**soproni egyetem**

**simonyi károly műszaki, faanyagtudományi és művészeti kar**

**informatikai és gazdasági intézet**



**új nemzeti kiválóság program**

**ÚNKP-18-1-I-SOE-27**

**dokumentáció**

**csanaki richárd**

**dr. pödör zoltán, dr. jereb lászló**

**kzeadr**

**sopron,**

**2019. március 3.**

**Ösztöndíjas időszak**

2018. november 1. – 2019. március 31.

**Vállalások**

1. Kutatási tervben megfogalmazott célok teljesítése
2. Havonta legalább egy magyar vagy idegen nyelvű szakmai anyag feldolgozása
3. Egyetemi ÚNKP rendezvényen való részvétel

**Feladat megfogalmazás**

A Soproni Egyetem SKK Informatikai és Gazdasági Intézménye által működtetett SensorHUB [1] infrastruktúrában tárolt szenzoradatok automatizált elemzése adatbányászati, Big Data megoldásokkal [2]. Valamint informatív adatvizualizáció az adathalmazokból kinyert ténylegesen új információk szakmai szempontú megjelenítésére [3]. Ezeken az információkon alapuló döntéshozatal támogatása, majd a döntéshozatali ciklusból a humánfaktor részvételének minimalizálása, gépi tanulás (machine learning), öntanuló algoritmusok és az emberi agy (intelligencia forrása, neuronhálózatok rendszere) modellezési lehetőségeinek vizsgálata, a lehetséges megoldások illesztése az adott feladathoz.

**Feldolgozott irodalom, anyagok, kurzusok**

1. Life 3.0 – Max Tegmark
2. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython – Wes McKinney
3. Python in 24 Hours – Katie Cunningham
4. Tidy Data – Hadley Wickham
5. Hands-On Machine Learning with Scikit-Learn & TensorFlow
6. Think Stats (Exploratory Data Analysis in Python) – Allen B. Downey
7. Pandas for Data Analysis, SciPy 2017 Tutorial – Daniel Chen
8. Introduction to Probability (Harvard Statistics 110) – Joseph K. Blitzstein

**Felhasznált technológiák**

* Python
  + Interpretált programozási nyelv, Big Data és ML területen standard
* Anaconda
  + Könyvtár és virtuális környezet csomag Python nyelven való fejlesztéshez
* CDS API
  + Copernicus adatbázisok lekérdező felülete
* SensorHUB
  + Intézeti adattároló környezet
* ecCodes
  + Terminál eszköztár GRIB és NetCDF formátumú fájlok kezelésére
* CMake
  + C alapú build eszköztár
* QGIS
  + GRIB, netCDF és további tudományos fájlok megjelenítésére képes open source alkalmazás
* Pandas
  + Tabuláris adatelemző könyvtár
* Matplotlib
  + Tudományos adatvizualizációs könyvtár
* NumPy
  + Nagy teljesítményű mátrix könyvtár
* Scikit-Learn
  + ML eszköztár
* Jupyter Notebook
  + Interaktív Python interpreter (Ipython Shell)

**Felhasznált technológiák – bővebben**

***Python***

A Python egy általános célú, high-level interpretált programozási nyelv. Főbb ismérvei a könnyű olvashatóság és az élénk fejlesztői közösség. Webfejlesztés, adatelemzés és gépi tanulás területeken standardnak számít.

***Anaconda***

Data Science területen standard Python disztribúció, könyvtár gyűjtemény. Több, mint 1000 open source csomagot tartalmaz, többek között a Pandas, NumPy, Matplotlib, Scikit-Learn és TensorFlow könyvtárakat.

***Pandas***

Python könyvtár könnyen használható, nagy teljesítményű adatstruktúrákkal, adatelemző eszközökkel.

***NumPy***

Python könyvtár N-dimenziós tömbökkel és ezek transzformációs eszközeivel. A hagyományos adatszerkezeteknél könnyebb és gyorsabb feldolgozhatóságot biztosít tetszőleg adattípusokkal.

***Scikit-Learn***

Klasszikus machine learning Python könyvtár, regressziós, klaszterezős és klasszifikációs modulokkal. Támogatja az ún. Supervised és Unsupervised gépi tanulás módszereket is.

***Jupyter Notebook***

IPython Shell alapú böngészős Python fejlesztői környezet, mely több integrált könyvtár segítségével és megoldással könnyíti meg a projektek dokumentálását: Markdown markup nyelv támogatása, Matplotlib grafikonok kirajzolása, adatszerkezetek automatikus megjelenítése.

***SensorHUB [1]***

Az Informatikai és Gazdasági Intézet által működtetett adattároló technológia, az országszerte kihelyezett 6 mérőállomás által küldött adatok tárolási helye, az elkészült általános Gateway alkalmazás [2] átmeneti futtató környezete.

***CDS API***

Copernicus Climate Data Store lekérdező felülete, Python környezetből elérhető, az egész Földre kiterjedő meteorológiai adatokat tárolja, könnyen hozzáférhető programozói interfészt biztosítva.

***ecCodes***

Az ECMWF által fejlesztett parancssori eszköztár, melynek segítségével a GRIB és BUFR kiterjesztésű fájlok feldolgozhatóak, átalakíthatóak más programok által is feldolgozható formátumba. A klasszikus GRIB-API bővített verziója.

***CMake***

C alapú build alkalmazás, hordozható tudományos alkalmazások telepíthetőek vele.

***QGIS***

Grafikus felhasználói felülettel rendelkező alkalmazás mely automatikusan képes GRIB és netCDF kiterjesztésű fájlokat megjeleníteni, elemezni.

A projektben részt vevő személyek

**Adatok**

Az elemzések első és legfontosabb lépése a sok és jó minőségű adat összegyűjtése.

*Milyen mennyiségtől tekintünk egy adathalmazt Big Data-nak?*

Tudományos körökön kívül, egyéni felhasználóként akár az 1-től 3 gigabájt (109 bájt) méretig terjedő adathalmazok már hatalmasnak számítanak, főleg a személyi számítógépek drága és limitált memóriakapacitása miatt. Az elemzéshez használt adatokat azért kell a memóriában tárolni, hogy az egyes Data Science könyvtárak által alkalmazott algoritmusok lefutása viszonylag gyorsan lefuthasson, pár GB-nyi adattal (1-5 évig terjedő adathalmaz) PC-ken is pillanatok alatt képesek ezek a programok eredményt generálni, főleg azért, mert ezen méreteknél nagyságrendekkel nagyobb mennyiségű adat feldolgozására lettek felkészítve.

Viszonytásképp a szuperszámítógépekkel felszerelt intézetek és vállalatok napi szinten petabájt (1015 bájt) nagyságrendű adatbázisokkal dolgoznak. Ebből következik, hogy maga a „Big Data” elnevezés erősen szituáció függő.

*Mitől lesz jó minőségű az adat?*

Az adatok összegyűjtése, azok szűrése és manipulálása külön tudományággá vált. Erről szól Hadley Wickham tanulménya is, mely kiinduló standarddá vált adat elmező körökben. Az ebben az anyagban megfogalmazott alapelvek:

1. Az egyes jelenségeket leíró adathalmazok feloszthatók tulajdonságokra és megvalósulásokra, azaz attribútumokra és ezek alapján rekordokra.
2. Az attribútumok a jelenség elemi tulajdonságait jelentik, így elemi adattípusokkal leírhatók.
3. A jelenséget leíró adathalmazok nem feltétlenül tabulárisak (táblázatos felépítésűek), bár a különböző adatbázis elmező szoftverek a tabuláris adatbázisok elemzésire már jól optimalizáltak.
4. Ne féljünk belenyúlni a megkapott adathalmazba! Ez az úgynevezett „data wrangling” folyamat, melyet jó közelítéssel adatmanipulációnak nevezhetünk. Ezen lépés teszi ki az adatelemzői procedúra oroszlánrészét, nem véletlenül: a hibás adatok (rossz mérések, duplikátumok, stb.) használata rossz és elvi hibás következtetések levonását eredményezik.
5. Ügyeljünk arra, hogy minimalizáljuk az adatmanipuláció során elvesztett adat mennyiségét, ugyanis a hiányzó adat sokszor hiányzó információt jelent.

***A projektben használt adatforrások***

* Az intézeti SensorHUB környezet 2015 év szeptemberéig visszamenőleg tárol adatokat, melyek forrása az Erdészeti Tudományos Intézet (ERTI) által működtetett 6 mérőállomás (Budapest, Mátrafüred, Püspökladány, Sárvár, Sopron, Kaszó). Ezen adathalmaz kb. 52 millió sornyi rekordot tartalmaz.