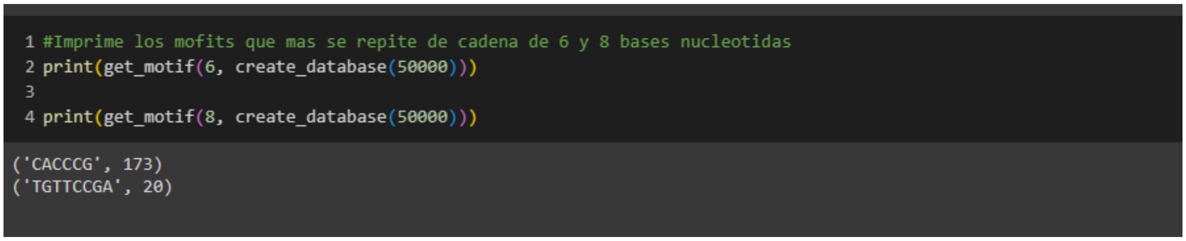
Workshop number 1 of systems analysis is based on making a fictitious database consisting of 50,000 genetic DNA sequences, where each DNA sequence has between 10 and 20 nucleotide bases. In the database the most repeated DNA sequences of 6 and 8 nucleotide bases (called motifs) are calculated, after which all the most repeated 6 and 8 base strands are printed.

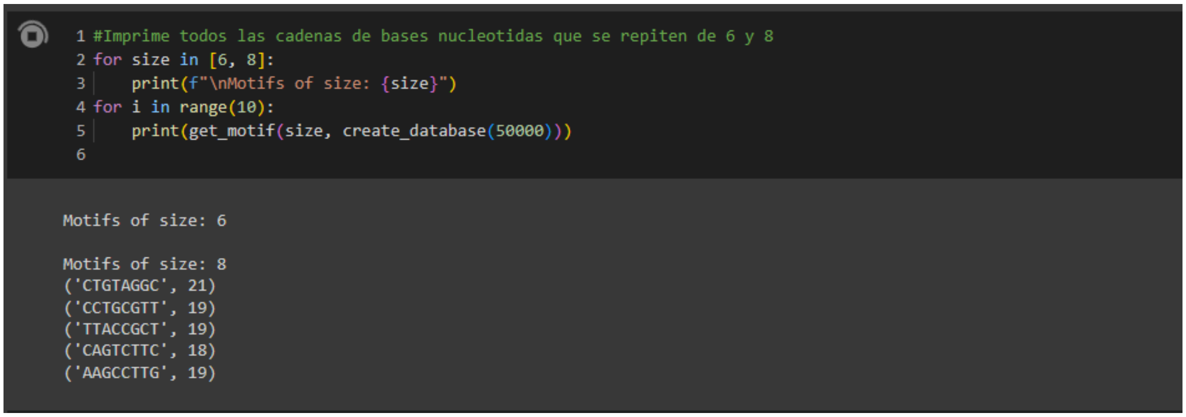
the results obtained were as follows.



```python
#Imprime los mofits que mas se repite de cadena de 6 y 8 bases nucleotidas
print(get_motif(6, create_database(50000)))

print(get_motif(8, create_database(50000)))
```

```
('CACCCG', 173)
('TGTTCCGA', 20)
```



```python
#Imprime todos las cadenas de bases nucleotidas que se repiten de 6 y 8
for size in [6, 8]:
    print(f"\nMotifs of size: {size}")
for i in range(10):
    print(get_motif(size, create_database(50000)))

```

```

Motifs of size: 6

Motifs of size: 8
('CTGTAGGC', 21)
('CCTGCGTT', 19)
('TTACCGCT', 19)
('CAGTCTTC', 18)
('AAGCCTTG', 19)
```

$$\sum_i p_i \log_2 \left(\frac{1}{p_i}\right)$$

Shannon's entropy equation

Using Shannon's entropy equation, we intend to calculate the DNA sequences that are less repeated. Substituting the formula with the procedure that we want to do, pi symbolizes the quantity of DNA sequences that have less frequency in the total of the 50000 DNA sequences, respecting a given condition, then we have to filter the DNA sequences that are less repeated from the 50000 that belong to the fictitious database.

the difficulties I have with the program is to know the condition to obtain the DNA sequences that are less repeated after using the Shannon entropy.