Sygnatury minhashowe

7 stycznia 2019

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 Dokładność sygnatur minhashowych *10p.*◇

Treść

Wykorzystując dane dotyczące problemu MSDC zweryfikuj dokładność sygnatur minhashowych. Dla pierwszych n użytkowników (np. $n = 100$) utwórz sygnatury minhashowe o długość d (np. $d = 100$). Następnie policz przybliżoną i prawdziwą wartość podobieństwa Jaccarda dla wszystkich par użytkowników (biorąc pod uwagę pierwszych n użytkowników).

Zadanie: Policz pierwiastek z błędu średniokwadratowego przybliżenia nie biorąc pod uwagę par użytkowników, dla których podobieństwo Jaccarda wynosi 0 (zauważ, że w takim przypadku przybliżenie obliczone na sygnaturze minhashowej będzie również równe 0). Przeprowadź eksperymenty dla różnej liczby użytkowników n oraz długości sygnatury d.

Punktacja:

- Realizacja zadania dowolnym podejściem (błąd dla $n = 100$ powinien wynieść ok. 0.018): 6p.
- Analiza wyników ze względu na rożne wartości n oraz d: 4p.