

Tantárgy neve: Big Data architektúrák és elemző módszerek	Kreditértéke: 5 kredit
A tantárgy besorolása: kötelező	
A tantárgy elméleti vagy gyakorlati jellegének mértéke : 50% (kredit%)	
A tanóra típusa : ea. / gyak. / konz és óraszám a: 2 / 2 / 1 az adott félévben , Az adott ismeret átadásában alkalmazandó további (sajátos) módok, jellemzők :: esetismertetések, prezentációk	
A számonkérés módja (koll. / gyj. / egyéb ¹): koll. / gyj. (összevont számonkérés) Az ismeretellenőrzésben alkalmazandó további (sajátos) módok : <i>órai és házi projekt feladatok</i>	
A tantárgy tantervi helye (hányadik félév): 5. félév	
Előtanulmányi feltételek (<i>ha vannak</i>): Programozási nyelvek I.	

Tantárgy-leírás: az elsajátítandó ismeretanyag tömör, ugyanakkor informáló leírása

Az informatika hatalmas fejlődésen ment keresztül az elmúlt tíz évben. Ma már életünk és világunk szinte minden entitásáról és változásáról valamilyen szintű digitális lenyomattal is rendelkezünk (pl. SmartCity, banki rendszerek, viselhető szenzorok, IoT, mobil hálózatok, közösségi hálózatok, stb.). Az informatikai rendszerek által generált egyre nagyobb adat tömegek a tárolási problémák mellett is számos kihívást jelentenek. Ezen adatok sokszor önmagukban semmitmondóak, értéket csak megfelelő tisztítás, elemzés, értelmezés és együttes használat során nyernek. A kis adattömegek esetén jól működő algoritmusok és módszerek sok esetben nagy, akár több terrabájtos adat mellett már nem nyújtanak megfelelő teljesítményt vagy egyáltalán nem alkalmazhatóak. Ezen folyamat eredményeképp az informatikai szektor is megváltozott. A piacon egyre nagyobb az igény olyan ún. adat tudós (data scientist) szakemberek iránt, akik az adatok megfelelő feldolgozását és értelmezését el tudják végezni, megfelelő gyakorlati és elméleti ismeretekkel rendelkeznek a gráf elemző és gépi tanulási módszerekről. A tárgy gyakorlati példákon keresztül vezeti be a hallgatókat a különösen nagy adattömegek elemzésének és adatbányászatának világába.

Részletes tematika:

1. Bevezetés, Mi is az a Big Data?
2. Adatelemző eszközök: Jupyter, iPython környezet
3. Adatelemző eszközök: Python és R használata
4. Adatelemző eszközök: Hadoop, Spark, Mahout, H2O, SciKitlearn
5. Alapfogalmak: Korreláció, mintavételezés, hipotézis vizsgálat, véletlen adat, KDE
6. Adatfeltáró elemzések Pandas, Numpy, Scipy
7. Adattisztítás fontossága, kiugró értékek kiszűrése, hiányzó értékek kezelése
8. Adattárolás: HadoopHDFS, replikáció, hibatűrés
9. Klaszter management: HadoopYarn
10. Batch feldolgozás: HadoopMapReduce
11. Memória alapú batch feldolgozás: Spark
12. Stream feldolgozás: Storm, Flink, Spark Stream
13. Gráf feldolgozás: Spark GraphX
14. Klaszterezés KMeans, Gaussian mixture, PIC, LDA, Streaming KMeans
15. Dimenzió csökkentés (SVD, PCA, RobustPCA, ICA)
16. Osztályozás, regresszió: Lineáris modellek (LASSO, Ridge, SVM, logisztikus regresszió, KNN)
17. Osztályozás, regresszió: Döntési fák, Random forest, Gradiens módszerek, AdaBoost, LogitBoost
18. Deep learning módszerek
19. Ajánlórendszerek
20. Adatvizualizáció és prezentáció matplotlib és d3.js, Zeppelin, Lightning

A legfontosabb kötelező, illetve ajánlott irodalom (jegyzet, tankönyv) felsorolása bibliográfiai adatokkal (szerző, cím, kiadás adatai, esetleg oldalak, ISBN)

- Raúl Garreta: Learning scikitlearn: Machine Learning in Python, Packt Publishing, ISBN:9781783281930
- Nick Pentreath: Machine Learning with Spark Tackle Big Data with Powerful Spark Machine Learning Algorithms, Packt Publishing, ISBN: 9781783288519
- Viktor Mayer Schonberger and Kenneth Cukier: Big Data: A Revolution That Will Transform, How We Live, Work, and Think, Eamon Dolan/Houghton Mifflin Harcourt, ISBN: 9780544002692
- Eric Siegel: Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, Wiley, ISBN: 9781118356852
- Tom White: Hadoop: The Definitive Guide
-

Azoknak az előírt szakmai kompetenciáknak, kompetencia-elemeknek (tudás, képesség stb., KKK 8. pont) a felsorolása, amelyek kialakításához a tantárgy jellemzően, érdemben hozzájárul

a) tudása

- Ismeri és érti az informatikai szakterület legfontosabb általános elméleteit, összefüggéseit, tényanyagát és az ezekhez szükséges felépítő fogalomrendszert, különösen az alábbi területeken: gépi tanulás és adatbányászat
- Ismeri az informatikai szakterület tervezési, fejlesztési, működtetési és irányítási folyamatainak alapvető feladatmegoldási elveit, módszereit és eljárásait, különösen - választott specializációjának megfelelően - a következő területeken: predikciós modellek és ajánlórendszerek fejlesztése, adat termék fejlesztése
- Ismeri és érti az informatikai szakterület legfontosabb általános elméleteit, összefüggéseit, tényanyagát és az ezekhez szükséges felépítő fogalomrendszert, különösen az alábbi területeken: elosztott számítások, map-reduce paradigma, nagy adattömegek feldolgozása
- Ismeri az informatikai szakterület tervezési, fejlesztési, működtetési és irányítási folyamatainak alapvető feladatmegoldási elveit, módszereit és eljárásait, különösen - választott specializációjának megfelelően - a következő területeken: Hadoop és Spark

b) képességei

- Képes az informatikai szakterület tudásanyagát alkalmazni algoritmusok tervezésére, elemzésére és implementálására a legfontosabb programozási paradigmák figyelembe vételével az alábbi területeken: adat alapú modellezés, gépi tanulás, adatbányászat.
- Képes az adattudomány módszereinek alkalmazására.
- Képes az informatikai szakterület tudásanyagát alkalmazni osztott rendszerek használata során, különös tekintettel osztott számítások elvégzésére.

Tantárgy felelőse (név, beosztás, tud. fokozat): Laki Sándor, Egyetemi adjunktus, PhD

Tantárgy oktatásába bevont oktató(k), ha van(nak) (név, beosztás, tud. fokozat):

Gombos Gergő, Egyetemi tanársegéd