CS 170 DIS 13

Released on 2019-04-23

1 Document Comparison with Streams

You are given a document A and then a document B, both as streams of words. Find a streaming algorithm that returns the degree of similarity between the words in the documents, given by $\frac{|I|}{|U|}$, where I is the set of words that occur in both A and B, and U is the set of words that occur in at least one of A and B.

Clearly explain your algorithm and briefly justify its correctness and memory usage (at most $\log(|A| + |B|)$). Can we achieve accuracy to an arbitrary degree of precision? That is, given any $\epsilon > 0$ can we guarantee that the solution will always be within a factor of $1 \pm \epsilon$ with high probability?

2 Lower Bounds for Streaming

- (a) Consider the following simple 'sketching' problem. Preprocess a sequence of bits b_1, \ldots, b_n so that, given an integer i, we can return b_i . How many bits of memory are required to solve this problem exactly?
- (b) Given a stream of integers $x_1, x_2, ...$, the majority element problem is to output the integer which appears most frequently of all of the integers seen so far. Prove that any algorithm which solves the majority element problem exactly must use $\Omega(n)$ bits of memory, where n is the number of elements seen so far.

3 Universal Hashing

Let [m] denote the set $\{0, 1, ..., m-1\}$. For each of the following families of hash functions, determine whether or not it is universal. If it is universal, determine how many random bits are needed to choose a function from the family.

(a) $H = \{h_{a_1,a_2} : a_1, a_2 \in [m]\}$, where m is a fixed prime and

$$h_{a_1,a_2}(x_1,x_2) = a_1x_1 + a_2x_2 \mod m$$

Notice that each of these functions has signature $h_{a_1,a_2}:[m]^2\to [m]$, that is, it maps a pair of integers in [m] to a single integer in [m].

- (b) H is as before, except that now $m=2^k$ is some fixed power of 2.
- (c) H is the set of all functions $f:[m] \to [m-1]$.