Filtr Blooma

10 grudnia 2018

Opis pliku z zadaniami

Wszystkie zadania na zajęciach będą przekazywane w postaci plików `.pdf`, sformatowanych podobnie do tego dokumentu. Zadania będą różnego rodzaju. Za każdym razem będą one odpowiednio oznaczone:

- Zadania do wykonania na zajęciach oznaczone są symbolem $\triangle$ – nie są one punktowane, ale należy je wykonać w czasie zajęć.
- Punktowane zadania do wykonania na zajęciach oznaczone są symbolem $\diamond$ – należy je wykonać na zajęciach i zaprezentować prowadzącemu, w wypadku nie wykonania zadania w czasie zajęć lub nieobecności, zadania staje się zadaniem do wykonania w domu ($\star$).
- Zadania do wykonania w domu oznaczone są symbolem $\star$ – są one punktowane, należy je dostarczyć w sposób podany przez prowadzącego i w wyznaczonym terminie (zwykle przed kolejnymi zajęciami).
- Zadania programistyczne można wykonywać w dowolnym języku programowania, używając jedynie biblioteki standardowej dostępnej dla tego języka.

1 Filtr Blooma *10p.*$\diamondsuit$

Treść

Zadanie polega na zaimplementowaniu i przebadaniu filtra Blooma, przy założeniu, że obiekty pochodzą z dziedziny liczb naturalnych w podanym zakresie.

Filtr Blooma jest probabilistyczną strukturą, która w sposób efektywny pamięciowo przechowuje (przybliżoną) informację o przynależności obiektów do zbioru S, będącego podzbiorem bardzo licznej dziedziny. Filtr Blooma jest parametryzowany poprzez liczbę funkcji mieszających k oraz rozmiar m, który musi być odpowiednio większy od zakładanej liczności n zbioru S.

Należy zaimplementować metodę dodawania elementu do zbioru oraz sprawdzania, czy dany element znajduje się w zbiorze. Porównaj filtr Blooma z dokładną reprezentacją zbioru (np. tabelą mieszającą). Policz dla danej instancji filtra Blooma liczbę obiektów prawdziwie pozytywnych, prawdziwie negatywnych, fałszywie pozytywnych oraz fałszywie negatywnych. Sprawdź, czy zaimplementowany filtr Blooma spełnia gwarancje teoretyczne oraz zweryfikuj jego złożoność pamięciową.

Do implementacji filtra Blooma można wykorzystać następującą rodzinę funkcji mieszających H. Niech $h \in H : \{1, \ldots, M\} \to \{0, \ldots, m\}$. Znajdź liczbę pierwszą $p > M$. Następnie zdefiniuj, dla każdego $a \in \{1, \ldots, p-1\}$ oraz dla każdego $b \in \{0, \ldots, p-1\}$, funkcję:

$$g_{a,b}(x) = ax + b \mod p\,.$$

Dla każdej takiej funkcji g zdefiniuj następującą funkcję mieszającą:

$$h_{a,b}(x) = g_{a,b}(x) \mod m\,.$$

Wartości a i b należy wylosować w sposób jednostajny z podanego przedziału. Powyższa rodzina H funkcji mieszających spełnia następującą własność:

$$\Pr_{h \in H}(h(x) = h(y)) \leq \frac{1}{m}\,,$$

dla każdej pary różnych $x, y \in \{1, \ldots, M\}$.

Punktacja:

- poprawna implementacja: 5p.
- efektywna implementacja: 3p.
- weryfikacja własności teoretycznych: 2 p.