BigData architektúrák és elemző módszerek GY.

Hadoop beadandó feladat (2023-24 őszi félév)

Határidő: 2023.11.05 (vasárnap éjfél)

Beküldés: canvas.elte.hu, „Hadoop beadandó” feladat

Értékelés: A feladathoz négy, egymásra épülő részfeladat tartozik. A beadandóra adott érdemjegy a megoldott részfeladatok alapján kerül meghatározásra, a program személyes bemutatása során. Bemutatásra a gyakorlat idejében van lehetőség a határidőnek megfelelően. Ha a program nem fut, az értékelés elégtelen.

Feladat:

A kmer_input1.txt, kmer_input2.txt és kmer_input3.txt fájlok az E. coli baktérium genomjának egy részét tartalmazzák (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 bemeneti fájlok 3 hosszú k-mereit (3-mer). A fájlokat ne öntsük egybe, mardjon meg a 3 különböző input txt fájl.
    - 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 és az előfordulásuk száma nagyobb, mint 100.
- (**Jó**) A k-merek előfordulásai után írjuk ki a kimeneti fájlba a szűrésen átment k-merek előfordulásainak összegét is.
- (**Kiváló**) A lokálisan megoldható összegzést végezzük el combiner segítségével.

(**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.)