Przetwarzanie i analiza dużych zbiorów danych - zadanie 4

Paweł Ciupka 234048

Bartosz Łuniewski 234086

Filip Woźniak 234131

Opis

Algorytm został zaimplementowany przy użyciu języka Python oraz oprogramowania do analizy danych Apache Spark. Poniżej została przedstawiona lista kroków algorytmu:

1. Usunięcie przedmiotów, które wystąpiły w mniej niż 100 sesjach.
2. Znalezienie wszystkich par i trójek przedmiotów.
3. Obliczono wartości ufności dla wszystkich znalezionych krotek.
4. Przesłanie wyników do plików.

Do obliczenia wartości ufności został wykorzystany wzór:

$$supp(X) = \frac{|\{t \in T : X \subseteq t\}|}{|T|} \quad conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

,

gdzie *supp* oznacza wsparcie, a *conf* oznacza ufność.

Wyniki dla dwójek

<DAI93865> <FRO40251> <1.0>

<GRO85051> <FRO40251> <0.999176276771005>

<GRO38636> <FRO40251> <0.9906542056074766>

<ELE12951> <FRO40251> <0.9905660377358491>

<DAI88079> <FRO40251> <0.9867256637168141>

Wyniki dla trójek

<DAI23334> <ELE92920> <DAI62779> <1.0>

<DAI31081> <GRO85051> <FRO40251> <1.0>

<DAI55911> <GRO85051> <FRO40251> <1.0>

<DAI62779> <DAI88079> <FRO40251> <1.0>

<DAI75645> <GRO85051> <FRO40251> <1.0>