234073
Numer indeksu
Aleksandra Kowalczyk
Imię i nazwisko

234102
Numer indeksu
Zbigniew Nowacki
Imię i nazwisko

234106
Numer indeksu
Karol Podlewski
Imię i nazwisko

Kierunek Informatyka Stosowana
Stopień II
Specjalizacja Data Science
Semestr 2

Data oddania 2 grudnia 2020

Przetwarzanie i analiza dużych zbiorów danych

Zadanie 4

1 Cel zadania

Celem zadania było stworzenie programu, który implementuje reguły asocjacyjne za pomocą algorytmu A-priori. Program powinien – na podstawie ostatnio przeglądanych przez użytkowników przedmiotów – identyfikować wszystkie pary i trójki przedmiotów, które wystąpiły wspólnie w ramach pojedynczej sesji co najmniej 100 razy, a następnie obliczyć ufność dla stworzonych reguł asocjacyjnych.

Miary istotności:

$$\text{wsparcie: } supp(X) = \frac{|\{t \in T : X \subseteq t\}|}{|T|} \tag{1}$$

$$\text{ufność: } conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{2}$$

Gdzie X i Y to zbiory elementów transakcji, t oznacza transakcję, a T jest zbiorem transakcji.

2 Opis implementacji

Do zaimplementowania zadania wykorzystano język Python 3 oraz API PySpark pozwalające wykorzystać możliwości Apache Spark w pythonowym kodzie. Wszystkie funkcje zostały zaimplementowane w programie, nie korzystano z zewnętrznych bibliotek.

3 Uzyskane wyniki

Pełne wyniki zostały przedstawione w plikach `result_doubles.txt` oraz `result_triples.txt`. Wyniki zostały posortowane malejąco po poziomie ufności oraz leksykograficznie rosnąco.

5 reguł o największym współczynniku ufności zostało przedstawionych w tabelach 1 i 2.

Tabela 1: Reguły asocjacyjne dla dwójek z najwyższym stopniem zaufania

Poprzednik	Następnik	Zaufanie
DAI93865	FRO40251	1,0
GRO85051	FRO40251	0,999176276771005
GRO38636	FRO40251	0,990654205607477
ELE12951	FRO40251	0,990566037735849
DAI88079	FRO40251	0,986725663716814

Tabela 2: Reguły asocjacyjne dla trójek z najwyższym stopniem zaufania

Poprzednik		Następnik	Zaufanie
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

4 Analiza wyników

- Produkt o identyfikatorze `FRO40251` pojawia się we wszystkich regułach asocjacyjnych o wysokiej ufności dla dwójek oraz w 4 na 5 reguł dla trójek.
- Trójki przedmiotów posiadają więcej reguł o wysokiej ufności – dla dwójek jej wartość spada o wiele szybciej pomimo ogólnie większej (co widoczne jest w plikach) liczby reguł.