Homework 4

Due 19th April Midnight

1. You want to filter out spurious emails. You have a bloom filter with n=8 billion bits (1 GB main memory). You have m=1 billion good email addresses. For each email, you have k hash functions and each hash function uniformly selects a bit and sets it to 1. Note you don't need to implement or use bloom filter. This question is only about understanding its behavior.
    a. Plot the false positive rate as you vary k from 1 to 20. [5 points]
    b. Explain your observations analytically. [10 points]

2. To get your 2 datasets use this file <hw4.pyc>.
   https://onedrive.live.com/redir?resid=EF92560AE8680184!75647&authkey=!ANLeWUndLiE2WiE&ithint=file%2cpyc
   This is a .pyc file. Run it as "python hw4.pyc last-2-digits-rollNo" in python. Eg, if your roll number is CS11B046, run it as "python hw4.pyc 46". It will generate 2 files x.txt and y.txt. Each line of the files contains an item Id. The files end with an empty line. Use x and y as the datasets.
    a. Count the top-k frequent items using the Space-saving algorithm that stores 10000 items treating dataset 1 and 2 as the streams. Compute the accuracy for each dataset at k=100, 500, 1000, 5000. Are the results better in one dataset than the other? [10 points]
    b. Explain the observations in part a. Specifically, if the results are better in one dataset, dig in to the datasets and identify why so using statistics or any other reasoning.[10 points]
    c. How can you provide an explanation like you did in part b when you are unable to store the stream (unlike dataset 1 and 2)? Back up your answer with actual experimental evidence. Specifically, treat the two datasets as streams.[10 points]