KMeans Colors

23 listopada 2017

Grzegorz Borkowski

1. WSTĘP

Zadanie zostało zaimplementowane w Pythonie 3.6.1
Do raportu został dołączony Jupyter Notebook z kodem źródłowym programu.

2. ZADANIE

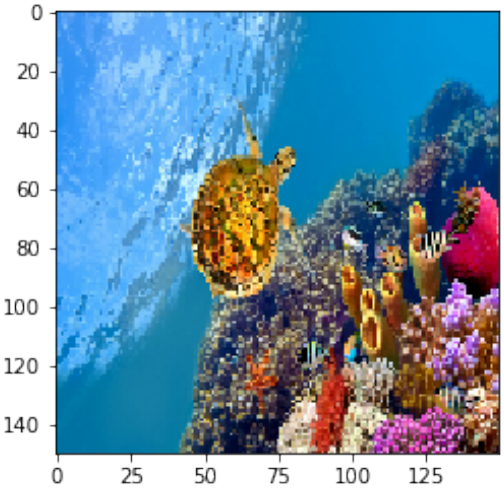
2.1 Treść zadania

Wybierzmy ładne, kolorowe zdjęcie. Potraktujmy każdy jego piksel jako obserwację w przestrzeni 3-D (po jednym wymiarze na każdy z kolorów). Zdecydujmy czy usuwamy ze zbioru duplikaty (piksele o takich samych wartościach RGB) - nasz wybór wpłynie na finalny wynik. Wykonajmy na takim zbiorze klasteryzację k-means, z następującymi założeniami: * jako środków klastrów używamy istniejące elementy zbioru, a nie ich średnie (czyli jest to w praktyce k-medians) - nie chcemy znaleźć kolorów, które nie wystąpiły na zdjęciu; * dobieramy wartość stałej k używając dowolnej zaproponowanej przez siebie metody.

Na koniec prezentujemy uzyskaną paletę, oraz wizualizujemy samą klasteryzację (np. rzutujemy punkty ze zbioru na 2D używając PCA, każdy z nich malujemy na pierwotny kolor, tło danego klastra malujemy na kolor go reprezentujący). W dyskusji zastanawiamy się, jak k-medians spisał się w tym zadaniu i czy wystąpiły problemy omawiane na zajęciach.

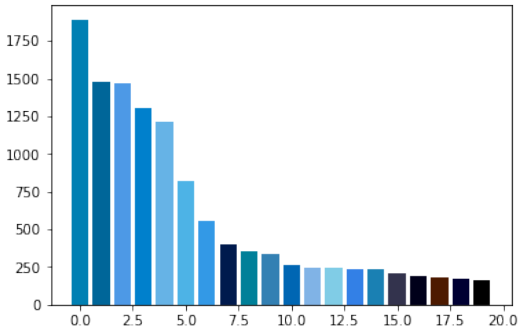
2.2 Rozwiązanie

Nie usuwam duplikatów ze zbioru wejściowego.

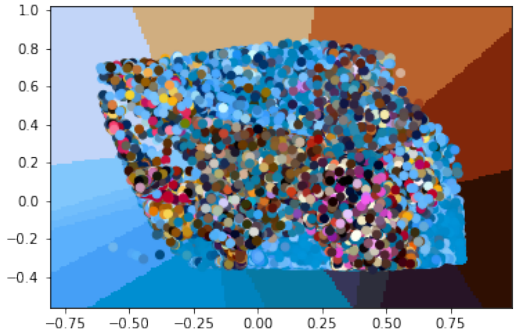


Rysunek 1. Zdjęcie wejściowe

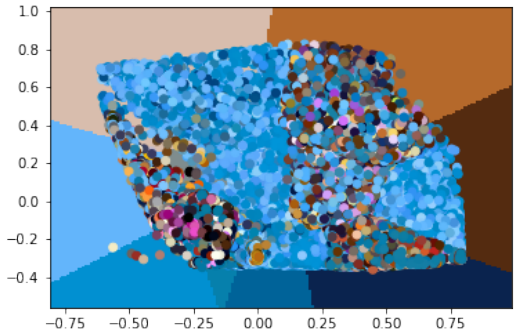
Dla k=5, metryka Silhouette'a: 0.2444
Dla k=10, metryka Silhouette'a: 0.3144
Dla k=20, metryka Silhouette'a: 0.226



Rysunek 2. Histogram wystąpień pikseli w dwudziestu najczęściej występujących kolorach

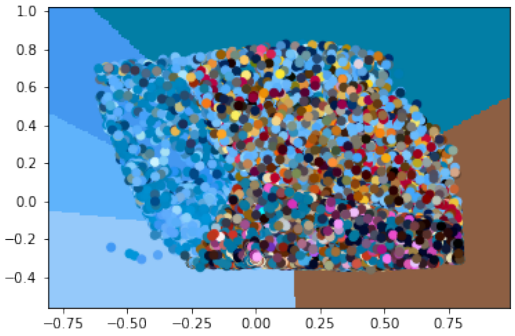


Rysunek 3. Wynik klastrowania dla k=20

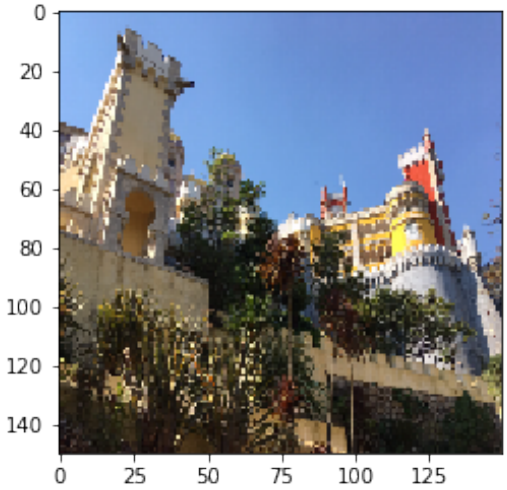


Rysunek 4. Wynik klastrowania dla k=10

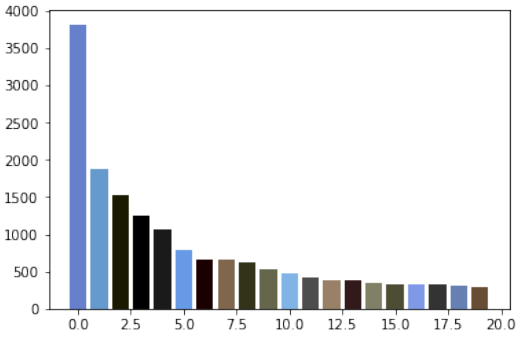
Powtórzyłem eksperyment dla innego zdjęcia wejściowego.



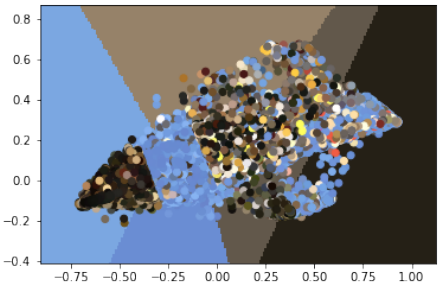
Rysunek 5. Wynik klastrowania dla k=5



Rysunek 6. Drugie zdjęcie testowe



Rysunek 7. Histogram kolorów dla drugiego zdjęcia



Rysunek 8. Klasteryzacja dla drugiego zdjęcia z k=5

3. WNIOSKI

(1) KMeans słabo radzi sobie jeżeli istnieją klastry dużo bardziej liczne niż inne
(2) Klastry nie koniecznie odpowiadają najczęściej występującym kolorom (dla pierwszego zdjęcia i k=10 występują klastry w kolorach brązowym, jasnobrązowym i różowym, pomimo, że te kolory nie należą do dziesięciu najpopularniejszych)
(3) KMeans nie nadaje się do wizualizacji palety kolorów lub popełniłem błąd w implementacji
(4) Piksele nie odpowiadają kolorom klastrów do których należą
(5) Indeks Silhouetta słabo sobie poradził w tym zadaniu, dla pierwszego zdjęcia wyniki to 0.2-0.3, wynik nie wskazuje ogromnej liczby punktów źle przypisanych (tak byłoby gdyby wynik był ujemny)