BigData architektúrák és elemző módszerek GY.

Hadoop beadandó feladat (2022-23 őszi félév)

Határidő: 2022.10.30

Beküldés: canvas.elte.hu

Értékelés: A feladathoz négy, egymásra épülő részfeladat tartozik. A beadandóra adott érdemjegy a megoldott részfeladatok alapján kerül meghatározásra. Ha a program nem fut, az értékelés elégtelen.

Feladat:

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.

- (**Elégséges**) Készítsen egy programot, amely összeszámolja és kiírja a kmerInput.txt 3 hosszú k-mereit (3-mer).
    - Példa bemenet: AGCTTTTC
    - Példa kimenet:
      AGC 1
      GCT 1
      CTT 1
      TTT 2
      TTC 1
- (**Közepes**) Csak azok a 3-merek szerepeljenek a kimenetben, amelyek tartalmazzák a T betűt (map szűrés) és az előfordulásuk száma nagyobb, mint 100 (reduce szűrés).
- (**Jó**) Használjunk lokális redukálást (combinert) a feladat elvégzéséhez. megj.: nem egyenlő reducer kódjával!
- (**Kiváló**) Készítsünk saját rekordtípust (Writable), amely tárolja, hogy a sorban hányadik pozíciótól kezdődött a k-mer. Eredményül adjuk meg az előfordulások számát és hogy mi a legkisebb index, ahol a k-mer szerepel.
    - példa bemenet AGCTTTTC
    - példa kimenet:
      k-mer (előfordulás, pozíció)
      GCT (10,1)
      CTT (12,2)
      TTT (20,3)
      TTC (10,5)

(**Megjegyzés:** a k-mer-ek elkészítésekor elég csak az adott sort vizsgálni, azaz nem kell egy sor utolsó karakterét összefűzni a rákövetkező sor első karaktereivel.)