

A TANTÁRGY ADATLAPJA

1. A képzési program adatai

1.1 Felsőoktatási intézmény	Babeş–Bolyai Tudományegyetem
1.2 Kar	Matematika és Informatika Kar
1.3 Intézet	Magyar Matematika és Informatika Intézet
1.4 Szakterület	Informatika
1.5 Képzési szint	Mesteri
1.6 Szak / Képesítés	Adatelemzés és modellezés / Analiza datelor și modelare Data analysis and modelling

2. A tantárgy adatai

2.1 A tantárgy neve	Nagy adathalmazok elemzése Analiza datelor masive Analysis of massive datasets						
2.2 Az előadásért felelős tanár neve	Dr. Darvay Zsolt egyetemi docens						
2.3 A szemináriumért felelős tanár neve	Dr. Darvay Zsolt egyetemi docens						
2.4 Tanulmányi év	1	2.5 Félév	1	2.6. Értékelés módja	vizsga	2.7 Tantárgy típusa	kötelező – alap
2.8 A tantárgy kódja	MMM8075						

3. Teljes becsült idő (az oktatási tevékenység féléves óraszám)

3.1 Heti óraszám	5	melyből: 3.2 előadás	2	3.3 szeminárium/labor/praktika	3
3.4 Tantervben szereplő össz-óraszám	70	melyből: 3.5 előadás	28	3.6 szeminárium/labor	42
A tanulmányi idő elosztása:					Óra
A tankönyv, a jegyzet, a szakirodalom vagy saját jegyzetek tanulmányozása					49
Könyvtárban, elektronikus adatbázisokban vagy terepen való további tájékozódás					14
Szemináriumok / laborok, házi feladatok, portfóliók, referátumok, esszék kidolgozása					49
Egyéni készségfejlesztés (tutorálás)					14
Vizsgák					4
Más tevékenységek:					
3.7 Egyéni munka össz-óraszám					130
3.8 A félév össz-óraszám					200
3.9 Kreditszám					8

4. Előfeltételek (ha vannak)

4.1 Tantervi	Nincs
4.2 Kompetenciabeli	Alapvető algoritmusok, programozási készségek, matematikai alapismeretek (algebra, valószínűségszámítás).

5. Feltételek (ha vannak)

5.1 Az előadás lebonyolításának feltételei	<ul style="list-style-type: none"> • Az előadásokhoz videoprojektor szükséges. • A példák kifejtéséhez és illusztráció számára tábla szükséges.
5.2 A szeminárium / labor	<ul style="list-style-type: none"> • A laboratóriumi órák alatt a diákok a számítógépet, az oktató a táblát

lebonyolításának feltételei	használja.
-----------------------------	------------

6. Elsajátítandó jellemző kompetenciák

Szakmai kompetenciák	<ul style="list-style-type: none"> Nagy adathalmazok elemzése információkinyerés céljából Adatbányászati algoritmusok elemzése és fejlesztése
Transzverzális kompetenciák	<ul style="list-style-type: none"> Önálló tanulás Munkamódszerek, módszertani kompetenciák Kritikus gondolkodás és reflexió

7. A tantárgy célkitűzései (az elsajátítandó jellemző kompetenciák alapján)

7.1 A tantárgy általános célkitűzése	<ul style="list-style-type: none"> A tantárgy célja a nagy adathalmazok elemzéséhez, feldolgozásához szükséges módszerek bemutatása.
7.2 A tantárgy sajátos célkitűzései	<ul style="list-style-type: none"> Az elemzési módszerek fogalmainak és algoritmusainak ismerete: <ul style="list-style-type: none"> Adatbányászati alapfogalmak Társítási szabályok bányászata Legközelebbi szomszédok gyors keresése A MapReduce modell Link-analízis Tanuló algoritmusok nagy adathalmazokon Adatbányászat adatfolyamokból és gráfokból

8. A tantárgy tartalma

8.1 Előadás	Didaktikai módszerek	Megjegyzések
1. Adatbányászati fogalmak, statisztikai modellezés, a Bonferroni-elv.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
2. Társítási szabályok bányászata nagy adatbázisokban: vásárlói kosár elemzése, az Apriori algoritmus.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
3. Legközelebbi szomszédok hatékony keresése: a Locality-Sensitive Hashing (LSH) metódus.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
4. Az LSH módszer elmélete.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
5. Osztott fájlrendszerek és a MapReduce modell.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
6. MapReduce algoritmusok és a MapReduce kiterjesztései.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
7. Tanuló algoritmusok nagy adathalmazokon: k-	Előadás, párbeszéd,	

legközelebbi szomszéd módszere (kNN), perceptron algoritmus.	problematizálás, tanári magyarázat	
8. Tanuló algoritmusok nagy adathalmazokon: szupport vektor gépek.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
9. Link-analízis: PageRank, topic-sensitive PageRank, link spam, HITS (Hubs and Authorities).	Előadás, párbeszéd, problematizálás, tanári magyarázat	
10. Adatbányászat adafolyamokból.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
11. Klaszteranalízis.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
12. Dimenzióredukciós módszerek: az SVD és CUR mátrixdekompozíciók.	Előadás, párbeszéd, problematizálás, tanári magyarázat	
13-14. Adatbányászat gráfokból: szociális hálók elemzése, szociális hálók klaszterezése, közösségek felfedezése, gráfparticionálás, a Simrank algoritmus, háromszögek számlálása, gráfok szomszédsági tulajdonságai.	Előadás, párbeszéd, problematizálás, tanári magyarázat	

Könyvészet

- [1] RAJARAMAN A., LESKOVEC J., ULLMAN J.D. *Mining of Massive Datasets*. Cambridge University Press, 2011.
- [2] BOTTOU L., CHAPPELLE O., DECOSTE D., WESTON J. *Large-Scale Kernel Machines*. MIT Press, 2007.
- [3] LIN J., DYER C. *Data-Intensive Text Processing with MapReduce*. Morgan & Claypool Publishers, 2010.
- [4] BALDI P., FRASCONI P., SMYTH P. *Modeling the Internet and the Web. Probabilistic Methods and Algorithms*. Wiley, 2003.
- [5] COOK D.J., HOLDER L.B. *Mining Graph Data*. Wiley, 2007.
- [6] SHAKHNAROVICH G., DARRELL T., INDYK P. *Nearest-Neighbor Methods in Learning and Vision. Theory and Practice*. MIT Press, 2006.

8.2 Szeminárium / Labor	Didaktikai módszerek	Megjegyzések
1. A véletlen hipersík alapú Locality-Sensitive Hashing módszer.	feladat, egyéni munka, párbeszéd	
2. Adatok indexelése.	feladat, egyéni munka, párbeszéd	
3. MapReduce alapú rendszerek: Hadoop.	feladat, egyéni munka, párbeszéd	
4. A PageRank algoritmus.	feladat, egyéni munka, párbeszéd	
5. Az SVD és CUR mátrixdekompozíciók.	feladat, egyéni munka, párbeszéd	
6. Az SVM optimalizálási feladatának megoldása ritka adatok esetén.	feladat, egyéni munka, párbeszéd	
7. Az SMO algoritmus.	feladat, egyéni munka, párbeszéd	
Könyvészet		
[1]–[6] +		

- [7] LANGVILLE A.N., MEYER C.D. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.
- [8] PERERA S., GUNARATHNE T. *Hadoop MapReduce Cookbook*. Packt Publishing, 2013.
- [9] HAN J., KAMBER M. *Adarbányászat. Konceptiók és technikák*. Panem, 2004.
- [10] <http://www.stanford.edu/class/cs246/handouts.html>

9. Az episztemikus közösségek képviselői, a szakmai egyesületek és a szakterület reprezentatív munkáltatói elvárásainak összhangba hozása a tantárgy tartalmával.

- Az előadás felépítése megegyezik a Stanford-on oktatott "Mining massive data sets" c. tantárgyával (<http://www.stanford.edu/class/cs246/>).
- A kurzus fontos fogalmakat és algoritmusokat mutat be, melyek szükségesek a rohamos mértékben bővülő, nagy adathalmazok feldolgozásához.

10. Értékelés

Tevékenység típusa	10.1 Értékelési kritériumok	10.2 Értékelési módszerek	10.3 Aránya a végső jegyben
10.4 Előadás	Írásbeli vizsga a félév végén	Írásbeli vizsga	60%
10.5 Labor	Programozási feladatok bemutatása	A megoldások pontozása	40%

10.6 A teljesítmény minimumkövetelményei

- Kötelező a pontok felének összeszedése minden kiértékeléskor (évközi kiértékelés (laborgyakorlatok), végső vizsga).

Kitöltés dátuma

Előadás felelőse

Labor / praktika felelőse

2024. ápr. 15.

Dr. Darvay Zsolt, egyet. docens

Dr. Darvay Zsolt, egyet. docens

Az intézeti jóváhagyás dátuma

Intézetigazgató

2024. ápr. 15.

Dr. András Szilárd, egyet. docens