Count-Min Sketch (Simple Implementation)

Assignment 1 - Stream Mining

Feras Shamasna - NIQWQX

In [130]:

```
import numpy as np
import pandas as pd
import seaborn as sns
%matplotlib inline
```

In [131]:

```
stream = ['A','B','C','A','B','A','B','C','C','A','C','C','A','B']
```

In [132]:

```
#Hash_entries table
data = [[0,1,3,4,2],[1,0,2,4,1],[0,3,2,1,4]]
hashes = ['h1','h2','h3','h4','h5']
input_stream = ['A','B','C']
values_of_hashes = pd.DataFrame(data=data,index=input_stream,columns=hashes)
values_of_hashes.head()
```

Out[132]:

	h1	h2	h3	h4	h5
A	0	1	3	4	2
B	1	0	2	4	1
C	0	3	2	1	4

In [133]:

```
values = list(set(df.values.reshape(-1)))
values = np.arange(min(values),max(values)+1)
```

In [134]:

```
# Initialize Count_min Sketch
countMinSketch = np.zeros((len(hashes),len(values)))
countMinSketch = pd.DataFrame(data = countMinSketch, columns=values,index=hashes)
countMinSketch
```

Out[134]:

	0	1	2	3	4
h1	0.0	0.0	0.0	0.0	0.0
h2	0.0	0.0	0.0	0.0	0.0
h3	0.0	0.0	0.0	0.0	0.0
h4	0.0	0.0	0.0	0.0	0.0
h5	0.0	0.0	0.0	0.0	0.0

In [135]:

```
for char in stream:
    for hash_ in hashes:
        idx_to_inc = values_of_hashes.loc[char][hash_]
        countMinSketch.loc[hash_][idx_to_inc] += 1
countMinSketch
```

Out[135]:

	0	1	2	3	4
h1	10.0	4.0	0.0	0.0	0.0
h2	4.0	5.0	0.0	5.0	0.0
h3	0.0	0.0	9.0	5.0	0.0
h4	0.0	5.0	0.0	0.0	9.0
h5	0.0	4.0	5.0	0.0	5.0

In [136]:

```
# get the frequancies for all the letters:
chars_freq = {}
for char in stream:
    Hashes_values_of_char = []
    for hash_ in hashes:
        idx_to_inc = values_of_hashes.loc[char][hash_]
        Hashes_values_of_char.append(countMinSketch.loc[hash_][idx_to_inc])
    chars_freq[char] = min(Hashes_values_of_char)

for char,freq in chars_freq.items():
    print(f'The freq. of {char} is: {freq}')
```

```
The freq. of A is: 5.0
The freq. of B is: 4.0
The freq. of C is: 5.0
```