BigData architektúrák és elemző módszerek GY.

Spark ZH

Megjegyzések:

Az adathalmazok megtalálhatóak a weboldalamon:

https://vargadaniel.web.elte.hu/kurzusok/bdgy23241.html

Az RDD feladatokat Spark RDD használatával kell megoldani, míg a DataFrame feladatokat DataFrame használatával, Spark vagy SQL lekérdezésekkel (minden feladat egy lekérdezés).

Spark RDD feladatok

Feladat 1. (2 pont)

A vasarlasok.txt egy kiskereskedelmi áruházban megvásárolt termékeket tartalmazzák. Minden sor egy vásárló kosarában lévő termékeket sorolja fel, vesszővel elválasztva. Határozd meg azt a három terméket, amelyekből a legtöbbet vásárolták és add meg hányat. Figyelj oda arra, hogy az adathalmazban kis és nagy betűk is előfordulhatnak. Azaz előfordulhat a "rizs" és a "Rizs" termék is. Ezeket azonos termékeknek kell tekintened.

Egy lehetséges kimenet: ('rizs', 120), ('narancs', 34), ('hal', 32)

Feladat 2. (2 pont)

Az online_retail_data.csv egy webáruház eladásait tartalmazza. Határozd meg, hogy melyik terméket melyik termékkel vásárolják gyakran együtt Franciaországban („France”). Azaz melyek a leggyakrabban előforduló termékpárok. Add meg a párok előfordulásának a számát és a 3 legtöbbször előfordulót ad vissza.

Egy lehetséges kimenet: [(('POSTAGE', 'RABBIT NIGHT LIGHT'), 130), (('POSTAGE', 'RED TOADSTOOL LED NIGHT LIGHT'), 124), (('PLASTERS IN TIN CIRCUS PARADE ', 'POSTAGE'), 116)]

Spark DataFrame feladatok

Az alábbi feladatok megoldásához a menu.csv és MenuCategory.csv fájlokat kell beolvasni. Az adathalmazok gyorséttermi menükről és azok kategóriáiról tartalmaz információkat.

Feladat 3. (1 pont)

Melyek azok az ételek, amelyek meghaladják az ajánlott napi zsír bevitelt? (Total Fat (% Daily Value))
Elvárt oszlopok: [Item]

Feladat 4. (1 pont)

Melyik ételnek van a maximális Sugars értéke?
Elvárt oszlopok: [Item, Sugars]

Feladat 5. (1 pont)

Hány elem van kategóriánként? Rendezzük csökkenő sorrendbe és adjuk meg a kategóriák nevét is.
Elvárt oszlopok: [Name, Cnt]

Feladat 6. (1 pont)

Átlagosan mennyi kalóriát tartalmaznak az egyes kategóriák? Adjuk meg a kategória nevét és rendezzük átlag alapján csökkenő sorrendbe. A 'Coffee and Tea' és a 'Beverages' kategóriákat ne vegyük figyelembe.
Elvárt oszlopok: [Name, AvgCal]