
 <p>دانشگاه صنعتی امیرکبیر</p>	<p>دوره آموزشی پردازش زبان طبیعی (NLP)</p> <p>تمرین دوم</p> <p>مهلت تحویل: ۲۳ مرداد ۱۴۰۰</p>	 <p>آکادمی همراه</p>
---	--	---

برای این تمرین فایل داده آموزشی و تست ارائه شده است که هر خط شامل یک جمله است. هر فاصله را جدا کننده بین کلمات در نظر بگیرید. که طریقه دریافت به صورت زیر است:

```
#https://drive.google.com/file/d/1qCVYpb67RuzUbyrJ3w-ohtCf-tzZKTal/view?usp=sharing
#https://drive.google.com/file/d/1dW5SkCYIFbXmNe3xKv4EhrLPDPlXy-IDy/view?usp=sharing
!gdown --id 1qCVYpb67RuzUbyrJ3w-ohtCf-tzZKTal
!gdown --id 1dW5SkCYIFbXmNe3xKv4EhrLPDPlXyIDy
```

نکته ۱: فاصله‌های پشت سرهم اضافی و فاصله در اول و آخر خطوط را در هیچ سؤالی در نظر نگیرید.

نکته ۲: خطوطی که کم‌تر از سه کلمه دارند را در هیچ سؤالی در نظر نگیرید.

نکته ۳: دقت کنید که جمله‌ها با خط جدید از هم جدا شده‌اند و مانند مدل‌های موجود در برخی منابع نباید جملات متفاوت را بهم بچسبانید. فرق این مساله در این موضوع است که گاهی به اشتباه کلمه آخر جمله قبل را با کلمه اول جمله فعلی را در یک ترکیب در نظر می‌گیرد که صحیح نیست.

سوال ۱ همانند قسمت آخر تمرین قبل جملات زیر را اینبار با روش Dirichlet Prior کامل کنید. دقت کنید فاصله بین خط تیره‌ها یعنی بیش از یک کلمه جا افتاده است.

$$\mu = 0.01 \text{ and } P_{BG} = 1/V$$

چون مشک سیه بود مرا هر دو ---

گر خورد سوگند هم آن ----

زانک نفس آشفته تر گردد از --

ازین زشت تر در جهان رنگ ----

دوست در خانه و ما گرد -----

شب است و شمع و شراب و -----

سوال ۲ Perplexity روش قبل را به کمک فایل test محاسبه کنید.

سوال ۳ سوال یک را این بار با روش مدل‌های شبکه عصبی تری‌گرم حل کنید.