mean', 'like', 'found', 'april', 'agency', 'win', 'war', 'surround
68 ['expressed', 'america', 'train', 'simon', 'responsible', 'reached', 'promise
69 ['sept', 'weeks', 'unspecified', 'tree', 'travel', 'suspended', 'stolen', 'st
70 ['put', 'driving', 'state', 'spokesman', 'seven', 'previous', 'pass', 'normal
71 ['letters', 'residence', 'past', 'oil', 'occurred', 'lines', 'jr', 'individual
72 ['three', 'program', 'university', 'second', 'represents', 'problems', 'perso
73 ['court', 'case', 'attorney', 'judge', 'federal', 'trial', 'law', 'state', 'c
74 ['aids', 'virus', 'blood', 'infected', 'number', 'hudson', 'users', 'spread'
75 ['day', 'britain', 'state', 'return', 'months', 'entrance', 'world', 'stopped
76 ['market', 'stock', 'dollar', 'late', 'trading', 'rose', 'index', 'new', 'yea
77 ['three', 'worth', 'joined', 'causing', 'back', 'tower', 'super', 'required',
78 ['version', 'television', 'aquino', 'transport', 'single', 'reverse', 'receiv
79 ['i', 'years', 'people', 'time', 'new', 'just', 'dont', 'like', 'first', 'get
80 ['recent', 'men', 'june', 'issued', 'indication', 'days', 'view', 'two', 'true
81 ['proposed', 'income', 'writing', 'weeks', 'safety', 'route', 'refused', 'rea
82 ['quickly', 'oct', 'months', 'level', 'wednesday', 'united', 'survey', 'retur
83 ['tv', 'december', 'written', 'workers', 'television', 'suffered', 'southwest', 'serious', 'secure
84 ['tuesday', 'die', 'agency', 'university', 'significant', 'showed', 'responsible', 'requirements',
85 ['women', 'watching', 'third', 'social', 'single', 'officers', 'miami', 'meet', 'hard', 'eventual
86 ['police', 'people', 'two', 'officials', 'killed', 'miles', 'fire', 'city', 'three', 'reported',
87 ['winning', 'operation', 'conducted', 'week', 'war', 'wanted', 'taken', 'smith', 'see', 'program',
88 ['trying', 'people', 'occurred', 'evening', 'british', 'violent', 'today', 'system', 'sure', 'supp
89 ['john', 'woman', 'chairman', 'workers', 'twenty', 'shopping', 'records', 'offered', 'ms', 'model
90 ['watching', 'ousted', 'member', 'body', 'wrong', 'turned', 'trying', 'treatment', 'spread', 'sou
91 ['employees', 'won', 'told', 'term', 'suffering', 'rules', 'new', 'man', 'local', 'failed', 'dome
92 ['skull', 'bones', 'yale', 'moderate', 'bonesmen', 'hours', 'aircraft', 'won', 'went', 'track', 'ti
93 ['million', 'agency', 'tv', 'tuesday', 'represented', 'pull', 'plans', 'pain', 'measures', 'leade
94 ['percent', 'year', 'million', 'billion', 'last', 'new', 'report', 'years', 'sales', 'government'
95 ['team', 'weapons', 'tiger', 'nuclear', 'site', 'oak', 'sites', 'ridge', 'department', 'contractor
96 ['statement', 'monday', 'material', 'wife', 'transportation', 'time', 'spent', 'situation', 'sell
97 ['age', 'weeks', 'total', 'step', 'provided', 'far', 'today', 'security', 'release', 'payments',
98 ['sunday', 'provided', 'program', 'gathered', 'day', 'warmus', 'two', 'th', 'special', 'smaller',
99 ['early', 'wing', 'turn', 'todays', 'step', 'star', 'shut', 'owners', 'live', 'informal', 'homes',
100 ['problems', 'faced', 'two', 'corps', 'back', 'able', 'york', 'women', 'widely', 'texas', 'sunday']

```

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در اپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

## ۴.۳ نتیجه‌گیری تغییر تعداد عنوان‌ها

در بخش ۱-۴ و ۲-۴ مدل‌های مختلفی با دیتاست‌های یک و دو ایجاد کردیم. در بخش ۱-۴ مدل‌های با دیتاست یک ایجاد کردیم که فقط در تعداد عنوان‌ها ( $T$ ) با یک دیگر متفاوت است. در بخش ۲-۴ هم مدل‌هایی با دیتاست دو ایجاد کردیم که فقط در تعداد عنوان‌ها با هم متفاوت است. در ادامه با مقایسه زمان اجرا و **Perplexity** بر حسب تعداد عنوان‌ها به بررسی تاثیر تغییر تعداد عنوان‌ها می‌پردازیم.

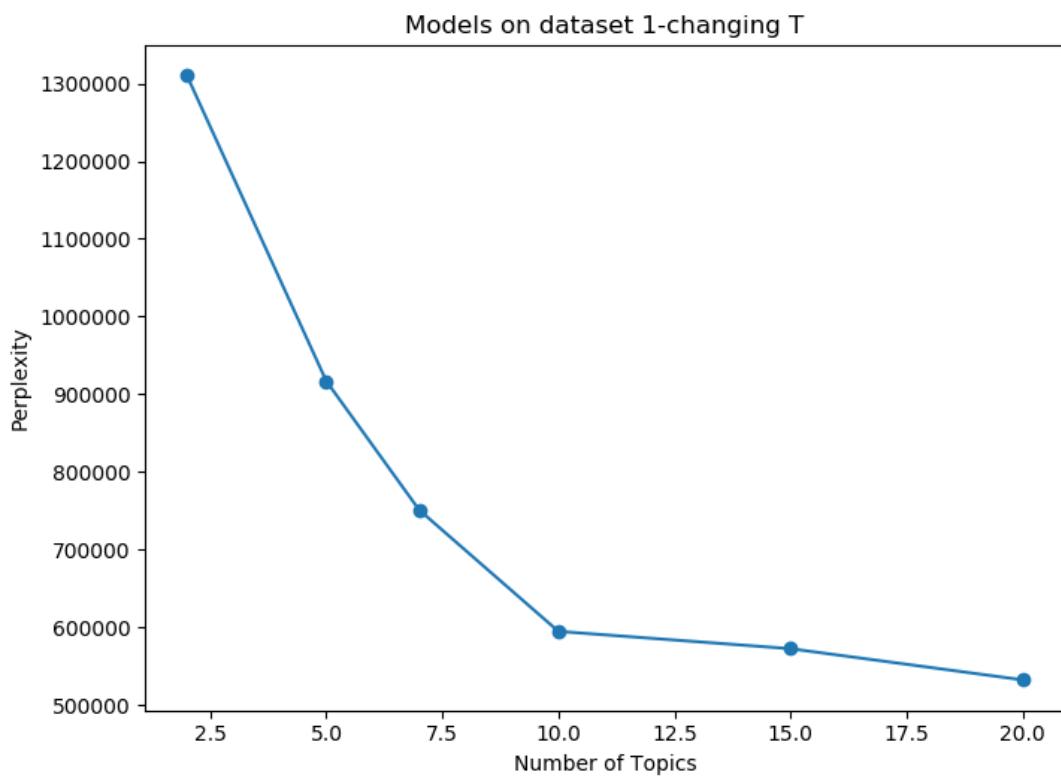
در بخش ۱-۴، مدل‌های شماره‌ی نه تا چهارده را با دیتاست یک ایجاد کردیم که فقط در تعداد عنوان باهم متفاوت‌اند، در جدول‌های ۲۱ تا ۳۲ می‌توان پارامترهای هر کدام از آن مدل‌ها و خروجی‌های مدل را مشاهده کرد.

زمان اجرا و **Perplexity** مدل‌های نه تا چهارده به این صورت است:

جدول ۷ مقایسه زمان اجرا و پرپلیسکی مدل‌های دیتاست یک با تغییر تعداد عنوان‌ها

مدل	تعداد عنوان ( $T$ )	زمان اجرای کل (S)	Perplexity نهایی
مدل نه	۲	۶۴۷,۶۷	۱۳۰,۹۹۸۳,۷۶
مدل ۵	۵	۶۵۱,۷۹	۹۱۶۹۱۰,۶۳
مدل یازده	۷	۷۲۵,۴۰	۷۵۱۰۳۳,۵۹
مدل دوازده	۱۰	۷۳۰,۷۵	۵۹۴۶۶۷,۱۰
مدل سیزده	۱۵	۸۰۵,۴۵	۵۷۲۲۳۵۲,۶۶
مدل چهارده	۲۰	۹۰۰,۲۸	۵۳۲۲۰۵,۱۱

همانطور که در جدول ۴۷ مشاهده می‌شود با افزایش تعداد عنوان‌ها زمان اجرایی زیاد شده است و **Perplexity** کم شده است، البته این نتیجه‌گیری نهایی نیست، در ادامه باید تغییرات را در مدل‌هایی که با دیتاست دو ایجاد شده‌اند را نیز بررسی کنیم و بعد نتیجه‌گیری کنیم.



در شکل بالا نمودار تغییر Perplexity بر حسب تغییر تعداد عنوان‌ها را در مدل‌های نه تا چهارده که با دیتاست یک ایجاده شده‌اند را مشاهده می‌کنید.

در بخش ۴-۲، مدل‌های شماره‌ی پانزده تا بیست و یکم را با دیتاست دو ایجاد کردیم که فقط در تعداد عنوان باهم متفاوت‌اند، در جدول‌های ۳۳ تا ۴۶ می‌توان پارامترهای هر کدام از آن مدل‌ها و خروجی‌های مدل‌ها را مشاهده کرد.

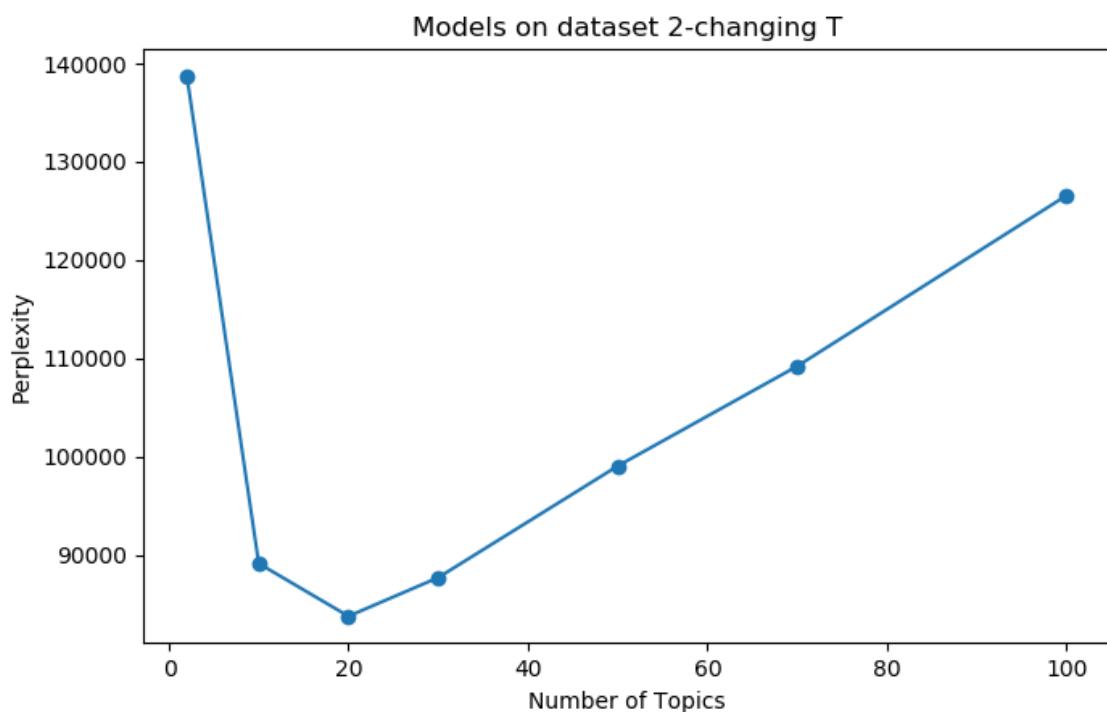
زمان اجرا و Perplexity مدل‌های شماره‌ی پانزده تا بیست و یکم به این صورت است:

جدول ۴ مقایسه‌ی زمان اجرا و پرپلیسکی مدل‌های دیتاست یک با تغییر تعداد عنوان‌ها

مدل	تعداد عنوان (T)	زمان اجرای کل (S)	Perplexity نهایی
مدل پانزده	۲	۱۱۰۹,۰۹	۱۳۸۶۷۲,۵۵
مدل شانزده	۱۰	۱۵۴۸,۴۰	۸۹۱۲۶,۶۲

۸۳۷۶۲,۸۶	۲۶۲۱,۲۶	۲۰	مدل هفده
۸۷۶۸۹,۵۰	۳۴۵۹,۲۷	۳۰	مدل هجده
۹۹۰۴۰,۶۳	۴۹۷۲,۲۲	۵۰	مدل نوزده
۱۰۹۱۹۰,۲۲	۶۵۶۶,۵۷	۷۰	مدل بیست
۱۲۶۵۴۸,۶۹	۹۰۸۵,۳۹	۱۰۰	مدل بیست و یکم

همانطور که در جدول ۴۸ مشاهده می‌شود با افزایش تعداد عنوان‌ها زمان اجرایی زیاد شده است ولی با زیاد شدن تعداد عنوان‌ها در ابتدا Perplexity کم شده است و در ادامه افزایش یافته است.



در شکل بالا نمودار تغییر Perplexity بر حسب تغییر تعداد عنوان‌ها را در مدل‌های پانزده تا بیست و یکم که با دیتاست دو ایجاده شده‌اند را مشاهده می‌کنید. در ابتدا پرپلسکی کاهش یافته است و بعد افزایش داشته است.

نتیجه گیری: با مقایسه‌هایی که در جدول ۴۷ و ۴۸ دو شکل بالا انجام دادیم به این دو نتیجه‌ی زیر می‌رسیم:

- با افزایش تعداد عنوان‌ها زمان اجرایی زیاد می‌شود.

- با زیاد شدن تعداد عنوان‌ها، Perplexity به دو صورت تغییر می‌کند، در حالت اول با زیاد شدن تعداد عنوان‌ها، پرپلیسکی کاهش می‌باید، در حالت دوم با زیاد شدن عنوان‌ها پرپلیسکی در ابتدا کاهش می‌باید و سپس افزایش می‌باید. **بر اساس تحقیقی که در اینترنت و بین مقاله‌های مختلف انجام دادم، متوجه شدم از این ویژگی برای تعیین مناسب تعداد عنوان استفاده می‌کنند.** در صورتی که با افزایش تعداد عنوان Perplexity به صورت مرتب کاهش بیابد، تعداد عنوانی که شبیب نمودار تغییراتش کم می‌شود را، تعداد عنوان بهینه انتخاب می‌کنند. در حالت دوم تعداد عنوانی که مقدار Perplexity شروع به افزایش می‌کند را به عنوان تعداد عنوان بهینه انتخاب می‌کنند.

## ۵ سوال ۳- نمونه برداری

-۲ Mixing در این سوال سه نمونه برداری را مورد بررسی قرار می‌دهیم: ۱- انتخاب یک نمونه بعد از رسیدن به Mixing ۲- انتخاب نمونه‌های بعد از رسیدن به Mixing ۳- انتخاب نمونه‌های بعد از رسیدن به Mixing به صورت پنج تا در میان.

برای این کار سه شبکه ۲۲، ۲۳ و ۲۴ را با دیتاست یک آموزش داده‌ایم.

### ۵.۱.۱ مدل بیست و دو

پارامترهای مختلف مدل:

جدول ۴ پارامترهای مدل بیست و دو

۰,۵	$\alpha$
۰,۵	$\beta$
۲۵	W (تعداد کلمه‌ها در دیکشنری)
۱۰	T (تعداد عنوان‌ها)
۱۶۰	Max Epoch (حداکثر تعداد ایپاک)
دیتاست یک	دیتاست
استفاده از یک نمونه	نمونه برداری

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوشه‌ی **model-22-dataset1** موجود است، خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت- تصویر ها : **Figure\_1.png**

تصویر عنوان‌های به دست آمده (Topics) : **Figure\_2.png**

تصویر Perplexity در حین آموزش **Figure\_3.png**

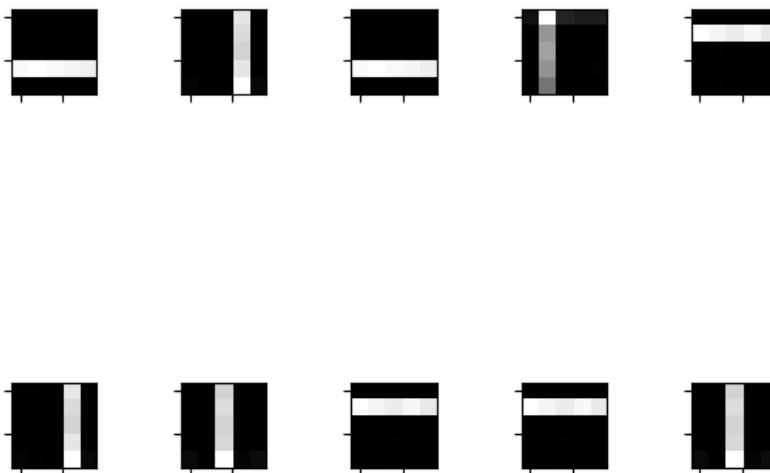
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$ ، تعداد کلمات، تعداد تاپیکها، Perplexity حداکثر تعداد ایپاکها، زمان میانگین انجام هر ایپاک،  $\theta$  ،  $\Phi$  ، میزان در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

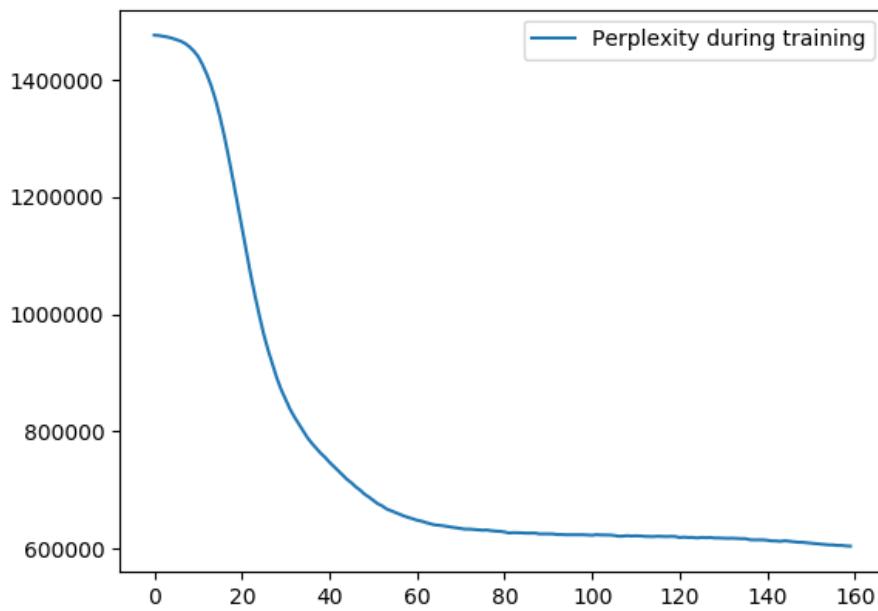
جدول ۵۰ خروجی مدل بیست و دو

تصویر تعدادی از دکیومنت‌ها	some randomly selected samples

تصویر عنوان‌های به  
دست آمده  
Topics



Perplexity  
در ایپاک‌های مختلف



زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۷۵۵,۱۰۱۱۳۴۳۰۰۲۳۱۹۵

زمان میانگین برای  
انجام یک ایپاک

۴,۷۱۹۳۸۲۰۸۹۳۷۶۴۵

160	ایپاکهای لازم برای رسیدم به حالت Mixing
۷۵۵,۱۰۱۱۳۴۳۰۰۲۳۱۹۵	زمان لازم برای رسیدن mixing به حالت
<pre>{   "total_time": 755.1011343002319,   "each_epoch_time": 4.71938208937645,   "W": 25,   "T": 10,   "alpha": 0.5,   "beta": 0.5,   "dataset": 2,   "phi": [     [       0.0010721688785113532,       0.018560434585785424,       0.0008339091277310524,       7.147792523409021e-05,       7.147792523409021e-05,       0.0002144337757022706,       0.04586500202520788,       7.147792523409021e-05,       0.00026208572585833075,       0.0002144337757022706,       0.00011912987539015035,       0.038907817302423105,       0.0026446832336613378,       0.0012627766791355936,       0.0007386052274189321,       0.00026208572585833075,       0.05777798956422291,       7.147792523409021e-05,       0.00016678182554621048,       2.222507507900200688e-05     ],     [       6.04545, 8684487643     ]   ] }</pre>	فایل model-dataset1.json در پوششی Model-22-dataset-1 که شامل $\Phi$ ، $\theta$ و سایر خروجی‌ها است
۶۰۴۵۴۵,۸۶۸۴۴۸۷۶۴۳	آخرین perplexity

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاکهای مختلف رو به کاهش است.

## ۵.۱.۲ مدل بیست و سه

پارامترهای مختلف مدل:

جدول ۵ پارامترهای مدل بیست و سه

۰,۵	$\alpha$
۰,۵	$\beta$
۲۵	W(تعداد کلمه‌ها در دیکشنری)
۱۰	T تعداد عنوان‌ها
۱۶۰	Max Epoch حداکثر تعداد ایپاک
دیتاست یک	دیتاست
استفاده از نمونه‌های بعد از mixing	نمونه برداری

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوشه‌ی **model-23-dataset1** موجود است، خروجی‌ها به شکل زیر اند:

: تصویر تعدادی از داکیومنت-تصویر ها **Figure\_1.png**

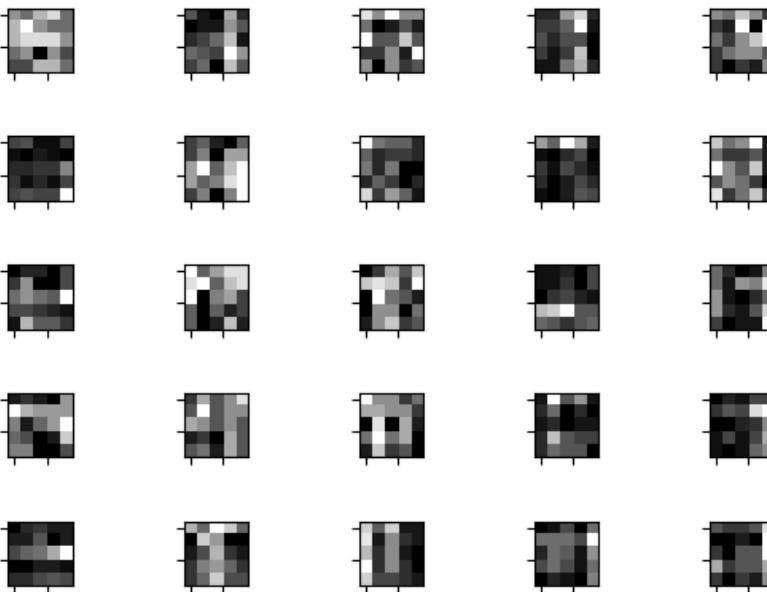
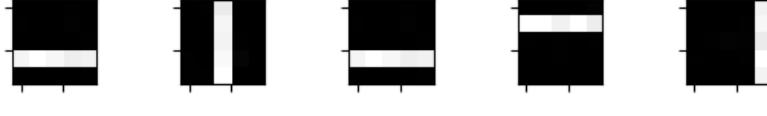
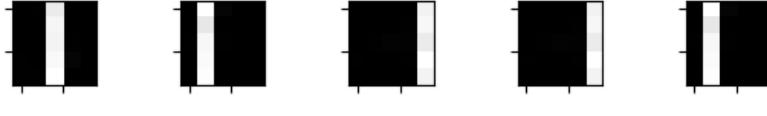
: تصویر عنوان‌های به دست آمده (Topics) **Figure\_2.png**

: تصویر Perplexity در حین آموزش **Figure\_3.png**

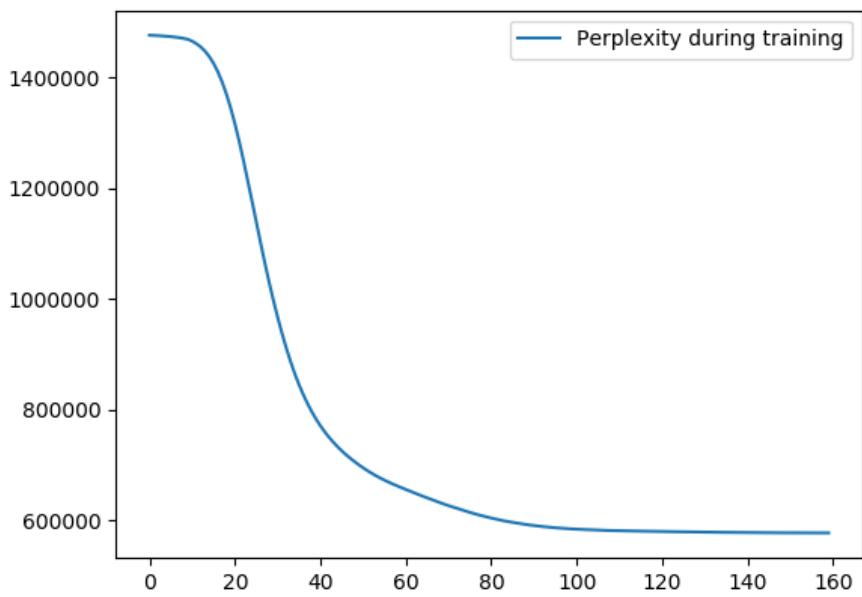
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ، میزان **Perplexity** در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۵۲ خروجی مدل سیزده

<p>some randomly selected samples</p> 	<p>تصویر تعدادی از دکیومنټها</p>
 	<p>تصویر عنوانهای به دست آمده Topics</p>

Perplexity  
در ایپاک‌های مختلف



زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۷۱۷,۶۷۰,۵۴۸۶۷۷۴۴۴۵ S

زمان میانگین برای  
انجام یک ایپاک

۴,۴۸۵۴۴۰,۹۲۹۲۳۴۰,۲۸ S

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

160

زمان لازم برای رسیدن  
mixing به حالت

۷۱۷,۶۷۰,۵۴۸۶۷۷۴۴۴۵ S

<pre>{     "total_time": 717.6705486774445,     "each_epoch_time": 4.485440929234028,     "W": 25,     "T": 10,     "alpha": 0.5,     "beta": 0.5,     "dataset": 2,     "phi": [         [             0.00013536257038970154,             0.00030679031717098737,             0.0009204088537255857,             0.00020052263607808443,             0.00042235752539055576,             0.0003116101822237337,             0.0022942707094225087,             0.003010597491130093,             0.00021092004424151867,             0.0004272192892956512,             7.537374535728126e-05,             0.0011348635364878824,             0.002774095966443263,             0.0022864669540708475,             0.0008406470783809787,             0.0012555049479763066,             0.0004628666928021042,             0.0007700805872871634,             ...         ]     ],     "perplexity": 577989.4493604839 }</pre>	فایل model-dataset1.json در پوشه‌ی Model-23-dataset-1 که شامل ، $\theta$ و $\Phi$ و سایر خروجی‌ها است
---	--

همین‌طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است.

### ۵.۱.۳ مدل بیست و چهار

پارامترهای مختلف مدل:

جدول ۵۳ پارامترهای مدل بیست و چهار

۰,۵	$\alpha$
۰,۵	$\beta$
۲۵	W(تعداد کلمه‌ها در دیکشنری)

۱۰	<b>T تعداد عنوان‌ها</b>
۱۶۰	<b>Max Epoch حداکثر تعداد ایپاک</b>
دیتاست یک	دیتاست
استفاده از نمونه‌های بعد از mixing به صورت پنج تا در میان	نمونه برداری

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوششی **model-24-dataset1** موجود است، خروجی‌ها به شکل زیر اند:

تصویر تعدادی از داکیومنت-تصویر ها : **Figure\_1.png**

تصویر عنوان‌های به دست آمده (Topics) : **Figure\_2.png**

تصویر Perplexity در حین آموزش : **Figure\_3.png**

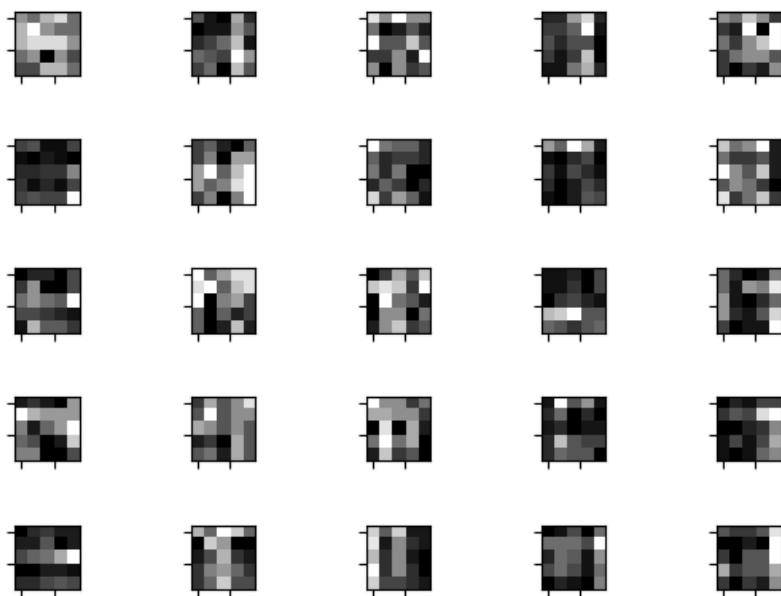
یک فایل json که با انواع text-editorها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++ پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$  ،  $\beta$  ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\Phi$  ،  $\theta$  ، میزان Perplexity در هر ایپاک است.

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

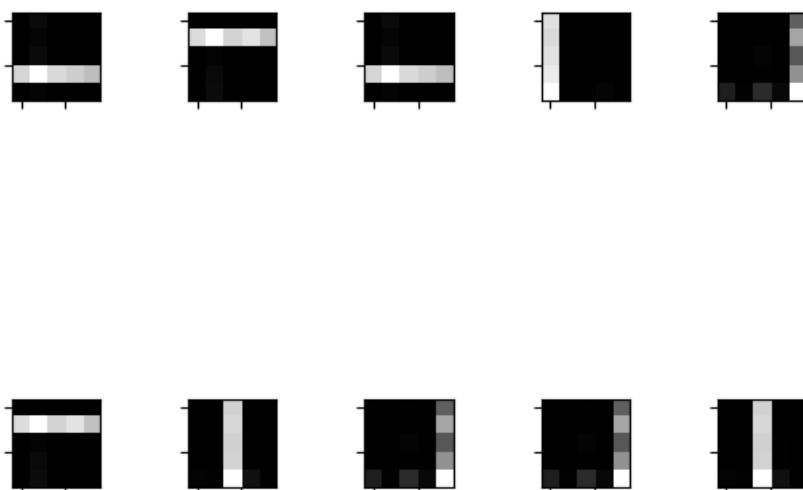
جدول ۴۵ خروجی مدل بیست و چهار

تصویر تعدادی از  
دکیومنټها

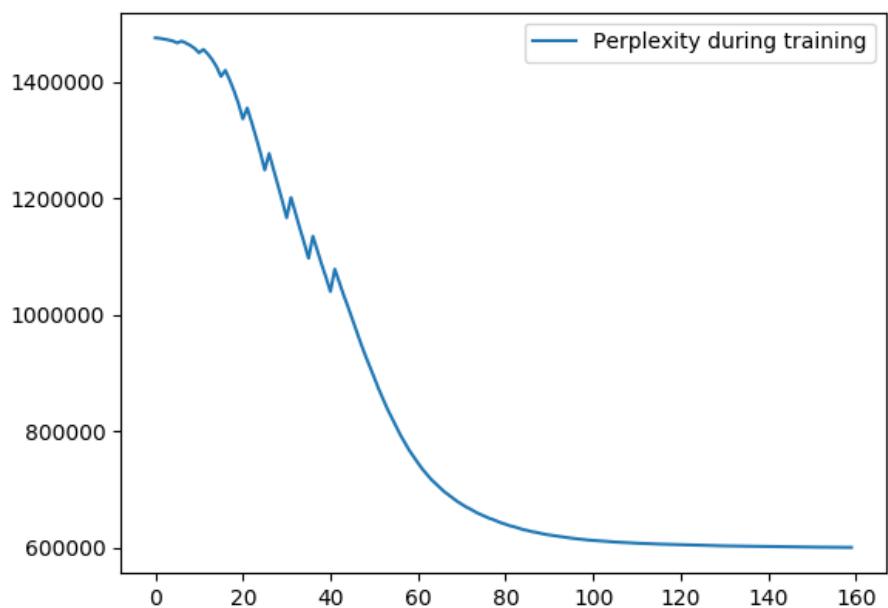
some randomly selected samples



تصویر عنوان‌های به  
دست آمده  
Topics



Perplexity  
در ایپاک‌های مختلف



زمان کل اجرای حداکثر  
ایپاک‌های مجاز

۷۲۱,۰۵۶۲۲۸۸۷۶۱۱۳۹ s

زمان میانگین برای  
انجام یک ایپاک

۴,۵۰۶۶۰۱۴۳۰۴۷۵۷۱۲ s

ایپاک‌های لازم برای  
رسیدم به حالت  
Mixing

160

زمان لازم برای رسیدن  
mixing به حالت

۷۲۱,۰۵۶۲۲۸۸۷۶۱۱۳۹ s

آخرین Perplexity

۶۰۱۱۲۴,۵۴۱۸۱۲۰۹۱۶

فایل

model-dataset1.json

در پوششی

Model-24-dataset-1

که شامل ،  $\Phi$  و

سایر خروجی‌ها است

```
{  
    "total_time": 721.0562288761139,  
    "each_epoch_time": 4.506601430475712,  
    "W": 25,  
    "T": 10,  
    "alpha": 0.5,  
    "beta": 0.5,  
    "dataset": 2,  
    "phi": [  
        [  
            0.00031214110703152925,  
            0.09300506853714541,  
            0.00021991066785030723,  
            0.00013781966833225246,  
            0.00010460445096408212,  
            0.00022401536414522175,  
            0.09308192284004906,  
            0.0004909038764644985,  
            7.463081422543085e-05,  
            0.00010036085881321172,  
            0.00010023758170102486,  
            0.09338352026961484,  
            0.0009554894548137066,  
            0.0004109728123353562,  
            0.00011325580523110406,  
            0.0001467377541864198,  
            0.10541696958502526,  
            7.428813481814228e-05,  
            0.00016395043982766688,  
            1.473620412802866e-05  
        ]  
    ]  
}
```

همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است.

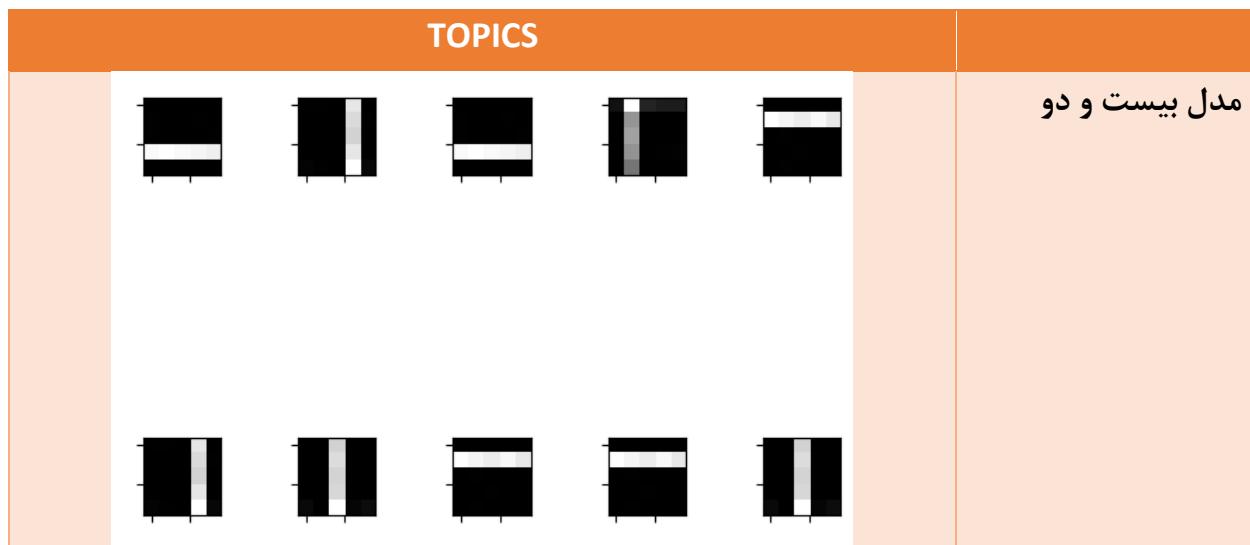
## ۵.۲ نتیجه گیری - روش‌های نمونه برداری

در بخش ۱-۵ سه مدل یکسان را با داده‌های دیتابست یک ایجاد کردیم. در جدول‌های ۴۹ تا ۵۴، پارامترها و خروجی‌های آن‌ها را مشاهده می‌کنید.

جدول ۵۵ مقایسه‌ی پرپلیسکی در روش‌های مختلف نمونه برداری

Perplexity نهایی	
۶۰۴۵۴۵,۸۶۸۴۴۸۷۶۴۳	مدل بیست و دو(یک نمونه)
۵۷۷۹۸۹,۴۴۹۳۶۰۴۸۳۹	مدل بیست و سه(نمونه‌های بعد از Mixing)
۶۰۱۱۲۴,۵۴۱۸۱۲۰۹۱۶	مدل بیست و چهار(نمونه‌های بعد از Mixing با فاصله‌ی ۵ در میان)

جدول ۵۶ مقایسه‌ی عنوان‌ها در روش‌های مختلف نمونه برداری





نتیجه‌گیری: همین طور که در جدول ۵۵ و ۵۶ مشاهده می‌شود و همین طور با توجه به تصادفی بودن بخش‌های مختلف، سه روش در Perplexity نهایی و عنوان‌های استخراج شده بسیار شبیه به هم عمل کرده‌اند در نتیجه بعد از عبور از Mixing ، روش نمونه برداری چندان مهم نیست و عملکردھایی نزدیک به هم از خود نشان می‌دهند.

## ۶ سوال ۴- خوشه بندی اسناد

برای خوشه بندی متن‌ها ابتدا مدل ۲۵، با پارامترهای زیر برای دیتاست دو، را آموزش می‌دهیم.

### ۶.۱ مدل بیست و پنج

پارامترهای مختلف مدل:

جدول ۵ پارامترهای مدل بیست و پنج

۰,۴	$\alpha$
۰,۵	$\beta$
۱۰۴۷۳	W (تعداد کلمه‌ها در دیکشنری)
۳۰	T (تعداد عنوان‌ها)
۱۲۰	Max Epoch (حداکثر تعداد ایپاک)
دیتاست دو	دیتاست

خروجی‌های حاصل از اجرای برنامه:

کلیه‌ی خروجی‌های حاصل از اجرای کد برای مدل یک در پوشه‌ی **model-25-dataset2** موجود است، خروجی‌ها به شکل زیر اند:

نمودار Perplexity در حین آموزش: **Figure\_1.png**

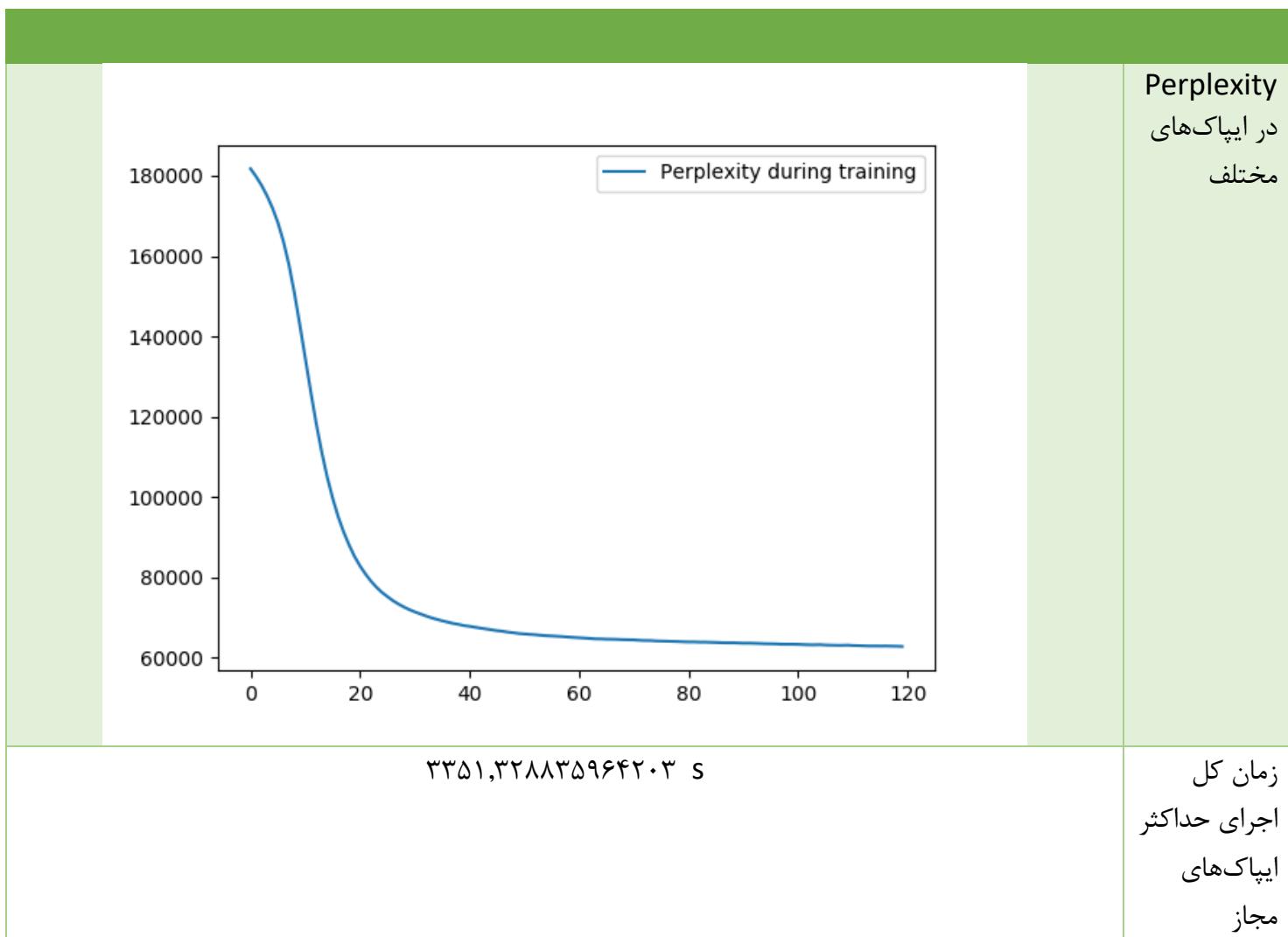
: یک فایل json که با انواع text-editor ها قابل باز کردن و مشاهده است، (برای نمایش بهتر Notepad++) پیشنهاد می‌شود). **این فایل شامل** پارامترهای  $\alpha$ ،  $\beta$ ، تعداد کلمات، تعداد تاپیک‌ها، حداکثر تعداد ایپاک‌ها، زمان انجام کلیه‌ی ایپاک‌ها، زمان میانگین انجام هر ایپاک،  $\theta$ ،  $\Phi$ ، میزان Perplexity در هر ایپاک است.

در فایل `model-dataset2.json`, مقدار  $\theta$  و  $\Phi$  به صورت عددی موجود است، در فایل `topic-topics.txt`

`topics.txt`، کلمه‌ها را از دیکشنری برمی‌داریم و در  $\Phi$  جایگزین می‌کنیم. کلماتی که احتمال بیشتر دارند را در ابتدا قرار می‌دهیم. به این ترتیب عنوان‌ها در فایل `topics.txt` قابل مشاهده‌اند. (برای نمایش بهتر از یک خوب مانند `notepad++` استفاده شود).

در زیر خلاصه‌ای از خروجی‌ها که در بالا به آن اشاره شد و را مشاهده می‌کنید:

جدول ۵۸ خروجی مدل بیست و پنج



27.92774029970169 s	زمان میانگین برای انجام یک ایپاک
120 Epochs	ایپاکهای لازم برای رسیدم به حالت Mixing
۳۳۵۱,۳۲۸۸۳۵۹۶۴۲۰۳ s	زمان لازم برای رسیدن به حالت mixing

فایل

model-dataset1.json

در پوششی

Model-21-dataset-2

که شامل ،  $\theta$ .

$\Phi$  و سایر

خروجی‌ها

است

```
"total_time": 3351.328835964203,  
"each_epoch_time": 27.92774029970169  
"W": 10473,  
"T": 30,  
"alpha": 0.4,  
"beta": 0.5,  
"dataset": 2,  
"phi": [  
    0.04592684673485657,  
    0.004661595211393209,  
    0.00012160683160156197,  
    0.014200975557026848,  
    0.00487778513424043,  
    0.0032563607128862707,  
    6.755935088975666e-05,  
    0.001553865070464403,  
    0.0038238592603602266,  
    6.755935088975666e-05,  
    0.008742180005134511,  
    0.004850761393884528,  
    4.053561053385399e-05,  
    0.0010133902633463498,  
    0.00020267805266926996,  
    0.0009863665229904472,  
    1.351187017795133e-05,  
    0.00014863057195746464,  
    4.053561053385399e-05,  
    0.007688254131254307,  
    0.0023105298004296774,
```

فایل

topics.txt

در پوششی

Model-21-dataset-2

که شامل

عنوان‌ها به

صورت کلمه

است

```

1 ['i', 'people', 'dont', 'think', 'like', 'just', 'years', 'going', 'get', 'time', 'say', 'says'
2 ['soviet', 'gorbachev', 'union', 'president', 'party', 'moscow', 'soviet', 'reagan', 'news', '
3 ['percent', 'year', 'million', 'market', 'billion', 'prices', 'stock', 'sales', 'rose', 'last',
4 ['workers', 'new', 'percent', 'plant', 'union', 'work', 'contract', 'labor', 'years', 'report',
5 ['court', 'charges', 'trial', 'attorney', 'case', 'prison', 'judge', 'federal', 'convicted', 'd
6 ['new', 'york', 'years', 'year', 'film', 'art', 'music', 'john', 'record', 'movie', 'first', 'w
7 ['women', 'rights', 'human', 'abortion', 'members', 'magazine', 'souter', 'club', 'new', 'men',
8 ['church', 'pope', 'john', 'catholic', 'vatican', 'king', 'paul', 'mexico', 'bishops', 'told',
9 ['ms', 'children', 'hospital', 'ms', 'care', 'family', 'mother', 'doctors', 'parents', 'mandel
10 ['police', 'people', 'government', 'killed', 'two', 'violence', 'reported', 'army', 'city', 'ca
11 ['drug', 'aids', 'health', 'computer', 'virus', 'disease', 'system', 'people', 'program', 'rese
12 ['air', 'navy', 'space', 'ship', 'force', 'two', 'military', 'accident', 'shuttle', 'launch', '
13 ['news', 'network', 'cbs', 'million', 'nbc', 'television', 'abc', 'tv', 'rating', 'week', 'show
14 ['trade', 'farmers', 'japan', 'united', 'japanese', 'farm', 'states', 'agriculture', 'agreement
15 ['fire', 'southern', 'miles', 'area', 'northern', 'people', 'water', 'central', 'high', 'offici
16 ['dollar', 'late', 'yen', 'london', 'gold', 'bid', 'tokyo', 'new', 'bank', 'dealers', 'dollars'
17 ['school', 'students', 'university', 'student', 'schools', 'education', 'board', 'college', 'te
18 ['house', 'bill', 'congress', 'senate', 'budget', 'committee', 'tax', 'billion', 'rep', 'year',
19 ['united', 'states', 'iraq', 'kuwait', 'iraqi', 'military', 'american', 'bush', 'saudi', 'gulf'
20 ['police', 'two', 'man', 'found', 'shot', 'yearold', 'authorities', 'killed', 'car', 'night', '
21 ['bush', 'dukakis', 'campaign', 'president', 'democratic', 'jackson', 'republican', 'presidenti
22 ['south', 'west', 'east', 'german', 'germany', 'africa', 'north', 'united', 'african', 'war', '
23 ['air', 'plane', 'airlines', 'flight', 'north', 'eastern', 'airport', 'pilots', 'airline', 'air
24 ['heart', 'wine', 'greyhound', 'medical', 'days', 'patients', 'hospital', 'dr', 'bus', 'surgery
25 ['oil', 'cents', 'futures', 'cent', 'lower', 'higher', 'prices', 'tons', 'crude', 'trading', 'm
26 ['government', 'party', 'political', 'opposition', 'president', 'national', 'minister', 'electi
27 ['israel', 'government', 'israeli', 'aid', 'army', 'military', 'rebels', 'peace', 'two', 'talks
28 ['water', 'scientists', 'species', 'venus', 'time', 'fish', 'state', 'river', 'earth', 'wildlif
29 ['company', 'million', 'new', 'billion', 'corp', 'inc', 'bank', 'stock', 'offer', 'co', 'chairm
30 ['court', 'state', 'law', 'case', 'filed', 'federal', 'ruling', 'judge', 'decision', 'legal', '
31

```

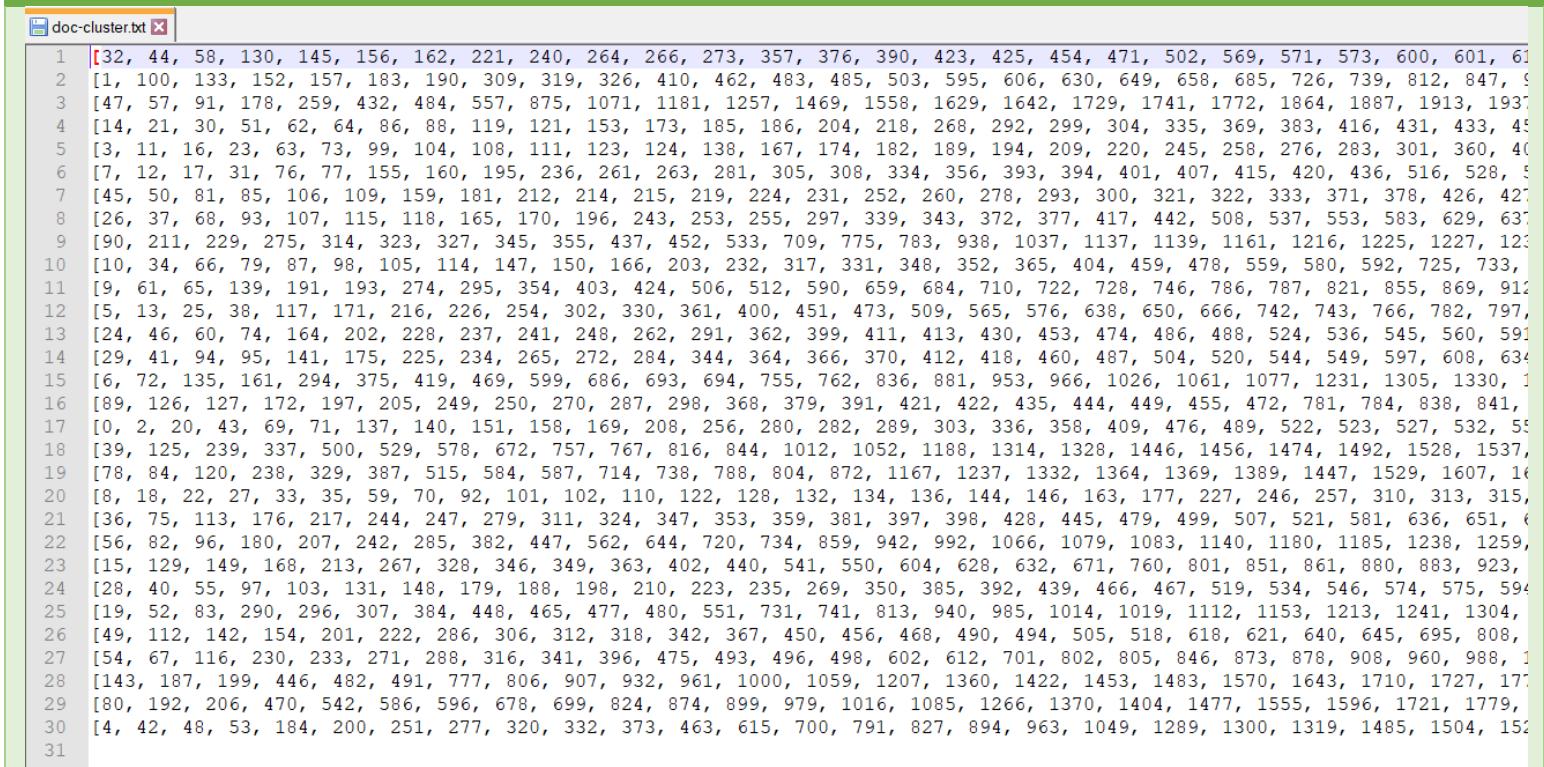
همین طور که در خروجی‌های بالا دیده می‌شود، perplexity در ایپاک‌های مختلف رو به کاهش است و کلمات مختلف در هر کدام از عنوان‌ها، به هم مرتبط‌اند.

## ۶.۲ خوشبندی با استفاده از تابع مدل بیست و پنج

بعد از ایجاد مدل بیست و پنج، با استفاده از الگوریتم Kmeans کتابخانه Sklearn استفاده می‌کنیم تا سندها را خوشبندی کنیم. بعد از آموزش مدل  $\theta_{25}$ ، را به دست می‌آوریم. تا یک ماتریس است که تعداد سطرهای آن برابر با تعداد سندهای آن برابر با تعداد ستون‌های آن برابر با تعداد عنوان هاست. تا اهمیت هر عنوان در هر سند را مشخص می‌کند. می‌تواند از تابع  $\text{A}^T \theta$  به عنوان وکتور ویژگی سند  $\text{A}$  استفاده کرد. به این ترتیب با استفاده از تابع کلاسیفیکر را انجام می‌دهیم.

در زیر خروجی کد خوشبندی را مشاهده می‌کنید، سندها را به ۳۰ خوشبندی تقسیم کرده‌ایم، هر سطر نشان‌دهنده‌ی یک خوشبندی است. در هر سطر شماره‌هایی قرار دارد که نشان دهنده‌ی شماره‌ی سندهایی که خوشبندی است.

### قسمتی از فایل خروجی doc-cluster.txt را شکل زیر مشاهده می‌کنید.



```

doc-cluster.txt
1 [32, 44, 58, 130, 145, 156, 162, 221, 240, 264, 266, 273, 357, 376, 390, 423, 425, 454, 471, 502, 569, 571, 573, 600, 601, 61
2 [1, 100, 133, 152, 157, 183, 190, 309, 319, 326, 410, 462, 483, 485, 503, 595, 606, 630, 649, 658, 685, 726, 739, 812, 847, 9
3 [47, 57, 91, 178, 259, 432, 484, 557, 875, 1071, 1181, 1257, 1469, 1558, 1629, 1642, 1729, 1741, 1772, 1864, 1887, 1913, 193
4 [14, 21, 30, 51, 62, 64, 86, 88, 119, 121, 153, 173, 185, 186, 204, 218, 268, 292, 299, 304, 335, 369, 383, 416, 431, 433, 49
5 [3, 11, 16, 23, 63, 73, 99, 104, 108, 111, 123, 124, 138, 167, 174, 182, 189, 194, 209, 220, 245, 258, 276, 283, 301, 360, 40
6 [7, 12, 17, 31, 76, 77, 155, 160, 195, 236, 261, 263, 281, 305, 308, 334, 356, 393, 394, 401, 407, 415, 420, 436, 516, 528, 5
7 [45, 50, 81, 85, 106, 109, 159, 181, 212, 214, 215, 219, 224, 231, 252, 260, 278, 293, 300, 321, 322, 333, 371, 378, 426, 427
8 [26, 37, 68, 93, 107, 115, 118, 165, 170, 196, 243, 253, 255, 297, 339, 343, 372, 377, 417, 442, 508, 537, 553, 583, 629, 637
9 [90, 211, 229, 275, 314, 323, 327, 345, 355, 437, 452, 533, 709, 775, 783, 938, 1037, 1137, 1139, 1161, 1216, 1225, 1227, 123
10 [10, 34, 66, 79, 87, 98, 105, 114, 147, 150, 166, 203, 232, 317, 331, 348, 352, 365, 404, 459, 478, 559, 580, 592, 725, 733,
11 [9, 61, 65, 139, 191, 193, 274, 295, 354, 403, 424, 506, 512, 590, 659, 684, 710, 722, 728, 746, 786, 787, 821, 855, 869, 912
12 [5, 13, 25, 38, 117, 171, 216, 226, 254, 302, 330, 361, 400, 451, 473, 509, 565, 576, 638, 650, 666, 742, 743, 766, 782, 797,
13 [24, 46, 60, 74, 164, 202, 228, 237, 241, 248, 262, 291, 362, 399, 411, 413, 430, 453, 474, 486, 488, 524, 536, 545, 560, 591
14 [29, 41, 94, 95, 141, 175, 225, 234, 265, 272, 284, 344, 364, 366, 370, 412, 418, 460, 487, 504, 520, 544, 549, 597, 608, 634
15 [6, 72, 135, 161, 294, 375, 419, 469, 599, 686, 693, 755, 762, 836, 881, 953, 966, 1026, 1061, 1077, 1231, 1305, 1330, 1
16 [89, 126, 127, 172, 197, 205, 249, 250, 270, 287, 298, 368, 379, 391, 421, 422, 435, 444, 449, 455, 472, 781, 784, 838, 841,
17 [0, 2, 20, 43, 69, 71, 137, 140, 151, 158, 169, 208, 256, 280, 282, 289, 303, 336, 358, 409, 476, 489, 522, 523, 527, 532, 55
18 [39, 125, 239, 337, 500, 529, 578, 672, 757, 767, 816, 844, 1012, 1052, 1188, 1314, 1328, 1446, 1456, 1474, 1492, 1528, 1537,
19 [78, 84, 120, 238, 329, 387, 515, 584, 587, 714, 738, 788, 804, 872, 1167, 1237, 1332, 1364, 1369, 1389, 1447, 1529, 1607, 16
20 [8, 18, 22, 27, 33, 35, 59, 70, 92, 101, 102, 110, 122, 128, 132, 134, 136, 144, 146, 163, 177, 227, 246, 257, 310, 313, 315,
21 [36, 75, 113, 176, 217, 244, 247, 279, 311, 324, 347, 353, 359, 381, 397, 398, 428, 445, 479, 499, 507, 521, 581, 636, 651, 6
22 [56, 82, 96, 180, 207, 242, 285, 382, 447, 562, 644, 720, 734, 859, 942, 992, 1066, 1079, 1083, 1140, 1180, 1185, 1238, 1259,
23 [15, 129, 149, 168, 213, 267, 328, 346, 349, 363, 402, 440, 541, 550, 604, 628, 632, 671, 760, 801, 851, 861, 880, 883, 923,
24 [28, 40, 55, 97, 103, 131, 148, 179, 188, 198, 210, 223, 235, 269, 350, 385, 392, 439, 466, 467, 519, 534, 546, 574, 575, 594
25 [19, 52, 83, 290, 296, 307, 384, 448, 465, 477, 480, 551, 731, 741, 813, 940, 985, 1014, 1019, 1112, 1153, 1213, 1241, 1304,
26 [49, 112, 142, 154, 201, 222, 286, 306, 312, 318, 342, 367, 450, 456, 468, 490, 494, 505, 518, 618, 621, 640, 645, 695, 808,
27 [54, 67, 116, 230, 233, 271, 288, 316, 341, 396, 475, 493, 496, 498, 602, 612, 701, 802, 805, 846, 873, 878, 908, 960, 988, 1
28 [143, 187, 199, 446, 482, 491, 777, 806, 907, 932, 961, 1000, 1059, 1207, 1360, 1422, 1453, 1483, 1570, 1643, 1710, 1727, 17
29 [80, 192, 206, 470, 542, 586, 596, 678, 699, 824, 874, 899, 979, 1016, 1085, 1266, 1370, 1404, 1477, 1555, 1596, 1721, 1779,
30 [4, 42, 48, 53, 184, 200, 251, 277, 320, 332, 373, 463, 615, 700, 791, 827, 894, 963, 1049, 1289, 1300, 1319, 1485, 1504, 152
31

```

به عنوان مثال سند شماره‌ی صفر و دو در خوشه‌ی ۱۷ قرار گرفته‌اند، در زیر بخشی از متن این دو سند را مشاهده می‌کنید:

A 16-year-old student at a private Baptist school who allegedly killed one teacher and wounded another before firing into a filled clas...

A gunman took a 74-year-old woman hostage after he was foiled in an attempt to steal \$1 million in jewelry belonging to the late Liberace, but police shot and kille....

هر دو متن موضوعی مشابه در مورد موضوع جنایی دارند.

## ۷ بخش‌های مختلف کد

جدول ۶۰ بخش‌های مختلف کد

 <code>read_dataset_1.py</code>	در این فایل دوتابع وجود دارد که با استفاده از آنها اطلاعات مربوط به دیتاست یک را می‌خوانیم.
 <code>read_dataset_2.py</code>	در این فایل یکتابع وجود دارد که با استفاده از آن اطلاعات مربوط به دیتاست دو را می‌خوانیم.
 <code>lda_by_sampling.py</code>	در این فایل تابع <code>LDA</code> نوشته شده است که از یک نمونه‌ی ایجاد شده بعد از رسیدن به <code>Mixing</code> برای محاسبه‌ی تتا و فی استفاده می‌کند.
 <code>lda_by_sampling_more_z.py</code>	در این فایل تابع <code>LDA</code> نوشته شده است که از بیش از یک نمونه‌ی ایجاد شده بعد از رسیدن به <code>Mixing</code> برای محاسبه‌ی تتا و فی استفاده می‌کند.
 <code>perplexity.py</code>	در این فایل تابع محاسبه‌ی <code>perplexity</code> قرار دارد.
 <code>titles_dataset2.py</code>	در این فایل تابعی قرار دارد که با دریافت فی، عنوان‌ها را با لغتهای درون دیکشنری بر می‌گرداند.
 <code>display_samples.py</code>	این تابع سندها و عنوان‌های دیتاست یک را به صورت تصویر و مرتب نشان می‌دهد.
 <code>run_lda_dataset_1.py</code>	این فایل با استفاده از تابع‌های بالا، دیتاست یک را می‌خواند و با استفاده از <code>lda</code> که از یک نمونه استفاده می‌کند، تتا و فی، زمان اجرای و ... را محاسبه می‌کند و آنها را نمایش می‌دهد و خروجی‌های لازم را در فایل ذخیره می‌کند.
 <code>run_lda_dataset_1_more_z.py</code>	این فایل با استفاده از تابع‌های بالا، دیتاست یک را می‌خواند و با استفاده از <code>lda</code> که از بیش از یک نمونه استفاده می‌کند، تتا و فی، زمان اجرای و ... را محاسبه می‌کند و آنها را نمایش می‌دهد و خروجی‌های لازم را در فایل ذخیره می‌کند.
 <code>run_lda_dataset_2.py</code>	این فایل با استفاده از تابع‌های بالا، دیتاست دو را می‌خواند و با استفاده از <code>lda</code> که از یک نمونه استفاده می‌کند، تتا و فی، زمان اجرای و ... را محاسبه می‌کند و آنها را نمایش می‌دهد و خروجی‌های لازم را در فایل ذخیره می‌کند.
 <code>clustering.py</code>	کد خوشه‌بندی در این فایل قرار دارد.