„Big Data" elemzési módszerek zárthelyi 2015

1 Adatelemzési és statisztikai alapok

1. Mi a strukturált/nemstrukturált/szemistrukturált adat? Mondjon példát a típusokra! **(5p)**

2 Vizuális analízis (15p)

2. Mi a dobozdiagram (*boxplot*)? Minek a szemléltetésére használjuk? Ábrán szemléltesse, hogy a dobozdiagram hogyan reprezentálja egy megfigyelés-halmaz alapvető leíró statisztikáit! **(4p)**
3. Mi a dobozdiagram mediánjának, „bajszainak" és „sarokpontjainak" (*whiskers and hinges*) kapcsolata a normális eloszlás paramétereivel? Diszkutálja, hogy alkalmas-e a dobozdiagram más eloszlások szemléltetésére is, és ha igen, milyen korlátokkal! **(5p)**
4. Kiszámítjuk egy folytonos változó értékeit tartalmazó adatsor mediánját, móduszát és átlagát. Válassza ki az igaz állításokat! **(6p)**
   - A medián biztosan kisebb az átlagnál.
   - Az átlag legfeljebb kétszerese lehet a mediánnak.
   - A medián és a módusz az adatsor egy-egy kitüntetett értékét jelölik.
   - A módusz megegyezhet a mediánnal.
   - Találunk olyan reguláris kategorikus változót, amelynek mediánja megegyezik az általunk kiszámolt mediánnal.
   - Találunk olyan reguláris kategorikus változót, amelynek módusza megegyezik az általunk kiszámolt mediánnal.

3 Nagyméretű adatok vizualizációja (9p)

5. Mik a disztributív, algebrai, holisztikus típusú statisztikai aggregátorok? Hová tartozik a szórás, az IQR és a percentilis? **(5p)**
6. Mi a *small multiplies* elv a vizualizációban? Miért különösen fontos ez a nagyméretű adatsorok vizualizációja témakörben? **(4p)**

4 A MapReduce algoritmusszervezési minta (10p)

7. Mi a "shuffle and sort" fázis feladata a MapReduce végrehajtás során? **(2p)**
8. A kiterjesztett MapReduce sémában mi a "combiner" feladata? Miért érdemes alkalmazni? **(3p)**
9. Tároljunk a HDFS-ben fix formátumú CSV állományokat, melyek n folytonos változó feletti megfigyeléseket írnak le egy időbélyeggel kiegészítve. Adjon Mapper és Reducer pszeudokódot az egyes megfigyelt változók időbeli maximum-helyének meghatározására! **(5p)**

5 Adatfolyam-feldolgozás

10. Web crawlerünkben Bloom-filtereket alkalmazunk a már látogatott URL-ek felismerésére. Bloom filterünk két hash függvényt használ, a következő paraméterekkel működik:
    - N = 11
    - Input: egész számok (az URL-eket reprezentálandó)
    - h1(x): a páros bitekből képezett y mod N (tehát h1(585) = h1(1001001001_2) = 01001_2 mod 11 = 9 mod 11 = 9)
    - h2(x): a páratlan bitekből képzett y mod N (tehát h2(585) = h2(1001001001_2) = 10010_2 = 18 mod 11 = 7)

    A rendszerünkben eddig a következő műveleteket hajtottuk végre: **Beszúr(25), Beszúr(159)**. Mit ad vissza a **KERES(118)** művelet? Interpretálja a végeredményt! **(8p)**