MAGYAR TUDOMÁNYOS AKADÉMIA
MTA IX. OSZTÁLY Statisztikai és Jövőkutatási Tudományos Bizottság

MEGHÍVÓ

Az MTA IX. Osztály Statisztikai és Jövőkutatási Tudományos Bizottságának
a Magyar Tudomány Ünnepére rendezett tudományos ülésére

2020. november 17. 10:00-13:00

online kapcsolódás:
https://us02web.zoom.us/j/89158118578?pwd=bTNieGZhK0JtelJrcmFCZ1FVVnF6UT09

„A mesterséges intelligencia és a gépi tanulás alkalmazásának kihívásai”

Program:

10:00-10:10 Megnyitó – *Sándorné Kriszt Éva (Ph.D, egyetemi tanár, BGE) a SJTB elnöke, levezető elnök*

10:10- 10:35 Mesterséges intelligencia és gépi tanulás a hitelkockázat előrejelzése területén
Kristóf Tamás *(Ph.D, egyetemi docens, BCE)*

10:35-11:00 Online forgatókönyvírás és a mesterséges intelligencia kapcsolata
Retek Mihály (PhD hallgató BCE)

11:00-11:25 A székesfehérvári vállalatok jövőjének technológiai és innovációs aspektusai
Márton András (PhD hallgató BCE)

II. szekció Levezető elnök: Hideg Éva *(MTA doktora, egyetemi tanár, BCE) az SJTB társelnöke*

11:25-11:50 A statisztika jövője a szoftverfejlesztők szemszögéből – Python, R
Daróczi Gergely (Ph.D, R fejlesztő, Rapporter)

11:50-12:15 A statisztika jövője a szoftverfejlesztők szemszögéből – Answerminer
Borbély Bence, Kalmár Péter, Magyar Nándor (Answerminer)

12:15-12:40 Gépi tanulás társadalomkutatói perspektívából
Németh Renáta (Ph.D, egyetemi docens, ELTE) – Raskovics Márton (ELTE)

12:40-13:00 Kérdések és reflexiók

Budapest, 2020. október 21.

Hideg Éva s.k.
SJTB társelnöke

Sándorné Kriszt Éva s.k.
SJTB elnöke

Elnök: Sándorné Kriszt Éva
Társelnök: Hideg Éva
http://www.ksh.hu/mta_sjtb

Titkár: Szilágyi Roland
+3630/2897612
roland.szilagyi@uni-miskolc.hu

Előadások összefoglalói

Mesterséges intelligencia és gépi tanulás a hitelkockázat előrejelzése területén
Kristóf Tamás (Ph.D, egyetemi docens, BCE)

A csődelőrejelzés egy bináris klasszifikációs probléma, amelynek lényege, hogy minél pontosabban különbséget tudjunk tenni a fizetőképes és a fizetésképtelen vállalatok két csoportja között. A csődelőrejelzés a vállalati pénzügyek, illetve a statisztika (adatbányászat) határtudományának tekinthető, amely a pénzügyi mutatószámokat magyarázó változóként felhasználva tesz kísérletet a vállalatok jövőbeli fizetőképességének előrejelzésére, arra alkalmas többváltozós módszerek alkalmazásával.
A hazai csődelőrejelzés mintegy 30 éves fejlődéstörténete alapján kijelenthető, hogy az napjainkra elérte a nemzetközi szakirodalom és gyakorlat színvonalát, a vizsgált kutatási kérdések, az alkalmazott módszerek és az empirikus eredmények tekintetében egyaránt.
A hazai csődelőrejelzés fejlődéstörténetében nyomon követhető az a fejlődési út, amely az egyszerűbb, keresztmetszeti adatokból felépülő, kisebb mintákon, klasszikus módszertanokkal történő csődelőrejelzéssel kezdődött, és napjainkra eljutott a dinamikus, through-the-cycle szemléletű tőkemodellek követelményeinek megfelelő vállalati minősítő rendszerek kialakításáig.
A szakterület aktuális elméleti, módszertani és gyakorlati kihívásait a gépi tanulás, az adatbányászat és a mesterséges intelligencia, valamint az új eljárások egymással történő kreatív kombinációjával a hibrid modellépítés dominálja. Az előadás célja átfogó képet nyújtani a kutatási terület legfrissebb eredményeiről, a hangsúlyt a hazai empirikus modellfejlesztésre helyezve.

Online forgatókönyvírás és a mesterséges intelligencia kapcsolata
Retek Mihály (PhD hallgató BCE)

Évről - évre egyre több forgatókönyvet készítenek a világban. Azok hasznosítása szempontjából fontossá válik a forgatókönyvek informatikai eszközökkel segített és online készítése. Ennek kapcsán a stakeholderek bevonásával készülő forgatókönyvek belső konzisztenciájának biztosítása, valamint a nagyszámú forgatókönyv feldolgozása és hasznosítási szempontú értelmezése válik hangsúlyossá. Ez jelentheti a participatív és online forgatókönyvírás folyamatában megjelenő egyszerűbb szövegfeldolgozó és elemző statisztikai módszerek használatát, valamint a forgatókönyvek értelmezésében és elemzésében a szövegbányászati (text mining) módszerek alkalmazását. E módszerek segítenek annak feltárásában, hogy a logikai és a tartalmi kapcsolatok konzisztensek-e a forgatókönyvírás teljes folyamatában. A folyamat lezártával más módszerek javaslatokat adhatnak a megrendelőknek, hogy az elkészített forgatókönyvek mennyire megbízhatóak a számukra.
Egyszerű statisztikai módszerek informatikai megoldásokkal segített vagy online forgatókönyvírásba épített alkalmazásával megoldható a legtöbbet kiválasztott hajtóerők, a legértékesebb hajtóerők, a legbefolyásosabb tengelyek és az értékes szcenáriók kiválasztása. A komplexebb alkalmazások már szövegbányászati módszereket is tartalmaznak. Ezekkel kimutathatók a forgatókönyvek és a hajtóerők közötti kapcsolatok valószínűsége, kiszűrhetők

a forgatókönyvekben levő haszontalan információk, valamint azok pozitív és negatív töltöttsége, továbbá elvégezhetők az egyes forgatókönyvek idő alapú elemzése.
Annak érdekében, hogy könnyű legyen a megrendelőknek is áttekinteni a kimeneti forgatókönyveket, vizuális ábrák segítségével is támogatni kell a teljes folyamatot. Ilyen ábrák lehetnek a forgatókönyvek és a hajtóerők szófelhői (word cloud), a hajtóerők fogalmainak megjelenése a forgatókönyvekben (lexical dispersion), a fontosabb fogalmak megjelenésének aránya (frequency).
Az előadás angol nyelvű példákon – köztük az előadó által alkalmazott és továbbfejlesztett változatokon - keresztül mutatja be a fent említett módszereket, amelyekkel a forgatókönyvírás készítésében alkalmazható módszerek és a kész forgatókönyv elemzések fél-automatikusan vagy teljesen automatikusan elvégezhetőek.

A székesfehérvári vállalatok jövőjének technológiai és innovációs aspektusai
Márton András (PhD hallgató BCE)

Székesfehérvár erőssége többek között a város hat ipari parkjában tömörülő termelő vállalatok gazdasági ereje. Az itt működő nagyvállalatok számos modern technikai eszközt (pl. napelemeket, autós hangrendszereket, motoralkatrészeket) gyártanak, a kisvállalatok pedig általában egy kisebb szegmenst próbálnak kiszolgálni. Kutatásunkban feltártuk a KKV-k és nagyvállalatok innovációs, technológiai és munkaerőpiaci kihívásait is, különös tekintettel a mind nagyobb hangsúlyt kapó fenntarthatósági kritériumokra. Általánosságban elmondható, hogy a helyi vállalatok innovációs gyakorlatát nagymértékben befolyásolja a vállalatméret és a tulajdonosi kör: a nagyvállalatok sok esetben élenjáró technológiát alkalmaznak, de a külföldi tulajdonosok megtartják a kutatás-fejlesztést és innovációt az anyavállalat feladatának; míg a kisvállalatok sokszor kevésbé jó (illetve modern) technológiai háttérrel indulnak, de helyi, országos vagy európai uniós pályázati forrás felhasználásával lehetőségük nyílik innovációra. Ez a különbség a fenntartható fejlődéshez való vállalati hozzáállásban is tükröződött, hiszen a nagyvállalatok közül az bizonyult „zöldebbnek", amely tulajdonosai (akár magyarok, akár külföldiek) eleve érdeklődőek, támogatóak a fenntarthatóság irányában, ugyanakkor a KKV-k közül az a meghatározó szempont, hogy valamilyen forrásból meg tudják-e oldani a nagyobb volumenű zöld beruházások (komplex energetikai korszerűsítés, napelemek felszerelése stb.) finanszírozását. Mérettől függetlenül megegyeztek abban a székesfehérvári cégek döntéshozói, hogy a munkaerő elvándorlása, a szakképzett munkaerő hiánya olyan problémákat fog okozni, amelyre csak részben tud megoldást nyújtani a fejlődő, közte a mesterséges intelligenciát is alkalmazó ipari technológia.

A statisztika jövője a szoftverfejlesztők szemszögéből – Python, R
Daróczi Gergely (Ph.D, R fejlesztő, Rapporter)

A nyílt forráskódú adatelemző eszközök megjelenése és elterjedése új lehetőségeket nyitott a szoftverfejlesztők számára, és alapjaiban változtatta meg a szoftverfejlesztés folyamatát. Előbb az adatok feldolgozását végző algoritmusok megírása egyszerűsödött a mások által megírt programkódok felhasználásával, majd ez egyre jobban általánossá vált, és megjelentek a statisztikai módszereket, különböző adatvizualizációs megoldásokat implementáló modulok is, amelyek alkalmazására akár nyelvtől függetlenül is lehetőség nyílt (például R vagy Python

programkód meghívása egyéb programnyelvekből vagy programokból). Később a nyílt forráskódú eszközökből sokszor üzleti alapokon nyugvó szolgáltatás is született terméktámogatással és vállalt rendelkezésre állással együtt, amelyek tovább egyszerűsítették a statisztikai és egyéb módszerek integrációját. Ezzel párhuzamosan a ma fejlesztett szoftverek meghatározó része is szolgáltatásként vált elérhetővé a felhasználók számára, és az említett integrációk segítségével néhány kattintással végezhető például A/B tesztek online kiértékelése Thompson-mintavétel segítségével vagy MCMC szimuláció akár mobiltelefon indítva.

A statisztika jövője a szoftverfejlesztők szemszögéből – Answerminer

Borbély Bence, Kalmár Péter, Magyar Nándor (Answerminer)

Előadásunkban bemutatjuk azokat a konkrét ötleteinket és fejlesztéseinket, amelyeket beépítettünk az AnswerMiner-be, annak érdekében, hogy könnyebbé és érthetővé tegyük az adatelemzést. Sokszor szembesülünk azzal a problémával, hogy egy bonyolult és összetett statisztikai műveletet kell a felhasználó számára egyszerűen végrehajthatóvá tennünk. Ennek megoldásába szeretnénk betekintést nyújtani, valamit megosztani, hogy adatelemző szoftvert fejlesztő cégként, hogyan látjuk a statisztika és az adatelemzés jövőjét.

Gépi tanulás társadalomkutatói perspektívából

Németh Renáta (Ph.D, egyetemi docens, ELTE) – Raskovics Márton (ELTE)

A gépi tanulás társadalomkutatási felhasználása nem önmagában álló módszertani újításként, hanem az információs társadalom megjelenéséből és az azt kísérő új adatforrásból („big data", „found data") következő paradigmatikus változás. Előadásunkban az új paradigma jellegzetességeit tekintjük át, a sajátos adatforrás, elemzési módszer és kutatási logika tekintetében. Áttekintésünk során valós társadalomkutatási példákra támaszkodunk, és felvillantjuk az új nézőpontból adódó lehetőségeket, a hagyományos és az új közötti kapcsolódási pontok hangsúlyozásával.

A korábbiakhoz képest eltérő adatforrások lehetőségeinek és korlátainak kiegyensúlyozott értékelése kialakulóban van. A kezdeti „n=all" és „numbers speak for themselves" eufóriáját követő kijózanodás lehetőséget teremtett a módszertani problémák szisztematikus vizsgálatára. Amíg a mintavételes (survey-)kutatások során standard eljárásokra lehet támaszkodni, addig a talált, organikus adatok esetében még kérdéses a megfelelő adatminőség definíciója, mérése. Az látszik, hogy a reprezentativitás és az érvényesség szempontjai big data esetén sem kevésbé kritikusak, mint korábban.

Az adatelemzési módszer tekintetében is vannak eltérések a klasszikus statisztikai szemlélet és a gépi tanulás között. Leo Breiman sok vitát kiváltó Two cultures című cikke ma is aktuális ebben a tekintetben: Mit gondolunk az adatgeneráló mechanizmusról? Használunk-e valószínűségi modellt, vagy sem? Milyen jellegű összefüggésekre számítunk és milyeneket tudunk feltárni?

A fentiekre adott válaszok függvényében a kutatási logikában is eltérések vannak. Milyen szerepet játszik, hogy önbevallásos adatok helyett megfigyeléses adatokkal dolgozunk? Egy elméletorientált kutatási logikában az elmélet által motivált hipotézisek tesztelésére lehet szabni az adatfelvételt. Az adatvezérelt esetben meg lehet-e, meg szabad-e ezt fordítani? Az adatmennyiség növekedése sokszor együtt jár a jobb predikciós teljesítményű modellekkel, de ez nem feltétlenül jelenti az oksági kapcsolatok jobb megértését.

Az egyik legvilágosabb jele annak, hogy az adatok nem beszélnek magukért, hogy a legnagyobb sikereket elért gépi tanulási modellek továbbra is felügyelt tanuláson alapulnak, ami csak növelte a „címkézett" adatok iránti igényt. Ezért is aktívan kutatott téma az „active learning", az adatok jól célzott címkézése, valamint a mesterséges intelligenciára épülő alkalmazások tanító adatainak szakértői összeállítása, illetve azért is, mert több – a big data-t nem elég körültekintően kiaknázó – ambiciózus alkalmazásról derült ki, hogy a tanuló adatokban, vagy az algoritmusokban rejlő torzítás miatt ténylegesen káros következményekkel jártak.