Przetwarzanie i analiza dużych zbiorów danych 2020/21

Prowadzący: mgr inż. Rafał Woźniak środa, 11:45

Piotr Wardęcki	234128	234128@edu.p.lodz.pl
Paweł Galewicz	234053	234053@edu.p.lodz.pl
Bartosz Jurczewski	234067	234067@edu.p.lodz.pl

Zadanie 3

1. Cel zadania

Celem zadania była implementacja algorytmu k-średnich z uwzględnieniem dwóch miar - Euklidesowej oraz Manhattan - dla dwóch sposobów generowania centrów - losowego oraz maksymalnie oddalonych od siebie punktów (według odległości euklidesowej).

Dla każdej iteracji należało obliczyć funkcje kosztu $\phi(i)$ oraz $\psi(i)$, wygenerować wykresy oraz obliczyć procentową zmianę kosztu po 10 iteracjach algorytmu dla obydwu miar odległości z wskazaniem, które z dwóch początkowych rozmieszeń skupień pozwoliło uzyskać lepsze rezultaty.

Miara euklidesowa:

$$\text{odległość: } ||a-b|| = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2} \tag{1}$$

$$\text{koszt: } \phi = \sum_{x \epsilon X} \min_{c \epsilon C} ||x-c||^2 \tag{2}$$

Miara Manhattan:

$$\text{odległość: } |a-b| = \sum_{i=1}^{d} |a_i - b_i| \tag{3}$$

$$\text{koszt: } \psi = \sum_{x \epsilon X} \min_{c \epsilon C} |x-c| \tag{4}$$

2. Wprowadzenie

3. Opis implementacji

Do wykonywania zadania niezbędna była instancji *Apache Spark*. Aby ograniczyć liczbę zainstalowanych środowisk skorzystaliśmy z odpowiedniego obrazu dla Dockera [1], który zawierał także *Jupyter Notebook*, *Python* oraz *Miniconda*. Dodatkowo aby ułatwić tworzenie środowiska do kolejnych zadań i między naszymi komputerami skorzystaliśmy z narzędzia *Docker Compose* (nasz plik [2]).

4. Apache Spark

Apache Spark posłużył nam do wczytania pliku oraz przeprowadzenia na nim wszystkich niezbędnych operacji iteracyjnych w celu wyszukania rekomendowanych znajomości dla użytkownika.Bardzo pomocna okazała się funkcja groupByKey, która polega na łączeniu danych w sposób key-value. Dla każdego klucza pobierana jest wartość w sposób iteracyjny. Dodatkowo połączyliśmy inne wbudowane metody Sparka z samodzielnie zaimplementowaną logiką, w celu osiągnięcia zamierzonego kryterium zadania.

5. Wyniki

6. Wnioski

- Wniosek 2
- Wniosek 1

Bibliografia

[1] *Jupyter Notebook Python, Spark Stack* `https://hub.docker.com/r/jupyter/pyspark-notebook`
[2] *Plik Docker Compose do zadania 2* `https://github.com/jurczewski/PiADZD/blob/master/zad2/docker-compose.yml`