



# SPSS Nyári Iskola 2023.



**Pannon Egyetem Mérnöki Kara**  
Veszprém

Statistical Products Hungary Kft.  
2023. július 3-7.

ISBN 978-615-01-7590-4

**Az SPSS Nyári Iskola 2023 rendezvény szervezője:**

Clementine/Statistical Products Hungary Kft.

1115 Budapest, Bartók Béla út 105-113. 1/b.

**A rendezvény házigazdája:**

Körmendi György Olivér

Clementine

ügyvezető igazgató

**Szerkesztette<sup>1</sup>:**

Izsán Orsolya, Keresztesi Ildikó, Pancza Judit

<sup>1</sup>oktatas@clementine.hu

## Tartalomjegyzék

„Helyzet van”?! A statisztika oktatásának kulcs-szerep jut(hat) az oktatási rendszerünk megújításában!

*Pótó László*.....4

Az adatelemzés oktatásának néhány módszertani problémája

*Nyéki Lajos*.....6

SZIGNIFIKANCIA VÁLSÁG? Alternatív utak a Bootstrappingtól a Bayes Theoremán át a

Markov Chain-MonteCarlóig

*Jánosa András*.....7

Térökonometriai modellezés a GeoDa-val és az R programcsomaggal

*Csonka Arnold*.....8

Adatvezérelt Kékszalag

*Gergely Norbert*.....9

A BRCA mutáció kimutatása emlődagatos betegeknél – a logisztikus regresszió egy klinikai alkalmazása

*Pósa Szonja Polett, Nikolényi Alíz, Varga Zoltán, Szűcs Mónika*.....10

„Data Science @ Business Intelligence: „SPSS-em és egyéb állatfajták” – Világgazdasági és munkaerőpiaci trendek az adatvezérelt technológiák hátterében

*Keszi Roland*.....11

Kvalitatív tartalomelemzés és topikmodellezés egy darknet piac feltáró vizsgálatában

*Szigeti Ákos*.....12

A dezinformációs hadviselés elleni küzdelem mesterséges intelligenciával

*Richard Frank, Barry Cartwright*.....13

## Elvágópont kereső eljárások összehasonlítása háromdimenziós ROC analízis esetén

*Szűcs Mónika, Boda Krisztina, Bari Ferenc*.....14

## Túléléselemzés és asszociációs szabályok: Egy integrált megközelítés az egyetemi hallgatók lemorzsolódásának vizsgálatában

*Csalódi Róbert, Abonyi János*.....15

## PLS-SEM módszer alkalmazása a digitális térben

*Hargitai Dávid Máté*.....16

## Hálózatalapú modell- és adatredukciós módszer

*Kosztján Zsolt Tibor, Kiss Dénes, Fehérvölgyi Beáta*.....17

## Gépi tanulási modellek értelmezhetősége

*Ipkovich Ádám, Abonyi János*.....18

## Sok a szöveg?! Olvass inkább a sorok közt!

*Molnár Anna Enikő, Tamási-Mészáros Evelin*.....19

## i2 hálózatelemző eszközök áttekintése és az i2 Academic program

*Pancza Judit*.....22

## A hálózatos elemzések adat schema tervezése iBase alkalmazásban

*Csatlós Béla*.....23

## „Helyzet van”?!

### A statisztika oktatásának kulcs-szerep jut(hat) az oktatási rendszerünk megújításában!

Pótó László

Pécsi Tudományegyetem, Általános Orvostudományi Kar, Bioanalitikai Intézet, Orvosi Statisztika és Informatika Tanszék, Pécs, Magyarország, laszlo.poto@aok.pte.hu

*Kulcsszavak: első statisztika kurzus, elitképzés-tömegoktatás, kiválóság*

Évtizedekkel ezelőtt jelent meg a statisztika oktatásának megújításáról a „Moore-féle” program. (1. Cobb, 1992) Az elmúlt évek során többször tartottam e témáról előadást az SPSS Nyári Iskolán. A legelső címe az volt, hogy „Baj van-e a statisztika oktatással?” Majd ezt a témát többször folytattam a Moore-féle program általam megvalósított koncepciójának, konkrét megoldásainak bemutatásával. Egy 3 lépéses megoldási sémát javasoltam, amely a statisztikán túl tulajdonképpen bármely tantárgy képzésének a megújítására alkalmas lehet. Mostanában (különösen a COVID időszak online oktatásának tapasztalatai alapján) jelentősen felerősödött a közgondolkodásban már régóta jelen levő „vészjelzés”, hogy baj van a közoktatásunkkal, és a felsőoktatásunkkal egyaránt! Egy igen elgondolkodtató „tünetegyüttest” mutat erről egy viszonylag friss ÁSZ-elemzés (2. Németh és mások, 2022) Az elmúlt évek során a statisztika oktatásának megújítására használt koncepciót (3. Gerő, 2008) alkalmaztam (egyelőre csak elméletben) erre a problémára, és úgy tűnik, működhet! Sőt, pozitív visszacsatolásként kiteljesítette az „első statisztika kurzus”-ra alkalmazott megoldásomat (mely lényegében a Moore-program továbbfejlesztése). Ezt szeretném most bemutatni a kollégáknak, amely további 2 (immáron kari szintű) integrációs és fejlesztő lépést ad a már meglevő tantárgyi programhoz. Bár a vázolt koncepció elsősorban a felsőoktatás megújítását célozza, úgy tűnik, a közoktatás problémáinak megoldására is segítséget, támpontot nyújthat ez a gondolatmenet, ez a probléma-megoldó séma.

Ráadásul kiderült, hogy ebben a koncepcionális megújításában kulcsszerepet kaphat az első statisztika kurzusok statisztika oktatása. (Igazán lelkesítő kifejelet – ha megvalósulna –, ugye?) Ez könnyen érthető, ha belegondolunk, hogy az „első statisztika kurzus” által elsődleges célként kitűzött kimeneti eredmények egyúttal alapvető bemeneti feltételt jelentenek sok-sok felsőoktatási képzésünk oktatási programjában. Továbbá kimeneti céljai számos felsőoktatási szakmai képzésnek a mezőgazdaságtól a természet- és társadalomtudományokon át a műszaki felsőoktatásig, sőt, az orvosi képzésig. Azonban mindezen képzésekben ezeket a kiemelt kimeneti célok közt szereplő kompetenciákat fő tantárgyi célként talán egyedül a statisztikai kurzus fejleszti, (fejlesztjené).

Azért szükséges a kettős (kijelentő és feltételes) szóhasználat, mert ugyancsak az elmúlt évtizedek változása, hogy a felsőoktatás korábbi „elitképző” jellegét tömeges képzés váltotta fel. A felső-oktatásunk kiemelt problémájának tartom, hogy az egyetemi oktatási módszerek, eszközök és tananyag-koncepciók máig a régi alapokon nyugszanak. Márpedig a tömegképzés és az elitképzés minden fontos jellemzőjében eltérő. Ezt a problémát úgy javaslom kezelni, hogy a tantárgyak, tananyagaikban, követelményeikben és módszereikben a kétféle képzési cél szerint kettős (általam önkényesen A-val és B-vel jelölt) programmal oktassanak. Ez valójában azt jelenti, hogy a tömegképzési oktatásunkba integráljunk egy célzott elitképzést az oktatási-kutatói utánpótlásunk képzésére a teljes felsőoktatási kurrikulumban.

Az előadás egy 6 lépésből álló „receptet” javasol, ami egyfajta útvonal-séma lehet nemcsak a statisztika, hanem az egész felsőoktatás megújításához mindezen szakterületeken. A javasolt megoldás illusztrálására az előadásba illesztett kis „gyakorlatként” bemutatom, hogy egy pici statisztikai tananyag-elem oktatásában mi lenne a B (basic) tananyag, tanulási követelmény, és mi lenne a félév során újra és újra előjövő A (advanced) szintű követelmény és tananyag – ez utóbbi természetesen opcionálisan.

Ez egy vita-anyag, amint a korábbi előadásaim is azok voltak. Kíváncsian várom a reakciókat, és különösen a megvalósításban alapvető csapatmunkához szívesen csatlakozók jelentkezését!

#### Referenciák

1. Teaching Statistics, Cobb, G. in L. A. Steen (ed.) *Heeding the Call for Change: Suggestions for Curricular Action*, MAA Notes 22. (1992) Washington,DC: Mathematical Association of America.  
[https://www.maa.org/sites/default/files/pdf/CUPM/first\\_40years/1992Cobb.pdf](https://www.maa.org/sites/default/files/pdf/CUPM/first_40years/1992Cobb.pdf)
2. „Pink education” jelenség Magyarországon?! – A felsőfokú végzettséggel rendelkező nők túltreprezentáltságának tényezői és gazdasági-társadalmi hatásai; Dr. Németh Erzsébet, Füzi Beatrix, Pats Regina, Puskás Balázs, Váradi Eszter, Állami Számvevőszék, 2022. *július*, publ. szám: T/600  
[https://www.asz.hu/dokumentumok/E2202\\_A\\_felfok\\_u\\_vegzetts.pdf](https://www.asz.hu/dokumentumok/E2202_A_felfok_u_vegzetts.pdf)
3. Az élethelyzethez igazított tanulás; Gerő Péter, egyetemi tankönyv, Zrínyi Miklós Nemzetvédelmi Egyetem, 2008, ISBN: 978-963-7060-54-0

## Az adatelemzés oktatásának néhány módszertani problémája

Nyéki Lajos

Széchenyi István Egyetem, Győr, Magyarország, nyeki@sze.hu

*Kulcsszavak: adatelemzés, kutatómódszertan, Excel, PSPP, JASP*

A pedagógiai kutatás módszertana az egyetemi szintű mérnöktanár-képzés fontos tantárgya. A tárgy heti órakerete egy óra előadás és egy óra gyakorlat a nappali tagozatos képzés hatodik félévében. A gyakorlatokon oktatási statisztikai témaköröket dolgozunk fel számítógépes szoftverek felhasználásával. A hallgatók használhatják a Microsoft Office 365 programcsomag részeként az Excel programot, és letölthetik a hálózatról a szabadon felhasználható PSPP és JASP szoftvereket. A tantárgyhoz készült jegyzet az elméleti témakörök mellett külön fejezetben mutatja be a gyakorlati statisztikai számításokat az Excel adatelemző eljárásaival. Ezek a következők: leíró statisztika, egymintás t-próba, párosított mintás t-próba, kétmintás F-próba, kétmintás t-próba, egytényezős variancia-analízis, korreláció és regresszió. A tantermi gyakorlatokon minden hallgató a saját laptopját vagy számítógépét használja. A félév során a hallgatóknak otthoni munkával el kell készíteni egy-egy választható témájú számítógépes adatelemzési feladatot. A tantermi gyakorlatokon mindhárom programot használjuk. Így meg tudjuk mutatni az egyes szoftverek által igényelt bemeneti adattáblák közötti különbségeket és az egyes szoftverek által nyújtott eltérő szolgáltatásokat, elemzési lehetőségeket. A tantermi gyakorlatok témakörei a következők: centrális tendencia, variabilitás, normális eloszlás és z-pontok, z-eloszlás és valószínűségek, z-próba, egymintás t-próba, kétmintás t-próba, variancia-analízis, korreláció és regresszió. A kísérletek tervezése témakörben a kétcsoportos kísérleti terv, a kétcsoportos kovariancia terv, a véletlen blokk terv, a 2x2 faktoriális terv és a  $2^{3-1}$  típusú rész faktor terv regressziós modelljeit ismertetjük. A mintavétel módszerei témakörben bemutatjuk az egyszerű véletlen mintavétel Excel által nyújtott lehetőségeit is. A félév során egy alkalommal röviden ismertetjük a Bayes-féle következtetés alapjait is. Ennek során összehasonlítjuk a Bayes-féle és a gyakorisági megközelítés jellemzőit, ismertetjük a Bayes-féle terminológiát, levezetjük a Bayes-faktort és bemutatjuk a Bayes-féle elemzések diagramjainak értelmezését. A szűkös időkeret miatt csak a t-próbák, a korreláció és az egyszerű lineáris regresszió esetében van lehetőség a Bayes-féle módszerek gyakorlati bemutatására.

# SZIGNIFIKANCIA VÁLSÁG?

## Alternatív utak a Bootstrappingtól a Bayes Theoremán át a Markov Chain-MonteCarlóig

Jánosa András

Budapesti Gazdaságtudományi Egyetem, Módszertan tanszék, Budapest, Magyarország,  
ja8206@gmail.com

*Kulcsszavak: Frequentist inference, Visszatevées módszerek (Replacing Algorithm), Bayes, MonteCarlo szimuláció, Markov(chain)*

A statisztikus “világon” belül, még mielőtt kirobbant volna, már folyt egy szakmai vita a nullhipotézis szignifikancia teszt (NHST) néven emlegetett problematika körül. Valójában azonban 2016-ban robbant ki, amikor a Basic and Applied Social Psychology nagytekintélyű folyóirat (a neve ellenére a matematikai statisztikai módszertant követő elemzések egyik legtekintélyesebb folyóirata) 2016-ban megtiltotta a publikálni kívánók számára a nullhipotézis szignifikancia tesztek használatát. Ez persze nagy megdöbbenést váltott ki. Természetesen a magyar statisztikai tudományos közéletben is azonnal felszínre került a vita. Számos, mindeddig csak lappangó vitapontot hozott felszínre. Érdekes viszont, hogy ezek mellett bekerültek a szakmai köztudatba olyan tézisek, melyekről korábban a szakmai közélet keveset tudott, mintegy „alvó tézisek” voltak, annak ellenére, hogy a vitában a problémásnak mondott elemek helyett mintegy helyettesítő, „javító” szerepet tölthetnek be.

A Brad Efron tanulmányában javasolt „újramintavételezési” technikákra úgy tekintenek, mint amelyek alkalmasak az összefüggések megbízhatóságának javítására.

Ezek közül a Jackknife eljárás a legelterjedtebbek egyike. A módszer lényege, hogy minden egyes lépésben az eredeti mintából elhagy egy vagy több elemet (1-törléses jackknife, vagy  $d$ -törléses jackknife), így képezve a másodlagos mintákat.

A Bootstrap módszer is Efron nevéhez kapcsolódik. Leginkább akkor használják, amikor a statisztika eloszlása ismeretlen vagy a normalitás feltételei nem teljesülnek. Ráirányította a figyelmet arra is, hogy a számítógépek kapacitásának növekedésével a replikációs eljárások elméletileg is új megközelítéseket adnak: az ismételt mintavétel lehetővé tette az analitikus formulák helyett számítógépes módszerek alkalmazását. Esetleg enyhíthet a „minta problémán” is, hasznos lehet, ha nincs kellően nagy mintánk, vagy nem tudjuk a teszt adatokon megismételni az eljárást.

A Bayes Theorema a klasszikus, ún. frequentista statisztika egyes problémáira kínál új megoldásokat. Felfogása szerint a becsléstárgya nem egy rögzített érték, hanem valószínűségi változó. Megengedi a szubjektív valószínűséget is.

A Markov Chain MonteCarlo (MCMC) eljárás egyre népszerűbb módszer az eloszlásokkal kapcsolatos információk megszerzésére, különösen a Bayes-i következtetések utólagos eloszlásainak becslésére. Lehetővé teszi az eloszlás jellemzését anélkül, hogy ismernénk az eloszlás összes matematikai tulajdonságát, véletlenszerűen mintavételezve az értékeket az eloszlásból.



## Térökonometriai modellezés a GeoDa-val és az R programcsomaggal

Csonka Arnold

Magyar Agrár- és Élettudományi Egyetem, Agrár- és Élelmiszertudományi Intézet Kaposvári Campus, Kaposvár, Magyarország, csonka.arnold@uni-mate.hu

*Kulcsszavak: térökonometria, térstatisztika, térbeli függőség, területi klaszterek, térbeli egyenlőtlenségek*

Az elmúlt évtizedekben jelentősen megnőtt a térökonometriai és térstatisztikai vizsgálatok jelentősége a közgazdaságtanban. A Paul Krugman által meghirdetett Új Gazdaságföldrajz (New Economic Geography - NEG) megnyitotta az utat a térnek, mint gazdasági dimenziónak a mainstream gazdaságtudományokban. Erre építve növekvő számban és minőségben jelennek meg a gazdasági-társadalmi jelenségek térbeli mintázatát feltáró empirikus kutatások.

Az utóbbi pár évben a térgazdaságtani kutatásoknak egy újabb lökést adott az Evolúciós Gazdaságföldrajz (Evolutionary Economic Geography – EEG) megjelenése. Az EEG az evolúciós közgazdaságtan és a gazdaságföldrajz elméleteit kombinálja. Célja a regionális alkalmazkodás, illetve a regionális egyenlőtlenségek dinamikájának megértése. Az EEG szerint a regionális gazdasági fejlődésben kiemelt jelentősége van az útfüggőségnek (path dependency), a térbeli bezáródásnak (lock-in), a földrajzi közelségnek (proximity), a klaszteresedésnek (spatial clustering) és a centrum-periféria kapcsolatoknak. Az EEG, mint a térségi/regionális fejlődés vizsgálatának komplex keretrendszere, kiváló elméleti alapot nyújt a térbeli egyenlőtlenségek dinamikájának kutatásához és megértéséhez. Ezen belül az útfüggőség, a bezáródás és az agglomerációs előnyök térbeli vizsgálata nemcsak hazai, hanem nemzetközi viszonylatban is újszerű és hézagpótló témának tekinthető.

Az előadás inspirációt, valamint módszertani iránymutatást kíván adni az újszerű, térszemléletű kutatások iránt nyitott hallgatóknak és kutatóknak a témába illeszkedő kutatásaik megkezdéséhez. Az előadás során röviden érintjük a térgazdaságtan elméleti alapjait, de a fő hangsúly természetesen a módszertani alkalmazásokon, illetve az azt támogató, szabadon elérhető szoftvereken van.

A prezentáció során gyakorlati példákon, esettanulmányokon keresztül válnak megismerhetővé a térstatisztikai és térökonometriai elemzések kiinduló kérdései, módszertani alapjai és főbb lépesei. Reményeink szerint a résztvevők az előadásnak köszönhetően

- jobban megértik a térstatisztikai és térökonometriai kutatások értelmét, célját és lehetőségeit,
- világos rálátást kapnak a térbeli kutatások módszertanára,
- inspirációt és induló segítséget kapnak ahhoz, hogy saját maguk is térbeli elemzéseket végezzenek.

## Adatvezérelt Kékszalag

Gergely Norbert

Clementine, ZM Sailing Team, Budapest, Magyarország, ngergely@clementine.hu

*Kulcsszavak: Vitorlázás, Kékszalag, adatvezérelt stratégia, meteorológia*

A Kékszalag 155 km-es távjával Európa legnagyobb és legrégebbi tókerülő vitorlásversenye, amin együtt versenyzik a 126 éves Kishamis és több csúcstechnológiát képviselő, hipermodern, karbonszálal katamarán. Az 55. Kékszalag Nagydíj pont a nyáriiskola idejére esik, 2023. július 6-án dördül majd el a rajtot jelző ágyúlövés a balatonfüredi mólón, ami után 48 óra áll az indulók rendelkezésére, hogy szintidőn belül körbehajózzák a Balatont. Idén kivételesen izgalmas versenynek nézhetünk elébe, ahol számos kiváló hajó és csapat áll rajthoz. A jó szerepléshez elengedhetetlen a megfelelő technika, jó szélstratégia, versenytaktika, illetve erő- és állóképesség. Az abszolút győzelemre is esélyes egységek mögött többnyire komoly csapat dolgozik, hogy tökéletes versenyt futhassanak. Meteorológia számítások, statisztikai előrejelzések, hónapokon át tartó edzés és tesztelés előzi meg a rajt pillanatát, hogy az adott körülményeknek leginkább megfelelő beállításokkal, a legideálisabb vitorlaszettekkel és minden eshetőségre felkészülten vágjanak neki a tókerülésnek. Minden évben izgalmas küzdelemben dől el, hogy ki kapja az abszolút győztesnek járó kék szalagot. Azonban a Kékszalag nem csak a csúcshajókról szól, sőt. A sokszor 600-nál is több induló hajó fedélzetén több ezren versenyeznek, ők adják a verseny valódi rangját, nélkülük nem lenne Kékszalag a Kékszalag. De hogyan készül a 30+ különböző kategória és hajóosztály, akik korlátozottabb lehetőségekkel rendelkeznek és akik mögött nem állnak elemzők, meteorológusok? Az előadás során betekintést nyerünk az egyik legfiatalabb nemzeti hajóosztály, a Balaton 25-ös osztály két egységének felkészülésébe és előzetes versenysztratégiájába. Megnézzük, hogy milyen adatok keletkeznek a ZM Sailing Team - Zenit & More és a Solaris Sailing Team - Solaris nevű hajóján. Hogyan lehet és hogyan érdemes gyűjteni és elemezni a különböző szenzorok adatait? Milyen versenyszabályoknak kell megfelelni? Hogyan áll össze egy adatvezérelt, előzetes versenysztratégia? És mit lehet tenni, amikor összedől rögtön a rajt utáni pillanatokban? A Balatonfüred-Balatonkenese-Siófok-Keszthely-Balatonfüred táv hosszú és változatos, nincs két ugyanolyan verseny. A 48 órás szintidő alatt találkozhatunk teljes szélcsenddel, tomboló viharral, zuhogó esővel és kínzó UV sugárzással is, de nincs megállás. A 155 km alatt nincs kikötés, nincs motorozás, nincs külső segítség. Csak a hajó, a csapat és a Balaton van, akár két éjjelen és két nappal át. Rajt: 3 nappal az előadás után, 2023. július 6-án, 9:00-kor. Jó szelet!

## A BRCA mutáció kimutatása emlődaganatos betegeknél – a logisztikus regresszió egy klinikai alkalmazása

Pósa Szonja Polett<sup>1</sup>, Nikolényi Alíz<sup>2</sup>, Varga Zoltán<sup>2</sup>, Szűcs Mónika<sup>1</sup>

<sup>1</sup>SZTE SZAOK-TTIK Orvosi Fizikai és Orvosi Informatikai Intézet, Szeged, Magyarország,  
posa.szonja.polett@o365.u-szeged.hu

<sup>2</sup>SZTE SZAKK Onkoterápiás Klinika, Szeged, Magyarország

*Kulcsszavak: BRCA, logisztikus regresszió, klinikai onkológia, változóselekcio*

### Bevezetés:

Az emlőrák a nők körében leggyakrabban előforduló malignus megbetegedés, amely kialakulását jelentősen növeli az örökletes BRCA (1/2) génmutáció. A BRCA mutációt hordozó betegek szűrése genetikai teszteléssel lehetséges, azonban a molekuláris genetikai vizsgálat időigénye és költségvonzata akadályokat támaszthat a diagnosztikai folyamatban.

### Célkitűzés:

Jelen retrospektív kutatás célja, hogy statisztikai módszerekkel meghatározza azon faktorokat, amelyek alapján előre jelezhető az adott, emlőrákkal diagnosztizált beteg BRCA mutációt hordozó státusza, ezzel indikálva a genetikai vizsgálat szükségességét.

### Betegek:

A kutatásban összesen 253 emlőrákkal diagnosztizált beteg adatai kerültek feldolgozásra az SZTE Onkoterápiás Klinika adatbázisa alapján, ebből 61 (24,11%) betegnél igazolódott a BRCA 1/2 mutáció megléte, míg 192 (75,89%) esetben a vad típus volt jelen.

### Módszerek:

Az egyes kliniko-patológiai jellemzők és a mutációt hordozó státusz közötti kapcsolat megállapítására egyszeres, valamint többszörös logisztikus regresszióanalízist alkalmaztunk R (version 4.2.3, <https://www.r-project.org/>) programozási nyelv segítségével. Először az egyes lehetséges magyarázóváltozók egyenkénti hozzájárulását vizsgáltuk a BRCA mutáció jelenlétére, majd backward és forward stepwise módszerekkel történő modellépítéssel több faktor együttes hatásáról nyertünk információt ugyanezen vonatkozásban.

### Eredmények:

Mind az egyszeres, mind a többszörös elemzés megerősítette statisztikailag szignifikáns p-értékekkel, hogy a családi anamnézis (elsőfokú rokon), a rizikófaktorok száma (>2) és a tripla-negatív emlőrák megléte megbízható prediktor a BRCA-pozitív mutációs státusz előrejelzésére. Az életkor változó beépítése a többszörös regressziós modellbe szintén szignifikáns eredményt hozott, ez esetben azonban az életkor 1 egységgel, azaz 1 évvel való növelése 0,95-szeresére csökkenti a BRCA mutáció kimutatásának esélyét, amelyből azon következtetés vonható le, hogy a fiatalabb életkor tekinthető további rizikófaktornak.

## **„Data Science @ Business Intelligence: „SPSS-em és egyéb állatfajták” – Világgazdasági és munkaerőpiaci trendek az adatvezérelt technológiák hátterében**

Keszi Roland

Eötvös Loránd Tudományegyetem, BGGyK, FOTRI, Budapest, Magyarország,  
keszi.roland@barczi.elte.hu

*Kulcsszavak: üzlet intelligencia, adattudomány, mesterséges intelligencia, munkaerőpiac*

Az adattudomány (DS) és az üzleti intelligencia (BI) – történetüket tekintve – olyan ökoszisztémában kiteljesedett fogalmak, melyeket a nemzetközi közgazdasági és szociológiai szakirodalom (a 2010-es évek óta) ipar 4.0-nak nevez. Ebbe az általános megnevezésbe olyan diffúz gazdasági és társadalmi jelenség halmaz tartozik, amely a világgazdaságot évtizedek óta jellemzi (ld. pl. BIG DATA, IOT, 3D nyomtatás, CLOUD COMPUTING). Ezeknek a jelenségeknek egy részére, illetve az ökoszisztéma működésének bizonyos érdekességeire gazdaság- és munkaszociológusok, munkaközgazdászok már a 80-as évektől kezdődően felhívták a figyelmet, mind Európában, mind az Egyesült Államokban, illetve – kisebb részben – Japánban. Az említett folyamatokra – szűk szakmai közönség érdeklődésén kívül – relative szerényebb figyelem irányult.

Annál több figyelmet szenteltek a kutatók a mesterséges intelligencia „nyarainak és teleinek” tanulmányozására, annak lehetséges társadalmi, gazdasági hatásainak vizsgálatára.

Üzleti oldalról a figyelem fókuszának megváltozásának szintén a 2010-es évek jelentett egyfajta fordulatot, hangsúly váltást, elsősorban a mély neurális hálózatok újabb áttöréseinek, illetve súlypont áthelyeződéseinek nyomán. Napjainkban ugyanennek a súlypont módosulásnak az éles megjelenését és termék fázisba forduló megjelenését láthatjuk a generatív nyelvi modellek üzleti folyamatokban való feltűnésével. A generatív nyelvi modellek üzleti intelligencia folyamatokba való további, szerves integrálódása, vagyis a DS és a BI egymásba fonódása a szemünk előtt zajlik.

A DS és a BI kiemelt témái a kvantitatív alapú tudományos, illetve gyakorlati, üzleti narratíváknak oktatási oldalról is. Neves egyetemek (pl. MIT, Harvard, Yale) és multinacionális vállalatok (pl. Google, IBM, Microsoft), de kisebb méretű magán cégek is sorra indítják képzéseiket, online és blended learning kurzusaikat külön-külön vagy a két téma ötvöztetésével. Egyes multinacionális vállalatok, természetesen saját termékeik, szolgáltatásaik értékesítéseinek ösztönzésére is alkalmazzák az említett oktatási tevékenységeket, például saját felhő szolgáltatásaik használata felé terelve a tanulni vágyó, vagy a munkáltatójuk által előírtan kiképzésre kötelezett fogyasztókat (ld. pl. MS Azure, Amazon AWS, IBM Watson). A DS és BI tehát kiemelten váltak nagy értékű oktatási árucikké is.

Az előadás ezeket a szövevényes gazdasági-társadalmi jelenségeket kíséri meg egységes keretbe ágyazva, közgazdaságtani és szociológiai eszközökkel áttekinteni. Az előadás az adat alapú megközelítésre fókuszál, a szocioökonómiai jelenségek mögött meghúzódó megatrendeket azonosít, középpontba helyezve munkaerőpiaci jelenségeket.

## Kvalitatív tartalomelemzés és topikmodellezés egy darknet piac feltáró vizsgálatában

Szigeti Ákos

Nemzeti Közszerológati Egyetem, Rendészettudományi Doktori Iskola, Budapest,  
Magyarország, szigeti.akos@uni-nke.hu

*Kulcsszavak: darknet, kábítószer, bizalom, kvalitatív tartalomelemzés, latent dirichlet allocation*

A társadalom digitalizálódása az illegális kábítószerpiacon is megmutatkozik: a 2010-es években a darknet piacokon zajló kábítószer-kereskedelem volumene folyamatosan emelkedett. A növekedés elsődleges oka a darknet piacok megbízható működése volt, melyet számos egymással összefüggő, közösségépítő (bizalmi) faktor befolyásolt. A darknetes kábítószerpiacok forgalmának jelenlegi, 2020-as évek elejétől érzékelt csökkenése főként a bizalmi tényezők egyikének, a korábban jól működő megbízható kézbesítés COVID-19 miatti megingásával, illetve annak az eladók és vásárlók közötti bizalmi kapcsolatra gyakorolt negatív hatásával magyarázható. Doktori kutatásom a darknet piacok nehezen elérhető, anonimitásra szerveződő közösségének közvetlen, feltáró vizsgálatát kvalitatív tartalomelemzés és Latent Dirichlet Allocation (LDA) topikmodellezés segítségével valósította meg. A kutatás első, kvalitatív fázisa a Dark0de Reborn darknet piacról gyűjtött termékleírásokat (n=100) és vásárlói értékeléseket (n=500) elemezte, majd a topikmodellezés során a piac vásárlói értékeléseinek egy nagyobb mintáját (n=26.728) vizsgálta. A két kutatási fázis eredményei összességében azt mutatják, hogy a darknet piacok köré egy bizalmon alapuló közösség szerveződik, mely a kábítószer-ellátás egy biztonságosabb formáját valósítja meg, csökkentve mind a fizetési és a szállítási szakaszban jelentkező kockázatokat, mind a kábítószer-használat potenciális ártalmait. Mindez a bizalom befolyásolására irányuló célzott intervenciók potenciális sikerére, valamint a biztonságosabb ellátási programok kezdeményezésének szükségletére irányítja a szakpolitika figyelmét. Mindemellett a vizsgált platform kvalitatív tartalomelemzés segítségével történő, dokumentált megismerése, valamint a topikmodellezés eredményeinek további elemzési lehetőségei rámutatnak a digitális társadalomkutatás többes- vagy vegyes módszertanú megközelítésében rejlő előnyökre, hozzájárulva a természetes nyelvfeldolgozást alkalmazó kutatások fejlődéséhez.

## A dezinformációs hadviselés elleni küzdelem mesterséges intelligenciával

Richard Frank<sup>1</sup>, Barry Cartwright<sup>2</sup>

<sup>1</sup>Associate Professor, Simon Fraser University, Director, International CyberCrime Research Centre, Simon Fraser University, Burnaby, Canada, rfrank@sfu.ca

<sup>2</sup>Associate Director, International CyberCrime Research Centre, Simon Fraser University, Burnaby, Canada, bcartwri@sfu.ca

*Kulcsszavak: dezinformáció, social media, mesterséges intelligencia*

A projekt célja egy olyan mesterségesintelligencia-eszköz kifejlesztése, amely lehetővé teszi az ellenséges szereplők által a social media-ban (Facebook, Twitter...) elkövetett dezinformációs támadások felderítését. A folyamat a dez-, téves- és valós-információk forrásainak manuális felfedezésével kezdődik a különböző közösségi média platformokon, majd kinyerjük a tartalmat, hogy létrehozzuk az „igazságot”, amelyre a gépi tanulási modellek épülnek. A kinyert tartalom szöveget, képet és videót tartalmazhat. Lementjük a videókat, például Facebook-ról, és elemezzük. Az algoritmus érzékeli a fontos jeleneteket, frame-eket. Más algoritmusok megnézik a videót, és megmondják milyen tárgyak, emberek, szöveg, tevékenységek vannak benne. A képek („frame”), amiket kiszedünk a videóból, és lementünk (például Facebook-ról) azokat tovább analizálunk. A szöveget, mely le lett kaparva vagy kiemelve a videókból és képekből ezután elemezzük, hogy kivonjuk a főneveket, témákat és érzéseket. Ha más nyelven van, akkor lefordítható, de a nyelvspecifikus árnyalatok elveszhetnek, ezért előnyben kellene részesíteni az anyanyelvi elemzést. Ez a tartalom alkotja az „alap igazságunkat”. A további hírcsatornákat ezután folyamatosan feltérképezi a rendszer új tartalmak után kutatva, ugyanúgy elemzi a videókat, képeket és szöveget. Az AI eszköz 6 gépi tanulási algoritmusból áll az algoritmikus torzítás elkerülése érdekében, amelyek mindegyike minden új bejegyzés esetében „szavaz” arról, hogy a bejegyzés dez-, téves- vagy valós- információ-e. Végül az összesített (súlyozott) szavazatokat az új dezinformációs támadások azonosítására használjuk, amelyek kulcsszavak és témafelismerés segítségével elemezhetőek, hogy megértsük a támadás irányát és forrását.

## Elvágópont kereső eljárások összehasonlítása háromdimenziós ROC analízis esetén

Szűcs Mónika, Boda Krisztina, Bari Ferenc

SZTE SZAOK Orvosi Fizika és Orvosi Informatika Intézet, Szeged, Magyarország,  
szucs.monika@med.u-szeged.hu

*Kulcsszavak: multi-ROC, biomarker, elvágópont, Youden-index, konkordancia-index*

Klinikai kutatások során gyakran feltett kérdés, hogy egy-egy biomarker alkalmas-e a kontroll és a betegcsoport szétválasztására. A ROC (Receiver Operating Characteristic) analízis széles körben elterjedt módszer az alkalmazott tesztek hatékonyságának elemzésére két dimenzió esetén. Számos módszerrel becsülhető a két csoportot legjobban elválasztó érték, ezek közül a leggyakrabban használt eljárások a Youden-index, konkordancia index és az „ideális ponttól” vett legkisebb távolság módszere. Az előadásunkban a kétdimenziós módszerek rövid ismertetése után bemutatjuk három és magasabb dimenziós módszereket, illetve azok összehasonlítását normál és gamma eloszlású generált adatokon, majd egy klinikumból származó példán is.

Kétdimenziós esetben a ROC görbe alatti terület (AUC) jellemzi a teszt diagnosztikus hatékonyságát:  $AUC=0,5$  jelenti, hogy a vizsgált marker alkalmatlan a csoportok elkülönítésére,  $AUC=1$  esetén tökéletes szétválasztásról beszélhetünk. Abban az esetben, ha a teszt diagnosztikus hatékonysága megfelelő, szükséges a legjobb elvágópont meghatározása. Számos elvágópont-kereső eljárás közül az orvostudományban a legelterjedtebb a Youden-index. Gyakran használt a konkordancia index, illetve az (1;1) „ideális” ponttól vett legkisebb eltérés módszere is.

Három és magasabb dimenziós esetekben hasonlóan az AUC-hez definiálható a ROC felszín alatti térfogat (VUS). Három csoport esetén a VUS alkalmas a teszt diagnosztikai hatékonyságának mérésére:  $VUS=1/6$  a teszt diagnosztikai alkalmatlanságát,  $VUS=1$  a tökéletes szétválasztást jelenti. Megfelelő diagnosztikus hatékonyság esetén az elvágópont kereső eljárások általánosíthatók három dimenzió esetére.

Több tanulmány is vizsgálja különböző eloszlások, elemszámok esetén a három elvágópont kereső eljárás által meghatározott elvágópontok egymáshoz való viszonyát kettő dimenzióban. Előadásunkban normál, illetve gamma eloszlású adatok esetén összehasonlítjuk a három módszer által becsült elvágópontokat. A generált adatok alapján becsült (cest) és az empirikus (copt) elvágópontok összevetéséhez minden modell esetén kiszámoltuk a relative bias és az átlagos négyzetes eltérés (MSE) értékét.

Eredményeink alapján elmondható, hogy a Youden-index által becsült elvágópontok esetén a „legkisebb” és „legnagyobb” csoportban magas, a középső csoport esetén alacsony a helyes diagnózisok aránya. A konkordancia index és az ideális ponttól vett legkisebb távolság módszerek a középső csoport esetén ad magasabb helyes-diagnózis arányt.

Előadásunkban ismertetjük a több-csoportos ROC analízis és az elvágópont kereső eljárások alkalmazását colorectalis carcinoma esetén. Vizsgáljuk a calprotectin és MMP-9 (mátrix-metalloprotáz-9) szintek relevanciáját a colorectalis carcinoma, a premalignus állapot (adenoma) és az „egészséges” betegek elkülönítése esetén, majd az általánosított Youden-index alapján elvágó pontokat határoozunk meg.

## Túléléselemzés és asszociációs szabályok: Egy integrált megközelítés az egyetemi hallgatók lemorzsolódásának vizsgálatában

Csalódi Róbert<sup>1</sup>, Abonyi János<sup>2</sup>

<sup>1-2</sup>Pannon Egyetem, Mérnöki Kar, Folyamatmérnöki Intézeti Tanszék, <sup>1-2</sup>ELKH-PE Komplex Rendszerek Figyelemmel Kísérése Kutatócsoport, Veszprém, Magyarország,  
<sup>1</sup>csalodi.robert@mk.uni-pannon.hu, <sup>2</sup>janos@abonyilab.com

*Kulcsszavak: hallgatói lemorzsolódás, túléléselemzés, gyakori elemhalmaz keresés, asszociációs szabályok*

A hallgatói létszámcsökkenés Magyarországon komoly probléma, amire kiemelt figyelmet kell fordítani. A probléma egyik megoldása, ha komoly figyelmet fordítunk a meglévő egyetemi hallgatók életútjára és olyan eszközöket fejlesztünk, amelyek képesek előre jelezni a lemorzsolódást. Erre a célra egy adatalapú módszertant mutatunk be, amely integrálja a túléléselemzést a gyakori elemhalmazokból generált asszociációs szabályokkal. A túléléselemzés célja olyan valószínűségi eloszlásfüggvények identifikálása, amelyek megmutatják annak a valószínűségét, hogy a hallgató több időt tölt az egyetemen, mint egy bizonyos félév. Mivel a hallgató távozása történhet sikeres diplomaszerzés, illetve lemorzsolódás következtében is, ezért kiemelten fontos ezeket a kimeneteket versengő kockázatokként kezelni. A módszer a hallgatók mintatantervhez képest történő elmaradásukat eseményekként reprezentálja és ebben az eseménytérben gyakorielemhalmaz kereséssel tipikus mintázatokat keres. A kidolgozott módszerrel azonosíthatóak azok a tárgyak, amelyet a hallgatók sikertelenül teljesítenek. Asszociációs szabályok segítségével determinálható a becslendő versengő kockázat, így azok a tárgyak, amelyek nagy valószínűséggel a hallgató lemorzsolódásához vezet. A gyakori elemhalmazok támogatottsága alapján meghatározott valószínűségek és konfidenciák alapján közvetlenül becsülhető a túlélési függvény a különböző bukási mintázatokat produkáló hallgatók körében. A modell tovább bővíthető középiskolai és demográfiai információkkal, amelyek igazoltan szignifikáns különbségeket mutatnak a lemorzsolódás kockázatára vonatkozóan. A rendszer bevezetése a Pannon Egyetemen egy integrált tanulmányi menedzsment rendszer részeként folyamatban van.



## PLS-SEM módszer alkalmazása a digitális térben

Hargitai Dávid Máté

Pannon Egyetem, Üzleti Tudományok Intézete, Gazdaságtudományi Kar, Marketing Tanszék,  
Veszprém, Magyarország, hargitai.david@gtk.uni-pannon.hu

*Kulcsszavak: PLS-SEM, médiafogyasztás, vásárlási magatartás, fandomizáció*

A Hallyu, mint kulturális jelenség nemcsak a koreai kultúra iránti nyitottságot befolyásolja, hanem a rajongók vásárlói magatartását is. A magyar társadalomra is hatással van a K-pop és a K-dráma, ezért a digitális tér K-pophoz kapcsolódó jellemzőinek jelentőségét vizsgáltuk ebben a tanulmányban. A PLS-SEM módszert ma már széles körben alkalmazzák a menedzsment különböző területein, mint például az emberi erőforrás menedzsment, marketing- és stratégiai menedzsment, de a szűken vett kutatási témában is több szerző használta már ezt a statisztikai módszert. A kutatás elsődlegesen a fandom, médiaaktivitás és vásárlási szándék összefüggéseit kívánja bemutatni. A kutatás primer adatok segítségével történt, mely 2022 decemberében online kérdőív segítségével került lekérdezésre, melyre 495 főtől érkezett teljes kitöltés. A elemzés során csak ezek a válaszadók kerültek be a mintába, továbbá kritérium volt, hogy válaszaik szórása legalább 1 legyen (az egyes konstrukciók likert-skálán lettek mérve), és kitöltési idejük meghaladja az 5 percet. A modell két komponensből áll. Az első eleme a külső modell, amely az indikátorok és látens változók közötti összefüggéseket méri. A második komponense a strukturális vagy belső modell, amely a látens változók hatását vizsgálja a regressziós utak segítségével. A PLS-SEM módszert jelen kutatásban azért találtuk adekvátnak, mert strukturális modellünk összetett, sok konstrukciót és modellkapcsolatot jelenít meg, melyek között több formatívan mért konstrukció is van, melyet a CB-SEM kevésbé tud kezelni. Az eredmények azt mutatták, hogy a fandom faktora szignifikáns hatást mutat a média aktivitásra. A legerősebb kapcsolat a lojalitással hozható összefüggésbe ( $\beta=0,316$ ), de a vonzerővel is közepesen gyenge reláció mutatható ki ( $\beta=0,236$ ). Ezzel szemben a fandom involvement komponensei közül csak a lojalitás mutat szignifikáns, de gyenge kapcsolatot a vásárlási szándékkal ( $\beta=0,156$ ). A kutatás hozzájárul a Hallyu-val kapcsolatos tanulmányok irodalmához. Megállapítható, hogy a fandom bevonódás szintje pozitív hatással van a vásárlási szándéokra. Ennek oka, hogy a rajongók szinte minden digitális térben ki vannak téve a K-pop termékeknek, és a fandom elvárja tőlük, hogy vásároljanak. A fandom bevonódás és a média aktivitás pozitív hatására vonatkozó megállapításokat Suvittawat (2022) is alátámasztja, mivel a média posztjain és kommentjein keresztül a rajongók nemcsak tájékoztatják egymást, hanem ajánlják is a bálványokkal kapcsolatos termékeket.

## Hálózatalapú modell- és adatredukciós módszer

Kosztyán Zsolt Tibor<sup>1</sup>, Kiss Dénes<sup>2</sup>, Fehérvölgyi Beáta<sup>3</sup>

<sup>1-3</sup>Pannon Egyetem, Gazdaságtudományi Kar, Menedzsment Intézet, <sup>1-2</sup>Kvantitatív Módszerek Intézeti Tanszék, <sup>3</sup>Innovációmenedzsment Intézeti Tanszék, Veszprém, Magyarország,  
<sup>1</sup>kosztyan.zsolt@gtk.uni-pannon.hu, <sup>2</sup>kiss.denes@gtk.uni-pannon.hu,  
<sup>3</sup>fehervolgyi.beata@gtk.uni-pannon.hu

*Kulcsszavak: modellredukció, adatredukció, hálózatelemzés, nemparaméteres módszerek*

A modell- és adatredukciós módszerek az egyik leggyakrabban használt redukciós eljárások mind a társadalom, mind a természettudományi kutatásokban. E módszerek több mint egy évszázados múltra tekinthetnek vissza. Ugyanakkor a hálózatelemzés új távlatokat nyit az adatelemzés területén is. Az adatpontokat csomópontokként és a közöttük lévő kapcsolatokat élekként ábrázolva „adathálózatot” kapunk, mellyel megnyílik a lehetőség az exponenciálisan fejlődő hálózatos elemzés eszköztárának alkalmazására is. Egy hálózaton számolt modulok meghatározása egy modulkeresési probléma optimalizációjaként fogható fel, mely eredményül meghatározott számú, változó-, vagy megfigyeléscsoportokat kapunk. Szemben tehát a modellredukciós és a legtöbb klaszterezési eljárással a klaszterek, látens változók számát, vagy egy hasonlósági küszöbindexet, nem a módszer alkalmazása előtt kell megadnunk, hanem a csoportok száma és a csoportokban szereplő tagok az elemzés eredményeként adódnak. Tanulmányunkban egy új hálózatalapú modell- és adatredukciós módszert javasunk, mely egy olyan nemparaméteres eljárás, amely modellredukció esetén megadja a látens változók, adatredukció esetén a klasztercentrumok számát. A kialakított módszer robosztus, mert képes kevés megfigyelés alapján is meghatározni a változócsoportokat, valamint kevés változó alapján az adatcsoportokat. A javasolt módszer alkalmazható szimmetrikus és aszimmetrikus változó- és adat-távolságmértékek esetén is. A módszert szimulált és valós adatokon is teszteljük. Eredményeink azt mutatják, hogy a szimulált adatokon legtöbbször a mi módszerünk adja meg helyesen a látens változók számát, és sorolja be azokat helyesen. Módszerünket számos valós adatbázison is teszteltük. Vizsgáltunk keresztmetszeti és idősoros adatokat is. Előadásunkban két példán keresztül mutatjuk be módszerünk alkalmazhatóságát. Az egyikben az egyetemek kollaborációs és publikációs teljesítményét mérő indikátorait csoportosítottuk és minősítettük. A másik példánkban a kereskedelmi hálózatok hálózati paramétereinek időbeni mintázatait csoportosítottuk. Az elkészült módszer R és MATLAB programnyelvben validált csomagként is elérhető.

## Gépi tanulási modellek értelmezhetősége

Ipkovich Ádám, Abonyi János

ELKH-PE Komplex Rendszerek Figyelemmel Kísérése Kutatócsoport, Folyamatintézeti  
Tanszék, Mérnöki Kar, Pannon Egyetem, Veszprém, Magyarország,  
ipkovichadam@gmail.com

*Kulcsszavak: Feketedoboz modellek, hierarchikus modellek, modellek magyarázhatósága, Shapley érték.*

A komplex gépi tanulási modellek bemenetei és a kimenetei ugyan ismertek, de a belső mechanizmusok nagyrészt homályban maradnak. Ez az átláthatatlanság akadályozhatja a bizalmat, különösen az olyan nagy tétet jelentő alkalmazásokban, mint az orvosi diagnosztika vagy az autonóm vezetés. Ráadásul, a rejtett torzítások akár hibás döntésekhez is vezethetnek, és a modell működésének átláthatósága nélkül ezeket a torzulásokat szinte lehetetlen korrigálni. Ezen adat-alapú fekete doboz modellekhez köthető problémák megoldása érdekében előadásunk a magyarázható mesterséges intelligencia (explainable artificial intelligence, XAI) egyre növekvő területébe ad bepillantást. Megvilágítjuk az XAI-ban alkalmazott technikákat, beleértve a Feature Importance-t, a Partial Dependence Plots-ot és a modell-agnosztikus módszereket, mint például a LIME (Local Interpretable Model-Agnostic Explanations) és a SHAP (SHapley Additive exPlanations). Bemutatjuk a módszerek elméleti alapjait, kiemelve, hogyan járulnak hozzá az összetett modellek belső működésének feltárásához. Különös figyelmet fordítunk arra, hogy a Shapley értékekkel miként vizsgálhatjuk, hogy adott változó mekkora mértékben járul hozzá a modell kimenetéhez, és így miként határozhatjuk meg a bementetek fontosságát. Egy olyan új megközelítést vezetünk be, amely magában foglalja a Shapley-érték felhasználását az egyes változók kimenethez való közvetlen és közvetett hozzájárulásának meghatározásához.

A módszerek alkalmazhatóságát a Global Green Growth Institute által kifejlesztett magyar vízügyi modell elemzésén keresztül mutatjuk be, illusztrálva, hogy az XAI módszerek alkalmazásával miként lehetséges szakpolitikai beavatkozási pontok azonosítása és döntéstámogató modellek validálása.

## Sok a szöveg?! Olvass inkább a sorok közt!

Molnár Anna Enikő, Tamási-Mészáros Evelin

Statistical Products Hungary Kft. (Clementine), Budapest, Magyarország,  
amolnar@clementine.hu, emesaros@clementine.hu

*Kulcsszavak: szóbeágyazási modell, word embeddings, többnyelvű szövegelemzés, Python, duplikációkezelés, hálózat, entitáskinyerés*

Az egyre szélesebb körben hozzáférhető szöveges adatok térhódításának köszönhetően egyre többször szembesülhet egy elemző azzal a kihívással, hogy az egyezőknek számító tartalom figyelmen kívül hagyása torzíthatja a statisztikát. A statisztikai adatszolgáltatás egy univerzális elvként értelmezhető a különböző tudományterületeken, amely a nemzetközi és nemzeti jogi elvekben is tükröződik. Az adatok elemezhetősége és elérhetősége egy olyan elvárás, amelyet egy az adott tudományterület iránt érdeklődő ember támaszt a publikált statisztikákkal szemben. Azonban nagyon nehéz eligazodni az információk tengerében. Az EUROSTAT 2022 decemberében indult pályázatának az egyik célja is az volt, hogy felhívja a figyelmet az adatok sokszínűségére, ezzel lehetőséget biztosítva a lelkes elemzők számára, hogy szöveganalitikai képességeiket összemérjék.

A felhívás keretén belül az interneten fellelhető álláshirdetések között található egymást átfedő hirdetéseket kellett feltárni. A verseny célja összefügg azzal a törekvéssel is - amely az Európai Unió jogrendjében is megtalálható -, miszerint „Az (EU) 2016/589 rendelet 17. cikke elrendeli egy egységes rendszer létrehozását a tagállamokból származó állásajánlatok, állaspályázatok és önéletrajzok EURES-portálon történő összegyűjtése érdekében.”<sup>1</sup> Az azonos tartalmat tükröző álláshirdetések torzítyják a statisztikát, hiszen ha ugyanazon pozíció több helyen is publikálásra kerül, akkor felülreprezentálnak tűnhet az adatbázisban.

A szöveganalitikai kihívás első nehézségét az jelentette, hogy a vizsgálandó szövegek között 32 különböző nyelvet tudunk detektálni, melyekből az alábbi nyelvek szerepeltek leggyakrabban: angol, német, francia, holland, spanyol, olasz, portugál, svéd, észt, lengyel, magyar. A módszertan kialakításakor ezt az aspektust is figyelembe kellett venni, hiszen a különböző nyelvű szövegek előkészítéséhez eltérő eszközök szükségesek. A nyelvek sajátosságai meghatározzák a felhasználható elemzési eszköztárat, hiszen a korpusz előállításához szükséges lépések - mint a tokenizálás vagy a ragok eltávolítása - különböző nyelveken eltérő módon valósíthatók meg.

A kivitelezés Python nyílt forráskódú programnyelven zajlott szóbeágyazási modellek alkalmazásával, de különféle technikai és módszertani kihívás is felmerült. A többnyelvű szövegek ténye az elérhető szakirodalmakban való további tájékozódást indikált. Jelenleg kevés olyan szövegek közötti hasonlóságot mérő modell létezik, amely áthidalja ezt a problémát, jelen esetben a sentence-transformers csomagra esett a választás, mivel korábbi tanulmányok alapján jól teljesít többnyelvű adatokon is. A csomagon belül különféle előre tanított modellek állnak rendelkezésre, amelyeket saját adattal tovább lehet fejleszteni. A leírás szerint közel 50 nyelvű adatot használtak fel a tanításhoz, amelyek között a magyar is szerepel, így esett a választás a *paraphrase-multilingual-MiniLM-L12-v2* nevű alapmodellre a többnyelvű szövegösszehasonlításhoz, az angol nyelvűek esetén pedig az *all-MiniLM-L6-v2* nevű modellre.

<sup>1</sup> <https://eur-lex.europa.eu/legal-content/HU/TXT/?uri=CELEX%3A32017D1257&qid=1682459328615>

A tapasztalat azt mutatta, hogy a modellek segítségével különböző nyelvű szövegek között kevésbé azonosítható a hasonló tartalom, mint az azonos nyelvűek között, ezért az előkészítés első lépését a nyelvi egységesítés jelentette, miszerint a szövegek angol nyelvre kerültek lefordításra, majd ez képezte a további elemzés alapját. A feladat magában foglalt egy matematikai problémát is, miszerint minden lehetséges hirdetéspárt szükséges lett volna megvizsgálni, de ez a több mint 112000 álláshirdetésből álló adatbázis esetén a rendelkezésre álló hardverekkel nem volt kivitelezhető, ezért egy iteratív folyamat segítségével fokozatosan kerültek meghatározásra a lehetséges párok, amelyek után egy manuálisan meghatározott szabályrendszer segítségével az előre meghatározott kategóriákba kerültek besorolásra. Ezzel a módszerrel 575190 potenciális pár került feltárássra, amelyek részben, szemantikailag vagy akár teljesen egymás variánsai voltak.

A verseny beküldési időszaka 2023. március 31-ével véget ért ugyan, de a lelkesedés nem csökkent.

A koronavírus járvány 2020-as megjelenését követően a sajtó nagy érdeklődéssel követte az eseményeket, és az egész médiát áthatotta a téma. A téma nagysága felvetette a probléma aktualitását, hiszen a média különböző forrásai eltérő stílusban és formában teszik közzé a nyilvános híreket. A korábban részletezett módszertan alkalmasnak tűnt arra, hogy az álláshirdetések duplikációmentesítéséhez hasonlóan definiálásra kerüljenek azonos információtartalmak mint például a napi koronavírus statisztika, oltások körüli hírek, közéleti szereplők a járvány szemszögéből, vagy a koronavírussal kapcsolatos intézkedések. A kialakított módszertan finomítása volt szükséges, hiszen a szöveges adatok nyelve magyar volt. Az évek során gyűjtött hírekből álló adatbázis több mint 114000 hírt tartalmaz, ahol a vizsgálandó adatok tartalmazzák a hír eredeti szövegét, a megjelenés dátumát és a hír forrását. A tapasztalat azt mutatta, hogy a több nyelvre alkalmazott módszertanhoz képest az egynyelvű szövegek kissé eltérő megközelítést igényelnek.

Elméleti szempontból megközelítve a vizsgált adatok angol nyelvre való lefordítása javított a hasonló cikkek azonosításában, hiszen angol nyelvre illesztett előre tanított szóbeágyazási modellek számos helyen elérhetőek, és ezek a rendelkezésre álló adatokkal tovább finomíthatóak. Az álláshirdetések szövegei esetén relevánsnak tűnt a szakmák beazonosítása a hasonlóság típusának meghatározásánál, és ez a szótár alapú megközelítés és egy hasonlósági arányszám együttes kezelésével hatékonyan is bizonyult. A cikkek hasonlóságának feltárása összekapcsolódik egyben egy témameghatározással is, amelyben nagy szerepet kaphat az entitáskinyerés a szakmák definiálásával analóg módon. A nyílt forráskódú entitáskinyerés a Python alapú Spacy csomag segítségével került kivitelezésre, amelyből a későbbiekben a hasonlósági arányszám kiegészítésére szolgáló szótáralapú megoldás megszületett. A gyakorlati alkalmazás azt mutatta, hogy a Spacy magyar nyelven ugyan használható, viszont nagyon érzékeny olyan apró részletekre is mint a kis- és nagybetűk különbsége, egybe- különírás vagy akár a betűszintű elírások. Az entitáskinyerése azonban segítséget nyújtott egy komplett szótár kialakításában, amelyben a szakértők által beazonosított tévesztések korrigálhatóak, de az eredményként kapott helyek, személyek, dátumok és különféle közösségi médiában használatos hivatkozások (@-ok, #-ek) hozzájárultak a hasonlóság pontosabb meghatározásához, még ha azokat fenntartással is kellett kezelni.

A megvalósítás technikai hátterét tekintve az elemzés hardverigénye nem indokolta GPU használatát, elegendő volt CPU segítségével végezni a számításokat. Értelemszerűen nagyobb adatmennyiség esetén szükség lehet a futtatás optimalizálására, amely megnyilvánulhat akár a különféle elemzési lépések párhuzamosításában vagy akár a rendelkezésre álló hardverek bővítésében.

Összességében elmondható, hogy a módszertan finomítása után a tartalmi végeredmény mutatta a várt hipotézist, miszerint a rendelkezésre álló szóbeágyazási modellek saját adatokkal tovább tanított változatából kinyert hasonlósági arányszámok, a megjelenés dátuma és a közös

entitások segítségével feltárhatóak voltak lényegi tartalmi összefüggések a szövegek között. A cikkek nagy hálózata a különféle beazonosított tartalmak mentén olyan alhálózatokra bontható, ahol a szövegek szorosabb kapcsolatban állnak egymással, mint a hálózat többi tagjával. Ezáltal egy hasonló módszertan kivitelezése támpontokat nyújthat egy klaszterezési feladathoz is, de mindig fontos szem előtt tartani az elemzendő adatok sajátosságait.

## i2 hálózatelemző eszközök áttekintése és az i2 Academic program

Pancza Judit

Statistical Products Hungary Kft., Budapest, Magyarország, [jpancza@clementine.hu](mailto:jpancza@clementine.hu)

*Kulcsszavak: i2, hálózat vizualizáció, hálózat elemzés*

A hálózatelemző- és vizualizációs eszközök számos iparágban segítik hatékonyan az elemzők munkáját, legyen az a bűnüldözési vagy gazdasági terület.

A gazdasági szektorban a pénzügyi csalások elleni küzdelem egyre nagyobb kihívást jelent, becslések szerint a csalások mértéke elérheti akár az éves árbevétel 5-8 százalékát is. Az elkövetők egyre ügyesebben használják ki a különböző vállalati rendszerek gyengeségeit, sőt akár még a munkavállalókkal is összejártsanak. Mindez fenyegetést jelent a vállalati hírnévre, és közvetve a fogyasztókra is hatással van.

A bűnüldöző szervezetek esetében a nyomozati munka sikeréhez elengedhetetlen a bűnözői hálózatok feltérképezése, a kulcsfontosságú tagok azonosítása, a hálózatban betöltött szerepük és viselkedésük megértése.

Az i2 eszközök a különböző forrásokból származó adatokból egy egységes képet képesek kialakítani, így a gyanús tevékenységek teljeskörűen feltérképezhetők és olyan esetek is feltárhatók, amihez hasonlóan korábban még nem tapasztalt az adott szervezet. Ennek alapja, hogy a csalások és bűncselekmények a normálistól eltérő viselkedési mintázatok. Ha a normális, megszokott viselkedési mintázatok azonosításra kerülnek, akkor ezáltal rögtön láthatóvá válnak azok az események, amelyek rendellenes viselkedésre utalnak.

Az i2 termékcsalád számos eszközzel segíti az elemzők munkáját, melyek közül a két legfontosabb: az i2 Analyst's Notebook a hálózatok vizualizációját és elemzését teszi lehetővé, az i2 iBase a hálózatot felépítő entitások és kapcsolatok hatékony tárolásának és elérésének módját adja.

Az i2 Academic Program olyan oktatási intézmények számára indult el idén, amelyek egyetemi szintű képzést kínálnak például a kriminológia, a törvényszéki elemzés területén vagy egyéb i2-hoz kapcsolódó gazdasági, pénzügyi területeken. A program lehetővé teszi, hogy az oktatási intézmények ingyenesen használják az i2 Analyst's Notebook szoftvert oktatási célra, hogy a hallgatók már pályájuk elején megismerkedjenek az i2 termékcsaláddal.

## A hálózatos elemzések adat schema tervezése iBase alkalmazásban

Csatlós Béla

Statistical Products Hungary Kft., Budapest, Magyarország, bcsatlós@clementine.hu

*Kulcsszavak: iBase, adat schema, tervezés, elemzés, hálózatok, adatbázis*

A relációs adatbázisokban tárolt adatok feldolgozására számos ismert módszer, eszköz áll rendelkezésre. Az adatbázisok gyakran alkalmazásokat szolgálnak ki, adatszerkezetük, terhelhetőségük nem elemzési célokat szolgál, ezért az elemzések támogatására adattárházakat, adatpiacokat hoznak létre. Az adatok adatbázisból történő lekérdezésének módja az SQL lekérdezés, mely számos lehetséges kimenetet képes létrehozni, azonban hálózatos vizuális megjelenítést lekérdezéssel nem kapunk eredményként. Bizonyos felhasználási esetek azonban megkövetelik az adatok hálózatos elemzését.

A hálózatban az adatok ábrázolási tevékenysége túlmutat a gráfelméleti struktúra elemeihez történő egy vagy többdimenziós adat hozzárendelésen. Az elemzési lehetőségek elméleti határát a forrás adatbázisban szereplő adatok determinálják, az elemzési lehetőséget bővítő és egyben korlátozó tényező a szoftver, illetve a szoftverben alkalmazható módszertan és a benne kialakított adatstruktúra. A szűk keresztmetszetek, elemzési lehetőségek, az optimális struktúra, struktúrák kialakításának igénye, rámutatnak arra a szükségszerűségekre, hogy az elemzési kérdés, illetve egy témában felvethető több elemzési kérdésre választ adni képes struktúra megtervezése minden esetben az első lépés a hatékony adatbázisokon alapuló hálózati elemzések végrehajtása felé.

Az iBase hálózati adatbázis kezelő szoftver, adatbázis struktúra tervező felülete az iBase Designer. Az alkalmazás az ELP (Entities Links Properties) adatmodellezési koncepciót használja. Az entitások és linkek szabadon definiálhatóak éppúgy, mint azok tulajdonságai. A koncepciót kiegészítve a legjobb nemzetközi gyakorlatokkal - mint például az általános céllal is használható, szoftverbe beépített információ kiértékelési szabálykészlet, mely az EU 2016/794 rendelet szerinti információ forrás megbízhatóság és információ pontosság értékelést tesz lehetővé - szervezetre, együttműködő szervezetekre vonatkozó, csoportmunkát és egységes elemzési koncepciókat támogató módszertan hozható létre.

Az általánosan használt entitások (személy, cím, esemény/ügy) domain független tervezési lehetőségein keresztül, kiértékelésre kerül számos konkrét adatstruktúra, azok elemzési lehetőségével, korlátaival, és a korlátok feloldási lehetőségeivel.