Week 5: Filtering and Streaming

02807 Computational Tools for Data Science

For simplicity (not performance!), we recommend using the BitVector library (`https://pypi.org/project/BitVector/`) and MurmurHash (`https://pypi.org/project/mmh3/`) for Todays exercises.

For performance, implement the bitvector and hashing yourself using numba or cython and use a state-of-the-art universal hash function (`https://en.wikipedia.org/wiki/Universal_hashing$#$Avoiding_modular_arithmetic`). Different hash functions can be created by using different seeds.

The packages can be installed by:

```
pip3 install mmh3
pip3 install BitVector
```

Exercises

1 Bloom Filter Implement a Bloom Filter to filter good URLs from bad URLs. Use the provided UrlBloomFilter.py template and test data (can be found on the course page). To test your implementation using this, you must call (replace X with 1, 2 or 3):

```
python3 UrlBloomFilterTester.py bloom_goodX.dta
```

Your filter cannot have any false negatives, and must have a false positive rate of less than 5%.

2 Count-Min Sketch Implement a Count-Min Sketch for the $\varepsilon - HH$ problem. Test your count-min sketch on the provided data `dataX_CMS.dta`, $X = 1, 2, 3, 4, 5$, and measure the difference between estimates and the correct frequencies. No templates are provided for this exercise, so you must make it all yourself.