Ellenőrző kérdések a BigData elemzési módszerek zárthelyihez - 2017

1 Big Data: alapfogalmak

- Adja meg és definiálja a Big Data problémák „4V" modelljének négy jellemzőjét!
- Mit jelent az „at rest" Big Data? Mit tekintünk ellentétének a Big Data problémák kontextusában?
- Mit jelent a lazán csatolt, magas lokalitású párhuzamosítás az elosztott számítástechnikában? Alkalmazása során mi a „bring the computation to the data" elv szerepe?
- Mi a felhő szolgáltatások szerepe a Big Data adatfeldolgozásban? Jellemezze, hogy mely esetekben „éri meg" nagy adat problémákat felhő platformokon megoldani!

2 Adatelemzési és statisztikai alapok

- Milyen típusú változótípusokat különböztetünk meg? Hol van ezeknek szerepe? Milyen típusú változók fordulhatnak elő egy olyan adatsorban, amely egy magyarországi lakosok vásárlási szokásait felmérő, alábbi pontokat tartalmazó kérdőívből született:
    - nyilatkozó neme, életkora, lakóhelye, legmagasabb iskolai végzettsége;
    - vásárlási gyakorisága, hetente hányszor vásárol X terméket;
    - a standard vagy a prémium alterméket szereti?
- Mi a strukturált/nemstrukturált/szemistrukturált adat? Mondjon példát mindhárom típusra!
- Mi a felderítő és mi a megerősítő statisztikai elemzés? Mondjon példát mindkét megközelítésre!
- Centrális tendencia jellemzésére mi a robosztusabb operátor: az átlag vagy a medián? Miért?

3 Vizuális analízis

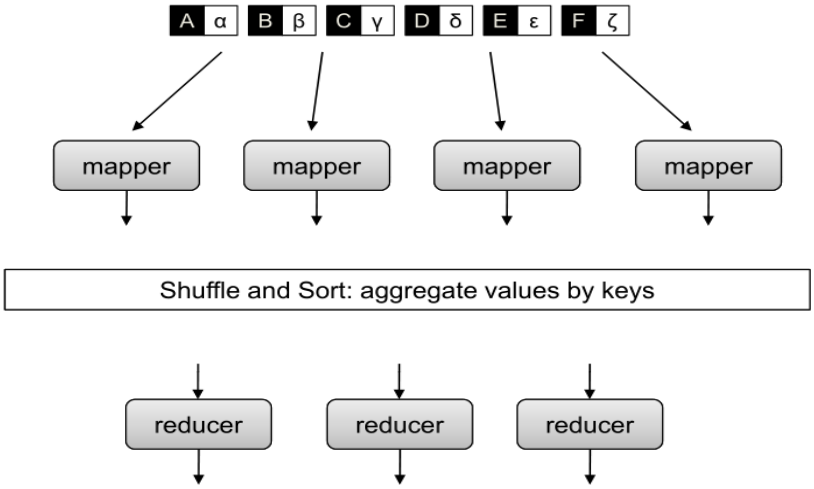
- Mik a fő különbségek az EDA és a CDA között az adatelemzés során?
- Mi a dobozdiagram? Minek a szemléltetésére használjuk? Ábrán szemléltesse, hogy a dobozdiagram hogyan reprezentálja egy megfigyelés-halmaz alapvető leíró statisztikáit!
- Mi a dobozdiagram mediánjának, „bajszainak" és „sarokpontjainak" (*whiskers and hinges*) kapcsolata a normális eloszlás paramétereivel? Diszkutálja, hogy alkalmas-e a dobozdiagram más eloszlások szemléltetésére is, és ha igen, milyen korlátokkal!
- Mi a SPLOM? Miért használjuk a vizuális EDA során? Mik alkalmazásának legfőbb korlátai?
- Mozaik-diagram: szöveges definíció, szemléltetés ábrával, jellemző alkalmazási esetei és alkalmazásának korlátai.
- Korrelogram: szöveges definíció, szemléltetés ábrával.
- Treemap: szöveges definíció, szemléltetés ábrával.
- Párhuzamos koordináták: szöveges definíció, szemléltetés ábrával, jellemző alkalmazási esetei és alkalmazásának korlátai.
- Mit jelent, ha egy párhozamos koordináta diagram két szomszédos tengelye között futó szakaszokra illesztett egyenesek egy pontban metszik egymást?

- Mit jelent, ha egy párhozamos koordináta diagram két szomszédos tengelye között futó szakaszokra illesztett egyenesek két pontban metszik egymást? Milyen hipotézist állítana fel ebből a megfigyelésből?
- Mik a trellis plotok? Jellemzően miért/mikor használjuk őket?

4 Nagy méretű adatok vizualizációja

- Mik a disztributív, algebrai, holisztikus típusú statisztikai aggregátorok? Hová tartozik a szórás, az IQR és a percentilis?
- Milyen típusú vizualizációkat alkalmazunk szokásosan a Big Data vizualizáció során?
- Ismertesse és indokolja, hogy a vizuális analízis klasszikus diagram-típusai közül melyek alkalmazása válik nehézkessé, illetve értelmetlenné „Big Data" kontextusban!
- A Big Data vizualizáció során a megjelenítést és a diagram-leíró adatok kiszámítását jellemzően szétcsatoljuk és az előbbit is több lépésben végezzük. Vázolja fel és ismertesse a Big Data vizualizációs „csővezeték" (*pipeline*) főbb lépéseit!

5 A MapReduce algoritmusszervezési minta



- Hogyan érjük el a MapReduce séma alkalmazásánál az adat és kód kolokációját?
- Mi a "shuffle and sort" fázis feladata a MapReduce végrehajtás során?
- A kiterjesztett MapReduce sémában mi a "combiner" feladata? Miért érdemes alkalmazni?
- Tároljunk a HDFS-ben fix formátumú CSV állományokat a következő szerkezettel: *timestamp,sensor_1_val,sensor_2_val,sensor_3_val*. Adjon Mapper és Reducer pszeudokódot a három szenzorral is megfigyelt jelenség (pl. hőmérséklet) óránkénti átlagos értékének meghatározására!

6 Adatfolyam-feldolgozás

- Ismertesse az adatfolyam-feldolgozás elemi blokkjának tekintett "stream processor" mintát! Hogyan történik ezekkel a bejövő adatfolyamok feldolgozása?
- Milyen problémák merülhetnek fel adatfolyamok mintavételezésénél? Kulcs és érték mezőkre osztható feldolgozandó n-esek esetén hogyan valósítaná meg a kulcstér feletti mintavételezést?

(Azaz a kulcsok halmazán mintavételezünk – pl. felhasználók, ha (*user,search_query*) alakú megfigyeléseink vannak - és minden a mintába eső kulcshoz tartozó n-est továbbengedünk.)

- Mik a Bloom filterek? Hogyan alkalmazzuk őket halmazba tartozás közelítő ellenőrzésére adatfolyam-feldolgozásban?
- Milyen rendszerintegrációs mintát valósít meg az Apache Kafka? Milyen megfontolások vezethetnek minket arra döntésre, hogy adatfolyam-feldolgozási topológiákat Kafka segítségével integráljunk, egyfajta „szuper-topológiát" képezve?

7 Spark

- Ismertesse a Spark számításszervezési-modelljének alapelemét, a *Resilient Distributed Dataset*-et (RDD)!
- Mit jelent, az hogy a Spark támogatja az ezekből szervezett számításleíró gráf lusta kiértékelését (*lazy evaluation*)?
- A Spark alapvetően két fajta *operációt* támogat, a *transzformációkat* és az *akciókat*. Definiálja az előbbi kategóriát és adjon három példát!
- A Spark alapvetően két fajta *operációt* támogat, a *transzformációkat* és az *akciókat*. Definiálja a második kategóriát és adjon három példát!
- Mit jelent az, hogy a Spark nem „valódi" folyamfeldolgozást valósít meg, hanem mikrokötegelést (microbatching)? Mik a mikrokötegelés előnyei és hátrányai?

8 ML/DM alapok

- Röviden ismertesse az adatlemzés alapfeladat-kategóriáit!
- k-means klaszterezés: alapötlet, definíció, (szekvenciális) pszeudokód
- Mi a sűrűségalapú klaszterezés alapötlete? Miben különbözik alapvetően a távolaságalapú klaszterezéstől?
- Bináris osztályozási döntések jóságának mérése: a konfúziós mátrix (igazságmátrix, eset-kontroll tábla; *confusion matrix*)
- Feladatunk prediktív karbantartás támogatása; ennek érdekében egy gyártósor esetén osztályozót építünk arra, hogy a mért és historikus paraméterek alapján a gyártósor el fog-e romlani a követlező ütemezett karbantartás előtt. Osztályozónk érzékenysége vagy specifikussága a fontosabb tulajdonság? Válaszát érveléssel támassza alá!
- Asszovciációs szabályok fogalma, support, confidence és lift definíciója.
- Lineáris regresszió: alapötlet, definíció
- Linráris regresszió: R^2 mérték (determinációs koefficiens), konfidencia-intervallum és predikciós intervallum fogalma.
- Mit csinálunk a Principal Component Analysis (PCA) művelet során? Jellemzően mire használjuk?
- Mit jelenítenek meg a biplotok? Segíthet-e a biplot-reprezentáció a klaszterek felismerésében? Miért?