Wyszukiwanie najbliższych sąsiadów

16 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.
- 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).

1 Dokładne i przybliżone 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 kolejnym kroku zaimplementuj algorytm wyszukiwania przybliżonych najbliższych sąsiadów dla odległości Jaccarda wykorzystując metodę LSH i technikę minhash. 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. Pokaż jak można sterować podejściem LSH w celu wyszukiwania użytkowników o różnym poziomie podobieństwa.

Punktacja:

- Realizacja wyszukiwanie dokładnych najbliższych sąsiadów dla 100 pierwszych użytkowników dowolnym podejściem: 3p.
- Optymalizacja kodu pod względem szybkości działania: 2p.
- Znalezienie stu najbliższych użytkowników dla wszystkich użytkowników: 2p.
- Wyszukiwanie przybliżonych najbliższych sąsiadów z wykorzystaniem LSH i techniki minhash dla wszystkich użytkowników przy zadanym poziomie podobieństwa: 2p.
- Sterowanie podejściem LSH w celu wyszukiwania użytkowników o różnym poziomie podobieństwa wraz z analizą wyników: 1p.
- Weryfikacja podejścia (policzenie czułości przy zadanym poziomie podobieństwa) z listą 100 pierwszych użytkowników: bonus 1p.
- Weryfikacja podejścia (policzenie czułości przy zadanym poziomie podobieństwa) dla 10 000 pierwszych użytkowników: bonus 2p.
- Algorytm liniowy wyszukiwanie k najbliższych sąsiadów dla pojedynczego użytkownika: bonus 2p.

Do implementacji funkcji minhashowych 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 $x, y \in \{1, \ldots, M\}$.