BigData architektúrák és elemző módszerek GY.

Spark ZH

Megjegyzések:

- Az RDD feladatokat Spark RDD használatával kell megoldani, míg a DataFrame feladatokat DataFrame használatával, Spark vagy SQL lekérdezésekkel (itt minden feladat egy lekérdezés).
- A megoldásaidat gyűjtsd össze egy .txt fájlba, amelyet a ZH végén canvasba kell feltölteni.
- Mindig másold be a megoldásodat és a futtatásod eredményét. Ha egy megoldáshoz nincs futási eredmény csatolva, az 0 pontot ér.

Ponthatárok:

- Elérhető pontok száma: 8
- 1-3 (elégtelen), 4 (elégséges), 5 (közepes), 6 (jó), 7-8 (kiváló)

Spark RDD feladatok

Feladat 1. (2 pont)

A kmerInput.txt az E. coli baktérium genomjának egy részét tartalmazza (A, T, G és C karakterek sorozata). A feladat egy k-mer számoló program elkészítése. A bioinformatikában k-mer-nek nevezzük a k karakter hosszú részsztringeket. Pl: A “AGCTTTTC” 3-mer-ei a következők: AGC, GCT, CTT, TTT, TTT, TTC.

Készítsen egy programot, amely összeszámolja a kmerInput.txt 3 hosszú k-mereit (3-mer). Csak azok a 3-mereket vegye számításba, amelyek tartalmazzák a T betűt és az előfordulásuk száma legalább 100. A program írja ki a szűrésen átment 3-merek darabszámát.

Elvárt kimenet: egyetlen szám.

Feladat 2. (2 pont)

A weboldalak.txt weboldalakat és azok szöveges tartalmát tartalmazza. Minden sor egy weboldal címével kezdődik, aztán pedig a tartalma következik. A felhasználónk az "ELTE" kifejezésre keres rá, nekünk pedig vissza kell adnunk számára a legrelevánsabb találatot. Határozod meg, hogy a weboldalak.txt-ben található weboldalak közül melyikben szerepel legtöbbször a keresett kifejezés. Csak azokat a weboldalakat vedd figyelembe, amelyek tartalma több, mint 10 szóból áll.

A programod írja ki azt a weboldalt, amelyekikben a legtöbbször szerepel a keresett kifejezés. Add meg továbbá a keresett kifejezés előfordulásainak a számát és a weboldal tartalmában lévő szavak számát is.

Egy lehetséges kimenet: ('wikipedia.hu', 10, 99), amelynek jelentése, hogy a wikipedia.hu weboldal tartalmában a keresett kifejezés 10-szer szerepel és összesen 99 szóból áll.

Spark DataFrame feladatok

Az alábbi 4 feladat megoldásához a books.csv és orders.csv fájlokat kell beolvasni. A fájlok egy könyveket árusító webshop adatait tartalmazza (forrás: kaggle.com). A books.csv minden sora egy könyv adatait tartalmazza, az orders.csv minden sora pedig egy adott vásárlási folyamat egy adott könyvhöz tartozó interakcióit adja meg (kattintások száma, kosárba helyezett darabszám, megvásárolt darabszám).

books.csv oszlopai:

- rowID – a sor egyedi azonosítója
- itemID – egy könyv egyedi azonosítója
- title – a könyv címe
- author – a könyv szerzője
- pupblisher – a könyv kiadója

orders.csv oszlopai

- rowID – a sor egyedi azonosítója
- sessionID – egy vásárlási folyamat azonosítója
- itemID – egy könyv egyedi azonosítója
- click – a könyv oldalára történő kattintások száma a vásárlási folyamat során
- basket – az adott könyv kosárba helyezett darabszáma a vásárlási folyamat során
- order – az adott könyvből megrendelt darabszám a vásárlási folyamat során

Feladat 3. (1 pont)

Add meg azt a három szerzőt (author), akik a legtöbb könyvet írták (a null nem számít szerzőnek).
Elvárt oszlopok: [szerző neve, könyvek száma]

Feladat 4. (1 pont)

Add meg azokat a szerzőket, akiknek több mint 35 különböző kiadónál is jelent meg könyvük.
Elvárt oszlopok: [szerző neve, kiadók száma]

Feladat 5. (1 pont)

Add meg, hogy a felhasználók munkameneteiben (sessionID) átlagosan hány kattintás történik.
Elvárt oszlopok: [átlagos kattintások száma]

Feladat 6. (1 pont)

Melyik könyvből vásárolták meg a legtöbbet? Add meg a szerző nevét, a könyv címét, és a megvásárolt könyvek darabszámát.
Elvárt oszlopok: [szerző neve, könyv neve, eladott darabszám]