Wyszukiwanie najbliższych sąsiadów metodą LSH

14 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 Wyszukiwanie najbliższych sąsiadów metodą LSH

10p. ◇

Treść

Zaimplementuj algorytm wyszukiwania przybliżonych najbliższych sąsiadów dla odległości Jaccarda wykorzystując metodę LSH i technikę minhash. Do wykonania tego zadania wykorzystaj dane i kod z poprzednich dwóch zajęć. W celu zweryfikowania podejścia należy porównać jego wynik z wynikiem wyszukiwania dokładnego. Można to wykonać obliczając czułość (ang. *recall*) dla zadanego poziomu podobieństwa, tj. z listy dokładnych i przybliżonych sąsiadów wybieramy tylko tych o podobieństwie większym lub równym od zadanego, a następnie sprawdzamy jaką część wybranych dokładnych sąsiadów stanowią wybrani przybliżeni sąsiedzi.

Zadanie: Znajdź przybliżonych najbliższych użytkowników dla wszystkich użytkowników przy zadanym poziomie podobieństwa. Porównaj wynik z dokładnym wyszukiwaniem licząc czułość przy zadanym poziomie podobieństwa. Pokaż jak można sterować podejściem LSH w celu wyszukiwania użytkowników o różnym poziomie podobieństwa.

Punktacja:

- Wyszukiwanie przybliżonych najbliższych użytkowników dla wszystkich użytkowników przy zadanym poziomie podobieństwa: 5p.
- Sterowanie podejściem LSH w celu wyszukiwania użytkowników o różnym poziomie podobieństwa wraz z analizą wyników: 2p.
- Weryfikacja podejścia (policzenie czułości przy zadanym poziomie podobieństwa) z listą 100 pierwszych użytkowników: 1p.
- Weryfikacja podejścia (policzenie czułości przy zadanym poziomie podobieństwa) dla 10 000 pierwszych użytkowników: 2p.