Computing at Scale in Machine Learning: Distributed computing and algorithmic approaches (14038)

B-TU Cottbus-Senftenberg · WiSe 2023
Prof. Dr. Alexander Schliep, Nathalie Gocht, M.Sc. and Aleksandra Khatova, M.Sc.

Problem set 3 from November 8, 2023 · Due on November 17, 2023

Problem 1 (Deleting items from a bloom filter, 4 Points)**.** Standard Bloom filters do not support deletion of previously inserted items. Please analyze the following proposed deletion operation: For a deletion of item x we set the $h_1(x), ..., h_k(x)$ bits for item x to zero. Assume that the Bloom filter B has $|B| = b$ bits, that the k hash function $h_i(x)$ are uniform, and that a total of $n + 1$ items have been inserted into B.

a) (2 pt) What is the probability that deleting one item from B does not cause false negative results in `query()` for any of the remaining n items in B?

b) (2 pt) What is the expected number of bits which are set due to insert of the first n items and which are unset by the delete of item $(n + 1)$? This is the same as asking the expected number of collisions when inserting the $(n + 1)$-st item.

Problem 2 (Bloom Filter Implementation, 6 Points)**.** You are counting frequencies of 1 billion items. You observe that 83.64% of the items appear once, 9.78% appear twice, 3.11% appear three times, and 3.47% appear four or more times. You're only interested in the frequencies of items appearing four or more times.

a) Suggest an efficient solution using Bloom filters to accelerate the counting; please determine k, the number of hash functions, and b, the size of Bloom filters, when you assume that there are a total of $n = 1$ billion items in the input. Choose error rates to obtain accurate counts.

b) Implement your solution using `PyBloom_live` (a maintained version of PyBloom).

Note: Please hand in:

- A PDF containing:
  - the answers to problem 1)
  - the answer to 2a)
- A Python script named `problem3_2.py`

Python code must be handed in as '.py'. Code submission in PDF will be graded with 0 points. Your code files can be accompanied with PDF.

You are allowed to use the example solutions provided where they might be helpful for answering question on this problem set.

Please feel free to discuss the problems and approaches to solving them with other students. However *you* should understand results of discussions well enough to write up the solution by yourself and possibly explain the solution. When you do collaborate, please make other's contributions clear. Present your solutions concisely.