Peti domaći zadatak iz Tehnika i metoda analize podataka

Danilo Veljović, broj indeksa 1120

Opis skupa podataka

Skup podataka korišćen za testiranje implementiranog algoritma je skup podataka o sertifikovanim biznisima sa Small Business Services sajta. Dataset sadrži informacije o tome ko su vlasnici biznisa, tj kog su pola, kog su manjinskog ili većinskog porekla, na kojoj lokaciji se nalaze i koliko ima takvih biznisa na toj lokaciji.
Dataset sadrži 1147 instanci.

Način implementacije algoritma

Ključni deo algoritma je dat u funkciji runApriori() od linije 81 do linije 89. Najpre se generiše kandidat skup, nadovezivanjem podskupa iz prethodnog koraka. Zatim se izvodi testiranje i prunning tog skupa, ako se tu nalazi neki ne-česti (infrequent) podskup. Na kraju se računa najčešći skup pribavljanjem onih elemenata koji zadovoljavaju minimalnu potporu (minimal support). Slika 1 ima isečak iz koda koji radi upravo ovo (u kodu linije 82 - 89)

```
largeSet[k - 1] = currentLSet
    currentLSet = joinSet(currentLSet, k)
    currentCSet = returnItemsWithMinSupport(
       currentLSet, transactionList, minSupport, freqSet
    )
```

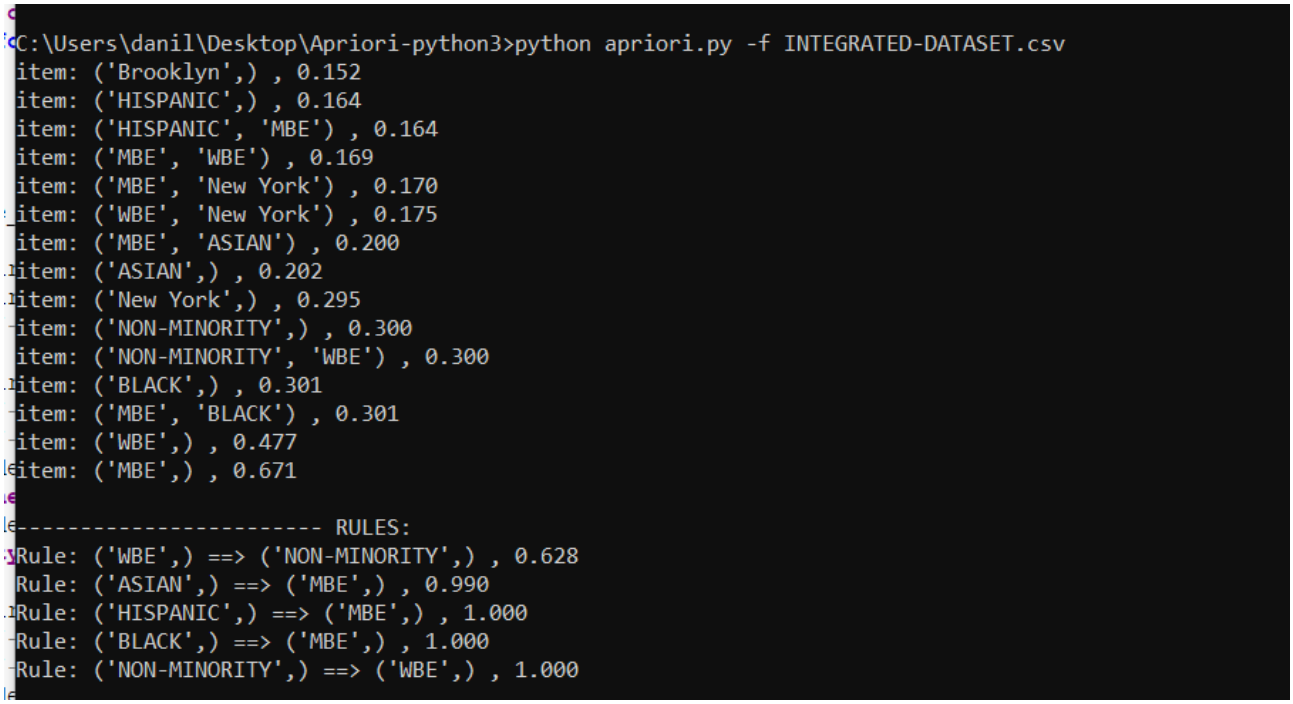
Sliika 1 – Srž Apriori algoritma

Generisanje itemseta (skupova) se radi funkcijom joinSet() kojom se kreira unija svih elementa koji su odredjene dužine.

Funkcijom returnItemsWithMinSupport() se određuje i vraća podskup elementa skupa čija je potpora veća od unapred zadaate minimalne potpore. Potpora se računa formulom na slici 2.

$$support(I) = \frac{Number\ of\ transactions\ containing\ I}{Total\ number\ of\ transactions}$$

Na kraju algoritma se finalni skup elemenata još jednom filtrira tako da se sad računa poverenje za svaki element podskupa i ako je vrednost nekog elementa veća od minimalne zadate vrednosti, taj element će biti vraćen. Na slici 3 je data izlazna vrednost algoritma.



Slika 3 – Rezultat izvršenja Apriori algoritma

Prvi elementi izlaza su vezani za sve elemente izlaznog skupa i odgovarajuće vrednosti njihovig potpora.
Drugi deo rezultata (od RULES podnaslova) vezan je direktno za pravila koja su zaključena i za njihov nivo poverenja.

U programu se može zadati i minimalni nivo poverenja, kao minimalni nivo potpora. Defaultne vrednosti su *minSupport* = 0.15 i *minConfidence* = 0.6.