St Number: G47609664

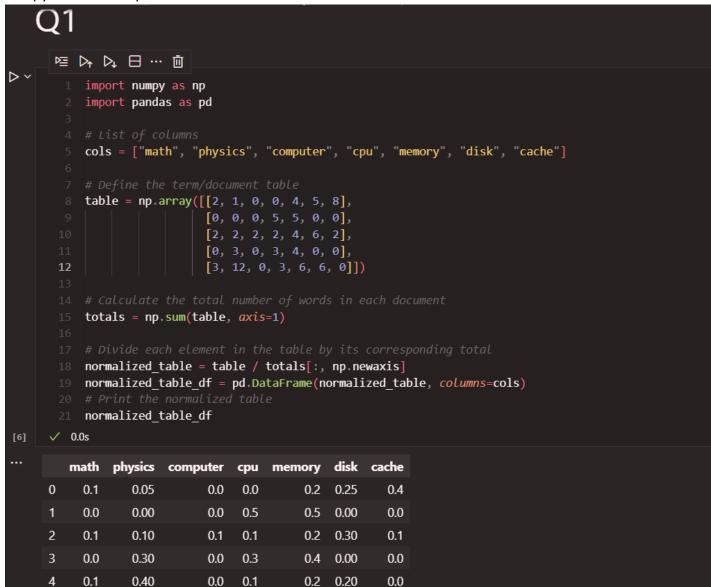
Exercise 6 Indexing

Given the following term/document table where values in table are actual term frequencies in each document:

Terms	math	physics	computer	cpu	memory	disk	cache
Doc1	2	1	0	0	4	5	8
Doc2	0	0	0	5	5	0	0
Doc3	2	2	2	2	4	6	2
Doc4	0	3	0	3	4	0	0
Doc5	3	12	0	3	6	6	0

1. Rewrite term/document matrix normalizing all documents by dividing by total number of words in each document

The python code to perform this:



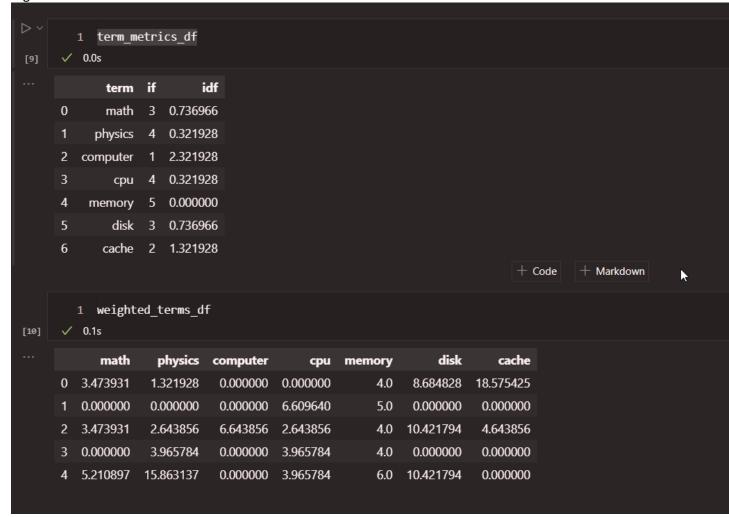
St Number: G47609664

2. Using original matrix calculate the IDF for Computer, Memory and cache showing values. Then modify original matrix using those to calculate weights.

The below formula was used to calculate the weights:

WEIGHT_{ii} = $TF_{ii} * [Log_2(n/IF_i) + 1]$

Below are the python code and its results:



3. Using original matrix calculate the Signal weight for physics, memory and cache showing each value. Then update matrix using those values to calculate weights.

The signal weights formula was used as follows:

Weight_{ik}=
$$TF_{ik} * [Log_2(TOTF_k) - TF_{ik}/TOTF_k Log_2(TF_{ik}/TOTF_k)]$$

Here is the Python code:

```
D v
        2 cols = ["math", "physics", "computer", "cpu", "memory", "disk", "cache"]
        5 table = np.array([[2, 1, 0, 0, 4, 5, 8],
                           [0, 0, 0, 5, 5, 0, 0],
                            [2, 2, 2, 2, 4, 6, 2],
                           [0, 3, 0, 3, 4, 0, 0],
                            [3, 12, 0, 3, 6, 6, 0]]
       12 total_num_terms = np.sum(table, axis=0)
       15 term_freqs = table / total_num_terms
       18 term_idfs = np.log2(term_freqs)
       21 term_idfs[np.isinf(term_idfs)] = 0
       24 info_val = (-1.0 * term_freqs) * term_idfs
       27 ave_info = np.sum(info_val, axis=0)
       30 signal_weights = table * (np.log2(total_num_terms) - ave_info)
       34 signal_weights_df = pd.DataFrame(signal_weights, columns=cols)
       35 signal_weights_df
[44]
     ✓ 0.0s
```

And here is the output:

NAME: Toghrul Tahirov St Number: G47609664 [44] 🗸 0.0s C:\Users\togru\AppData\Local\Temp\ipykernel_6028\558800666.py:20: RuntimeWarning: divide by zero encountered in log2 term_idfs = np.log2(term_freqs) physics computer disk cache math cpu memory 0 2.501396 2.765247 0.0 0.000000 8.890333 12.537997 20.8 0.000000 1 0.000000 0.0 8.892083 11.112916 0.000000 0.0 2 2.501396 5.530493 2.0 3.556833 8.890333 15.045597 5.2

0.000000

0.0

0.0

K

0.0 5.335250 8.890333

0.0 5.335250 13.335500 15.045597

3 0.000000

4 3.752095 33.182958

8.295740