

# Idősorok iteratív, semi-supervised protokoll szerinti klasszifikációja

- Idősorok osztályozása különböző felismerési feladatok közös elméleti háttere (kézírás, beszéd, EKG, agyhullámok)
- Adott
  - tanító adathalmaz (ismert osztálycímkék), és
  - felismerendő (teszt) adathalmaz (nem ismertek az osztálycímkék)
- Ötlet: a felismerendő halmazbeli osztálycímkék nem ismertek, de az objektumok igen, amennyire lehetséges, próbáljunk „tanulni” ebből az információból.

# Idősorok iteratív, semi-supervised protokoll szerinti klasszifikációja (folyt.)

- Konkrétan:
  - 1. lépés: készítünk egy legközelebbi szomszéd osztályozót (3 eset:  $k=1$ ,  $k=5$ , és  $k=10$ )
  - 2. lépés: Osztályozzuk a teszt idősorokat az előbbi osztályozóval, és keressük meg azon idősort, melynek osztályozása során „legbiztosabb” az osztályozó, azaz: azon  $t$  idősort keressük, melyre legkisebb a legközelebbi szomszédoktól mért távolságok összege. Osztályozzuk  $t$  idősort.
  - 3. lépés: vegyük hozzá a  $t$  idősort a tanító halmazhoz a 2. lépésben meghatározott címkével, és ismételjük az 1. lépéstől.

# Idősorok iteratív, semi-supervised protokoll szerinti klasszifikációja (folyt.)

- Biztosítjuk a feladathoz:
  - Hasonlósági mátrix 45 idősor-adatbázisra (összesen kb. 8 GB!)
  - A tanítóhalmazbeli idősorok címkéi
- A feladat megoldásaként elvárjuk:
  - A felismerendő (teszt) halmazra kiszámolt címkék beadását mind semi-supervised, mind „konvencionális” módon történő számítás esetén
  - Szoftver (forráskód), dokumentáció, előadás fóliák beadását
  - Bemutató előadás tartását

# Szövegadatok bányászata

- Két részfeladat:
  - a) Szövegadatok hatékony indexálása a halmaz jól kereshetővé tételéhez.
  - b) Gyakori mintázatok keresése szövegadatokban
- A két részfeladat külön-külön és együttesen is megoldható

# Szöveghadatok bányászata (folyt.)

- Biztosítjuk a feladathoz:
  - Szöveges adatok nagy halmaza (több GB)
- A feladat megoldásaként elvárjuk:
  - A szövegek (szükség szerinti) előfeldolgozását
  - Indexálás esetén: gyorsítás (speed up) demonstrálása a naiv keresőhöz képest
  - Gyakori mintabányászat esetén: megtalált gyakori minták bemutatása
  - Szoftver (forráskód), dokumentáció, előadás fóliák beadását
  - Bemutató előadás tartását