**سوالات 1-3 و 5 و 6: لطفا برای "هر" سوال، یک فایل پایتون py. یا ipynb. تحویل دهید**

**سوال 4: لطفا توضیحات را در یک فایل text یا word بنویسید.**

در این سوال، فایل NGS.fasq.gz و check.txt در اختیارتان قرار داده شده‌است. فایل NGS.fasq.gz، نتیجه‌ی Illumina Sequencing روی یک نمونه‌ی زیستی است .ورودی دیگر فایل check.txt است. فایل check.txt یک فایل تکست عادی است که در هر خط یک توالی دارد.

همچنین فایل‌های زیر در اختیارتان قرار داده شده:

Tree.py: این فایل دارای کد مرتبط با Node درخت است.

Read_NGS.py این فایل دارای کد مرتبط با خواندن فایل NGS.fastq.gz است. **(توجه داشته باشید که برای کار کردن کد باید فایل کد و فایل NGS.fasq.gz هر دو در یک فولدر باشند)**

2 عدد فیلم. (لطفا فیلم‌ها را پس از رسیدن به قسمت درخت مشاهده کنید چراکه در ابتدای سوال نیازی به آن نیست)

در این سوال می خواهیم بررسی کنیم که چه تعداد از توالی هایی که در فایل check.txt هستند در فایل NGS.fasq.gz حضور دارند. همچنین می خواهیم بررسی کنیم که هر کدام از توالی‌های مشترک بین NGS.fasq.gz و check.txt چند بار در داده‌ی توالی‌یابی (فایل NGS.fasq.gz) حضور دارند. در قسمت زیر، یک مثال از ورودی و خروجی مورد انتظار این برنامه نشان داده شده است:

فرض کنید فایل NGS.fasq.gz دارای توالی های زیر است:

GCGA

CGTA

CCGG

GCGA

TTTT

همچنین فرض کنید فایل check.txt دارای توالی های زیر است:

AAAA

GGGG

GCGA

CGTA

TTTT

TTTT

در این صورت باید خروجی شما در کونسول به صورت زیر print شود:

GCGA: 2

CGTA: 1

TTTT: 1

TTTT: 1

برای حل این سوال از دو الگوریتم متفاوت استفاده می کنیم. در ابتدا این مسئله را با یک الگوریتم ساده (Naive) حل کنید و سپس باید با استفاده از درخت مسئله را حل کنید.

# حل مسئله با الگوریتم Naive

1) همانطور که می دانید. در داده‌های توالی‌یابی، با توجه به غلظت هر توالی در نمونه، هر توالی نمونه می‌تواند چندین بار حضور داشته باشد. برنامه‌ای بنویسید که پس از خواندن داده‌های NGS.fasq.gz، دو list تولید کند. یک لیست باید مربوط به توالی‌ها باشد (آن را **listOfSequence** بنامید) و یک لیست باید تعداد حضور این توالی در داده‌ی توالی یابی باشد (آن را Occurrence بنامید). (توجه کنید که در این صورت در list اول، همه‌ی توالی ها باید یکتا باشند و هیچ توالی ای نباید تکرار شود. برای چک کردن اینکه یک توالی خاص در لیست هست یا نه، می توانید به صورت مقابل کنید: **if** sequence **in** **listOfSequence**. در صورت استفاده از دستور ذکر شده، اگر توالی sequence در **listOfSequence** وجود داشته باشد، مقدارش True می‌شود و if اجرا می‌شود. (2 نمره)

برای مثال اگر فرض کنید که فایل توالی یابی شامل توالی‌های زیر باشد:

```
AAAA
GGGG
TTTT
CCCC
CCCC
CCCC
TTTT
```

محتویات لیست‌های شما باید به صورت زیر باشد:

```
listOfSequence = [AAAA, GGGG, TTTT, CCCC]
Occurrence = [1, 1, 2, 3]
```

2) برنامه‌ای بنویسید که فایل check.txt را بخواند و همه‌ی خط‌های آن را در کنسول print کند. (1 نمره)

برای خواندن فایل check.txt می توانید از دستور مقابل استفاده کنید: **(فایل کدتان با فایل check.txt باید در یک folder باشند)**

```
myFile = open("check.txt", 'r')
```

سپس می‌توانید در همه‌ی خطوط را به صورت مقابل چک کنید:

```
for line in myFile:
```

اگر لوپ for را به صورت فوق بنویسید، در هر دور لوپ، محتویات line برابر با یک خط از خطوط check.txt است.

3) برنامه‌ای بنویسید که الگوریتم Naive را اجرا کند. الگوریتم Naïve به این صورت عمل می کند که باید چک کند هر توالی ای که در فایل check.txt است، را در **listOfSequence** می‌تواند پیدا کند یا خیر و این کار باید این گونه انجام دهید که به ترتیب اجزای **listOfSequence** را چک کنید و اگر با توالی مورد بررسی یکسان بود توالی باید به همراه تعداد حضورش print شود و سراغ توالی بعدی check.txt بروید. ورودی نمونه و خروجی نمونه‌ی این سوال در ابتدای سوال ذکر شده است (صفحه اول). (2 نمره)

# حل مسئله با استفاده از درخت (ابتدا کلیپ کوتاه را مشاهده کنید)

4) چرا چک کردن درخت الگوریتمی سریع تر و بهینه تر از الگوریتم Naive است؟ (با یک مثال توضیحات خود را توجیه کنید) (2 نمره)

5) الگوریتم Naïve که در بالا بررسی شد، یک الگوریتم کند است زیرا هر چه اندازه‌ی listOfSequence بزرگ‌تر باشد، گشتن در listOfSequence مقدار زیادی طول خواهد کشید. برای اینکه الگوریتم بهینه تری برای حل این سوال بنویسیم، باید از ساختار داده درخت استفاده کنیم. در این سوال از ساختار درخت با توجه به توضیحات داده شده در کلیپ باید استفاده کنید.
تابع/برنامه‌ای بنویسید که فایل NGS.fasq.gz را بخواند و یک درخت از توالی‌ها به همراه تعداد هر توالی درست کند. (2 نمره)

6) حدر این قسمت حال، با استفاده از کدی که در سوال قبل نوشته‌اید، برنامه‌ای بنویسید که با استفاده از درختی که برای داده‌های Illumina Sequencing درست کرده‌اید، توالی‌هایی که در فایل clear.txt وجود دارند و در NGS.fasq.gz هم وجود دارند را print کند و همچنین تعداد حضورشان در داده‌ی NGS را هم print کند. (بنابراین این سوال و خروجیش کاملا با سوال3 یکسان است، فقط به جای الگوریتم naive، باید با چک کردن درخت حضور یا عدم حضور یک توالی را بررسی کنید). (1 نمره)