Wyszukiwanie najbliższych sąsiadów

15 czerwca 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 Dokładne wyszukiwanie najbliższych sąsiadów *10p.*◇

Treść

Wykorzystując dane dotyczące problemu MSDC rozwiąż problem wyszukiwania najbliższych użytkowników. Podobieństwo pomiędzy użytkownikami należy określić używając współczynnika Jaccarda na zbiorach odsłuchanych utworów muzycznych.

Podobieństwo Jaccarda pomiędzy dwoma zbiorami jest zdefiniowane jako iloraz mocy części wspólnej zbiorów i mocy sumy tych zbiorów:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

gdzie A i B są zbiorami.

Rozważmy następujący przykład. Niech zbiór $S_1 = \{s_1, s_3, s_4\}$ odpowiada pierwszemu użytkownikowi u_1, a zbiór $S_2 = \{s_2, s_3, s_6\}$ użytkownikowi drugiemu u_2. W postaci tabelarycznej możemy te zbiory zapisać następująco:

użytkownik	s_1	s_2	s_3	s_4	s_5	s_6	s_7
u_1	1	0	1	1	0	0	0
u_2	0	1	1	0	0	1	0

Dla powyższych danych współczynnik Jaccarda wynosi:

$$J(A,B) = \frac{1}{5}$$

Zadanie: Dla stu pierwszych użytkowników, pojawiających się w tabeli `facts`, znajdź stu najbliższych sąsiadów o podobieństwie Jaccarda większym od 0 (w całym zbiorze). Następnie postaraj się tak zmodyfikować algorytm, aby możliwe było znalezienie najbliższych użytkowników dla wszystkich użytkowników w sposób dokładny w sensownym czasie (ok. 2-5 godzin przy algorytmie sekwencyjnym). W tym celu należy zauważyć specyficzną własność przetwarzanych danych, która pozwala na znaczące przyśpieszenie obliczeń.

Punktacja:

- Realizacja wyszukiwanie dokładnych najbliższych sąsiadów dla 100 pierwszych użytkowników dowolnym podejściem: 5p.
- Optymalizacja kodu pod względem szybkości działania: 3p.
- Znalezienie stu najbliższych użytkowników dla wszystkich użytkowników: 2p.
- Algorytm liniowy wyszukiwanie k najbliższych sąsiadów dla pojedynczego użytkownika: bonus 2p.

2 Dokładność sygnatur minhashowych *5p.*◇

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 wśród 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): 3p.
- Analiza wyników ze względu na rożne wartości n oraz d: 2p.

3 Przybliżone wyszukiwanie najbliższych sąsiadów

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 zadań. W celu zweryfikowania podejścia należy porównać jego wynik z wynikiem wyszukiwania dokładnego. Można tego dokonać 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.