فایل داده تمرینی داده شده در هر خط شامل یک جمله است. هر فاصله را جدا کننده بین کلمات درنظر بگیرید.

آدرس فایل:

https://drive.google.com/file/d/1Hb58rR--Qjwr21cLK6A29ROp6Uy5pqq8/view?usp=sharing

نحوه دریافت فایل با کولب:

```
!gdown --id 1Hb58rR--Qjwr21cLK6A29ROp6Uy5pqq8
```

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

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

سوال ۱) ترکیب‌های unigram و bigram فایل داده شده را بشمارید.

نکته ۳: تعداد تکرار کلمات و جفت کلمات را بشمارید. ۲۰ تای پر تکرار را به ترتیب رنک به همراه تعداد تکرار شان و ضرب رنک در تعداد تکرار چاپ کنید که به این صورت است.

```
1	50936 50936 و

2	31213 62426 از
```

نکته ۴: استفاده از کتابخانه مجاز نیست.

سوال ۲) یک تابع بنویسید که برای هر عدد ورودی دلخواه n (بین ۱ تا ۵) تمام ترکیبات n-gram را برای آن عدد ورودی از فایل آموزش استخراج کند و به همراه تعداد در خروجی چاپ کند.

امضای تابع؛

```
def compute_freq (file:str, n:int):
```

نمونه فراخوانی:

```
compute_freq("cleaned_train.txt",3)
```

سوال ۳) احتمال عبارات زیر را برحسب گرامهای ۱ ۲ ۳ محاسبه کنید.

- چون تویی آید به زیبایی و شیرینی پسر
- دل در این درد و رنج پاره کنیم
- ای به آرام تو زمین را سنگ

- جان را زند آ باغ صلاهای تعالوا
- شاهد و شمع و شراب و مطرب آنجا بهترست
- شب است و شمع و شراب و شیرینی

نکته ۵) در هر خط یک احتمال و سپس جمله را درج نمایید. و ابتدا همه جملات را برای unigram انجام دهید بعد یک خط فاصله bigram ها را چاپ کنید و سپس trigram ها را چاپ کنید.

سوال۴) چرا نمی‌توان احتمال راخداد « شب است و شمع و شراب و شیرینی» را برحسب trigram محاسبه کرد؟ راه حل پیشنهادی خود را بنویسید.

سوال ۵) عبارات زیر را به کمک مدل bigram کامل کنید (برای هر جای خالی دو کلمه محتمل تر را پیشنهاد دهید):

چون مشک سیه بود مرا هر دو بنا ---

گر خورد سوگند هم آن -----

زانک نفس آشفته تر گردد از --

ازین زشت تر در جهان رنگ ----

نکته ۶) برای هر جمله چهار خط خروجی لازم است:

- خط ۱: جمله ناقص داده شده در متن سوال
- خط۲: لیست تمام گزینه‌های محتمل را به همراه تعداد تکرار به صورت مرتب شده نزولی
- خط ۳: جمله کامل شده با کلمه پیشنهادی اول
- خط ۴: جمله کامل شده با کلمه پیشنهادی دوم