DSC 40B - Discussion 01

Problem 1.

a) Given a collection of 200 billion unique objects, design a Bloom filter with a 5% false positive rate by choosing the size c of the bit array and the number of hash functions, k.

> **Solution:** We use the formulas from lecture to choose c and then k. To choose c, we evaluate:
>
> $$c = -n \ln \varepsilon / (\ln 2)^2,$$
>
> where $\varepsilon = .05$ and $n = 200 \times 10^9$. We find $c = 1.25 \times 10^{12}$, or roughly 1.25 trillion.
>
> Next, we find k using the formula $k = \frac{c}{n} \ln 2$. We get approximately $k = 4.23$. We will therefore use five hash functions, since using only four would yield a false positive rate higher than 5%.

b) How much memory will your Bloom filter require, approximately?

> **Solution:** We use a single bit array of size $c = 1.25$ trillion bits. This comes out to roughly 156 gigabytes.

c) Suppose your machine has 32 gigabytes of memory. What false positive rate must you accept in order to fit the Bloom filter in memory?

> **Solution:** 32 gigabytes is 2.56×10^{11} bits; this is the largest c can be. We then work backwards from the formula $c = -n \ln \varepsilon / (\ln 2)^2$. We find:
>
> $$\ln \varepsilon = -c(\ln 2)^2 / n$$
>
> Plugging in the value of c above and the value of n, we find:
>
> $$\ln \varepsilon = -.615$$
>
> Exponentiating, we get: $\varepsilon = .54$.

d) A Bloom filter using 32 gigabytes of memory has a false-positive rate similar to flipping a coin. In what ways is the Bloom filter better than a coin?

> **Solution:** Both the Bloom filter and a coin have a false positive rate of around 50%, true, but the coin has a *false negative* rate of 50%, while the Bloom filter has a false negative rate of 0% (there are no false negatives).