Mesterséges intelligenciák a plágiumkeresésben

Önellenőrzési és tanácsadási módszerek az egyetemi hallgatói munkák értékelésében

jun. Pitlik László (ELTE TTK, SZIE My-X Kutatócsoport), sen. Pitlik László (MY-X kutatócsoport)

# Vezetői összefoglaló

A hasonlóságelemzésre építő plágiumkeresés nyelv-független és univerzális (context-free) módszer, mely széleskörű nemzetközi felhasználásra alkalmas lehet. A hasonlóságelemzés, mint sikeres, korábban már bizonyított módszer alapjaira egy kis erőforrás-igényű, nagymértékben automatizált modul felépítése reális elvárás a plágiumkeresés kapcsán. A hasonlóságelemzés alapú plágium-keresés elsődlegesen nem az azonosságkeresés kiváltására, hanem indirekt gyanúk megfogalmazására készíthető: pl. egységes stílusú-e egy adott szöveg, vagy kellően tudományos stílusú-e egy adott szöveg? A hasonlóságelemzés keretében a fordítások és az egynyelvű szövegek értelmezése nem különbözik egymástól, amennyiben a fordítást a nyers, csak a kulcsszavakat (és szinonimákat is) jelentő fordításként értelmezzük. Ennek kapcsán felvetődik, és további kutatásokat igényel, vajon forráskódok logikai hasonlóságának mérésére alkalmassá tehető-e az eljárás?

# Bevezetés

„plágium: szellemi tolvajlás, más művének közlése saját név alatt, a mű alapgondolatának vagy részleteinek felhasználása a szerzőre való hivatkozás nélkül”  
(Magyar Értelmező Szótár)

Mint azt a szótári definíció is mutatja, a plágium fogalmának teljes körüljárása talán nem is olyan egyszerű, mint elsőre feltételezhető volna. A legnyilvánvalóbb eset az, amikor bekezdések, oldalak, fejezetek, horribile dictu a teljes szöveg más szerző munkájának jelöletlen átvétele, ugyanakkor – mint azt a szótár is megemlíti – elegendő lehet csupán az eredeti alapgondolatot átemelni, s azt új köntösben saját munkaként a világ elé tárni. Ez is lehet plágium, s bár ennek a bizonyítása a szövegek közötti azonosságok keresésével esetleg ki sem mutatható, a hírbe hozott szerző (akár tudományos, akár politikai) karrierjébe is kerülhet egy vád, legyen az akár alaptalan is. Mind a bírálók, mind a szerzők számára fontos tehát, hogy olyan eszköz álljon rendelkezésre, amelynek segítségével ezen a területen is megvalósítható/megközelíthető a „tényalapú szakpolitizálás”.

Az elmúlt évek plágiumbotrányai Németországban[[1]](#footnote-1) és Magyarországon[[2]](#footnote-2) is jelentős port kavartak, a kipattanó ügyek száma pedig tovább emelkedhet, hiszen a technika, technológia, tartalom fejlődésével az interneten elérhető szolgáltatások[[3]](#footnote-3) lehetővé teszik a szövegek közötti gyors azonosságkeresést. A kérdéskör, vagyis milyen szolgáltatásokon keresztül lehet a plágiumgyanúk mennyiségét csökkenteni azonban továbbra sem rendelkezi még zárt válaszokkal. A szerző aktuálisan egyetemi (ELTE) hallgatóként, s egyúttal leendő tanárként személyesen is érintettnek érezheti magát a témában, vagyis hallgatóként úgy tűnik, szükség volna egy, a plágium jelenségét nem csak azonosságkeresésben kutató, s elsődlegesen nem jogi értelemben vett ítéleteket nyilvánító módszerre. A társszerző a SZIE GTK tehetséggondozási rendszerének megújítása keretében[[4]](#footnote-4) keresi a választ arra, vajon a MY-X munkacsoport eddigi kutatási eredményei milyen pontokon kapcsolhatók az eddigi hazai fejlesztésekhez (vö. SZTAKI KOPI[[5]](#footnote-5))? Miként lehet a hazai oktatási rendszert egységes alapokon és keretek között kiszolgálni a plágium jelensége kapcsán a megfelelő megoldásokkal?

Jelen cikk célja tehát egy olyan módszer, illetőleg a módszer minél magasabb fokú automatizálásával egy olyan eszköz/szolgáltatás megalkotásának előkészítése, amely képes felhívni a készülő szöveg írójának figyelmét már a munkafolyamat során is bizonyos, ha nem is vétségként, de semmiképpen nem előnyként ható jelenségekre. Ilyen jelenségek a közhelyesség, a semmit mondás, az indokolatlan ismétlés gyanúja (mint deklaratív módon nehezen értelmezhető szabályok); továbbá az idézetek nagy terjedelme (mint pl. deklaratív módon is szabályozható jelenség), a szöveg kulcsszavainak kiemelkedően magas előfordulása az interneten fellelhető anyagok között, illetőleg (kulcs)kifejezések nagyfokú hasonlósága más, különösképpen egyetemi, tudományos munkákkal összevetve.

A módszertani cél tehát a gyanúgenerálás, illetve a gyanú eloszlatásának automatizálhatóvá tétele (vö. szimuláció), ezáltal az emberi szubjektivitás kizárása/minimalizálása ezen kényes kérdésből, egyúttal pedig az emberi erőforrások (pl. bírálók, lektorok) optimálisabb felhasználása. Más megfogalmazásban cél: támogatni tudni gyanún felül álló (egyedi) munkák létrejöttét, illetve ráirányítani a gyanúsnak minősülő dokumentumokra a figyelmet. Azokra, amelyek esetében szükség lehet az intuitív, emberi vizsgálatra a gyanú végső eloszlatásához, avagy az esetleges plágium vádjának végső (jogi értelemben is helyt álló) bizonyításához.

A legegyszerűbb plágium-elkerülési stratégiák az ideális dokumentumok megalkotásának keretszabályaiban lelhetők fel (vö. academic writing skills). Amennyiben egy kiadó ki tudja kényszeríteni, hogy csak olyan állítás köthető a mindenkori szerzők nevéhez, melyet a szerző tételesen bizonyított, s minden más kijelentés/deklaráció esetére forrás(ok) megadása az elvárás[[6]](#footnote-6), akkor lényegében a plágium jelenségköre jelentősen szűkül. A szakirodalmi feldolgozások rövid idézetei és ezek szerzői értelmezései jelenleg is kizárnák elvileg már az idegen tollakkal való ékeskedést, de mivel már arról is vita folyik, kell-e, s mennyit idézni, vagy éppen a saját szavakkal való megfogalmazás az értékteremtő, így először intézményi, hatósági szinten kell rendet teremteni a szerzői szabadságjogok számára világos kereteket szabva, de ezeket mégsem megkurtítva.

Ennek megfelelően a célcsoport elsődlegesen a mindenkori egyetemi hallgatói-oktatói/kutatói réteg, akik kezébe a beadandó feladatoktól kezdve a diplomamunkákon át a doktori értekezésekig, egyéb tudományos közleményekig bármilyen munka esetén a lehetőség szerinti leginkább objektív, azonban az emberi intuíciók felhasználásával folyamatosan finomítható dinamikus értékelési támogatást kívánnak adni a fejlesztők. A végső cél pedig egy robot lektor megalkotása![[7]](#footnote-7)

# Helyzetértékelés

A latin *plagiarius* szó jelentése emberrabló, gyermekrabló; az I. században élt költő, Martialis[[8]](#footnote-8) volt az, aki a kifejezést eltulajdonított versei vonatkozásában is használta. A probléma végigkíséri az egész emberi történelmet, azonban míg kezdetben vagy a kevés rendelkezésre álló írott anyag miatt hamar lebukott a plagizátor, vagy a társadalmak alacsony mobilitása és relatív elszigeteltsége folytán párhuzamos felfedezésről is beszélhetünk, mára az interneten tulajdonképpen tetszőleges témában állnak rendelkezésre szakmailag releváns (egyetemi segédanyagok, jegyzetek, korábbi tudományos munkák digitalizált változatai) művek jelentős számban, melyekhez a hozzáférés is széles rétegek számára biztosított. Kellően sok forrás feldolgozásával „összeollózható” olyan munka, amely bár teljes átvett bekezdéseket esetleg sehonnan nem tartalmaz, mégsem tekinthető új, önálló alkotásnak[[9]](#footnote-9). Ez a fajta, mondhatni „stilisztikai” plágium a művészetekben új értéket is létrehozhat (vö. parafrázis), tudományos igényű munkáktól azonban inkább elvárható ennek kerülése, az újdonság (a bizonyított deklaráció) közlésére szorítkozás.

Jelenleg elérhető több plágiumkereső szolgáltatás is az interneten[[10]](#footnote-10), többek között a SZTAKI által fejlesztett (szövegdarabolással és szinonimák kezelésére alapozva működő) KOPI[[11]](#footnote-11) is. Ezen plágiumkeresők jelentős adatbázissal rendelkeznek (a KOPI „többtízezer regisztrált felhasználó által feltöltött dokumentumokban” keres, továbbá elérése van a Wikipedia, valamint számos digitális gyűjtemény felé), ezen adatbázisra támaszkodva vizsgálják a kérdéses dokumentumokat, szövegegyezéseket keresve, s minden egyes megvizsgált dokumentummal érzékenyebbé válik a keresésük. Jelentős számítási kapacitást igényel a dokumentumok tárolása, illetve a saját keresőrendszer felépítése.

A történeti áttekintés mellett a helyzetértékelés kapcsán érdemes utalni arra, hogy a Magyar Internetes Agrár/Alkalmazott Informatikai Újság[[12]](#footnote-12) keretében kialakításra került egy WIKI szolgáltatás. Ezen WIKI Példatár moduljába sok száz elvileg egyéni (innovatív, egyedi, speciális) elemzés került fel a mindenkori követelményrendszerek eredményeként az érintett hallgatók jóvoltából[[13]](#footnote-13). Noha az elvárások alaplogikája alapján a plágium esélye elvileg nagyon alacsony (lényegében nulla) illene, hogy legyen, egyes át nem látható motivációkkal rendelkező hallgatók mégis produkáltak gyanús megoldásokat. Ezek véletlenszerű felfedezését követően a manuális plágiumgyanú levezetése részévé vált a korábbi alkotás-orientált követelményrendszernek[[14]](#footnote-14). Így tehát mára rendelkezésre áll egy kézi vezérléssel és egyedi emberi erőforrások felhasználásával alapvetően jól áttekinthető teszt-környezet (mely a majdani hasonlóságelemzések során a referencia-adatbázis alapja lehet), illetve potenciális, emberi „szakértők” által levezetett gyanúmomentumok, melyek a mesterséges intelligencia megoldások benchmarkjaként is felfoghatók.

A My-X Kutatócsoport kapcsán szintén előzményként említendő a hasonlóságelemzés módszertanának kialakítása, mely probléma-megoldási szemléletmód meglehetős rugalmasságot kínál fel az emberi szakértők képességeinek robotok általi leképezésére[[15]](#footnote-15). A robot lektort tehát ebben az összefüggésrendszerben, mint reális célkitűzést kell tekinteni, mely képes lehet a jövőben: adott személy korábbi stílusjegyeinek ismeretét felhasználva egy-egy aktuális műről eldönteni, vajon valóban a szerző művéről van-e szó (eégszében, ill. bármely részletét tekintve). Emellett a tudományos közlés általános (érvelő, bizonyító, vitatkozó, leíró) stílusjegyeinek kellő mértékig való jelenléte hasonlóképpen elemzés tárgya lehet, mint ahogy a zenei szépség[[16]](#footnote-16) vagy az építészeti ideál[[17]](#footnote-17) kérdésköre már jelenleg is az.

# Kutatási problémák

Jelen munka célja a már meglévő szolgáltatások mellett egy új, szövegkorpuszként a mindenkor az interneten fellelhető, publikus tartalmat alapul vevő (hiszen a vizsgálandó szöveg szerzőjének is ez állt rendelkezésére digitális formában), a nyilvános keresőszolgáltatások erőforrásaira támaszkodó naplózó modul létrehozása, mely természetesen (erre vonatkozó felhasználó igény és a szükséges infrastruktúra együttes fennállásakor) tetszőlegesen egészíthető ki saját szövegkorpusszal is. A naplózó modul mellett szükséges a parafrázis/stílustalanság/egyediség (s még bármely más emberi fogalomként előnyös vagy káros jelenség) gyanújának levezetni tudása a naplóállományok alapján. Vagyis egy-egy olyan értelmező modul megalkotása, amely nem csak a teljes szövegegyezésekre, hanem a már korábban is említett stilisztikai, fogalmazási hasonlóságokra is érzékeny tanácsadóként funkcionál, s képes szimulációkat is végezni annak érdekében, hogy feltárja, a szerzőnek a dokumentum melyik részében mit kellene tennie azért (vö. tanácsadási funkció), hogy dokumentuma a korábban megvizsgált anyagokkal összehasonlítva egyre egyedibb legyen a robotizált gyanúgenerátor számára.

Mivel tehát – többek között a fent már jelzett hatalmas erőforrásigény miatt – nem cél jelenleg az önálló szövegkorpusz fejlesztése, vagy saját keresőmotor felállítása, a kutatás a következő irányokban indult meg:

1. a feldolgozandó szövegekből milyen módon nyerhetők olyan releváns keresési kifejezések, amelyek alapján az online keresőszolgáltatások eredményei további feldolgozásra kerülhetnek? (vö. szövegdarabolásos módszer egyre hosszabb szövegek bevonásával, ahol egy rel. hosszú szöveg lényegében visszahullik az azonosságkeresésbe)
2. a szövegbányászati statisztikák, illetve a feldolgozott keresési eredmények felhasználásával felállított COCO[[18]](#footnote-18) hasonlóságelemzési modellek eredményei milyen szöveges szakértői rendszerekkel megtámogatva válhatnak közvetlen tanácsadási, gyanúgenerálási eszközzé azon felhasználók számára is, akik magát a módszert nem kell, hogy kezelni tudják, csupán saját szövegük egyediségét/közhelyességét (gyanússágát) kívánják felmérni? (vö. több-dimenziós értékelések módszertani fejlesztése)

Ezen módszer saját szövegkorpusz helyett – mivel a COCO hasonlóságelemzés mindenkor relatív, tehát a többi, valaha feldolgozott dokumentumhoz képest értelmezett relatív gyanút fogalmaz meg – referencia-adatbázis felállítását követeli meg, ennek alapjait a kutatást végző SZIE My‑X Kutatócsoport az eddig is folyamatosan gyűjtött, és az interneten elérhetővé tett hallgatói feladatok[[19]](#footnote-19) demó-elemzésével állítja elő, ill. részben már elő is állította. A referencia-adatbázis a szolgáltatás működésbe lépésétől kezdve minden további vizsgált dokumentum mutatószámait és originalitás-indexét tárolhatja[[20]](#footnote-20), így továbbra is kis erőforrásigény mellett egyre pontosabb elemzések előállítását teszi lehetővé. Ez a működési mód egy fajta statikus szakértői rendszerként értelmezhető, míg a minden egyes elemzett dokumentumot azonnal feldolgozó, dinamikus gyanúmomentumokat generáló megoldás során a mindenkori utolsóként vizsgált dokumentum hatással lehet a gyanúmomentumok mértékére, a modellek hitelességére, az értelmezések kockázataira.

# Technológiai megvalósítás variációi

A vizsgálandó objektumot a hasonlóságelemzésen alapuló plágiumkereső többszintű ellenőrzésnek veti alá. Mivel a plágium gyanúja per definitionem kizárt az idézett szövegrészekben, elsődlegesen tehát a dokumentumban a hivatkozások mennyiségének (idézőjelek között lévő karakterek száma osztva a teljes dokumentum terjedelmével) felmérése történik. (S bár a hasonlóságelemzés eredendően a nem-deklaratív elveken alapuló gyanúgenerálás eszköze, itt – például akkor, ha ezt intézményi szabványok előírják – közvetlenül megállítható a folyamat és munka értékelése abban az esetben, ha a dokumentum a megengedettnél nagyobb arányban tartalmaz hivatkozásokat, s ilyen formán a plágiumkeresés eredményétől függetlenül elfogadhatatlan lenne.) Az idézetek kezelése kapcsán felmerül ezek felismerhetősége: természetesen sokféle jelölés képzelhető el az idézetek, az Olvasó számára félreérthetetlen kiemelésére (pl. a forrásdokumentumok színvilágában invertált képei), de a társadalmi hasznosság kapcsán talán nem irreális elvárás az SI mértékegység-harmonizáció mintájára az idézetet nemzetközi szabványként idézőjelben megadatni a jövőben minden szerző által.[[21]](#footnote-21)

Klasszikus szövegbányászati módszerekkel meghatározható a dokumentum leggyakoribb szavainak[[22]](#footnote-22), kifejezéseinek[[23]](#footnote-23) listája. Amennyiben ezen szolgáltatások a magyar nyelvet közvetlenül nem is támogatják, a módszer szempontjából még az sem jelent problémát, ugyanis az internetes szövegkorpusszal történő összevetés során minden vizsgált dokumentumot ugyanolyan mértékben érint ez a jelenség.

Lehetőség van a dokumentumok véletlenszerű megmintázására is, itt a szövegbeli gyakoriságot figyelmen kívül hagyva vizsgálható többek között az, hogy milyen hosszú karakterláncok, szavak, kifejezések mutatnak maximális hasonlóságot a szövegkorpusszal, tehát mennyire fogalmaz egyedien, újszerűen a szerző.

Szintén alkalmazható, a makrók segítségével pedig várhatóan nem túl nehezen automatizálható mintázási eljárás lehet végül a MS Word programban meglévő beépített kivonatoló[[24]](#footnote-24). Az alkalmazás kapcsán felmerülő probléma lehet a funkció támogatásának megszűnése a legutóbbi programverzióban[[25]](#footnote-25), azonban a tesztelési periódusban a jelenleg elérhető gépeken még meglévő modul is alkalmazásra kerül. Az internetes szövegbányászati eljárásokhoz hasonlóan a kivonatoló sem támogatja elsődlegesen a magyar nyelvet, ez azonban nem teszi alkalmatlanná a szógyakoriságok vizsgálatával (vö. szövegbányászat) a legkarakteresebb mondatok kiemelésére, továbbá – mint az már kifejtésre került – a mindenkori vizsgált dokumentumokat azonos mértékben érinti a nyelvi támogatás hiánya.

## Mutatószámképzés, OAM-építés[[26]](#footnote-26)

A COCO hasonlóságelemzés megfelelően strukturált adatvagyon esetén futtatható. Ez a speciális struktúrájú tábla az objektum–attribútum mátrix (továbbiakban: OAM), mely a hasonlóságelemzési modulba közvetlenül betölthető. Sorfejlécen szerepelnek a referencia-adatbázis építéséhez felhasználandó publikációk, mint objektumok, az oszlopfejlécen pedig a függő és független változók (attribútumok, mutatószámok).

Első lépésben a mutatószámok meghatározása szükséges, amely mindenkor emberi döntések következménye, ugyanakkor az általuk meghatározott elemzési térben minden versengő objektum azonos esélyekkel mérettetik meg. A mutatószámok adják a know how egyik alapvető rétegét, ezért itt csak mintaértékű példák felvillantására van mód:

* a szövegbányászatilag meghatározott pl. 10 leggyakoribb kulcsszó együttes előfordulását mutató Google (Bing, Yahoo) találatok száma;
* a szövegbányászatilag meghatározott pl. 10 leggyakoribb kulcsszó kötelezően együttes előfordulását mutató Google (Bing, Yahoo) találatok száma;
* a szövegbányászatilag meghatározott pl. 10 leggyakoribb kulcsszó együttes előfordulását mutató Google (Bing, Yahoo) találatok száma magyar névtérben (site:.hu);
* a szövegbányászatilag meghatározott pl. 10 leggyakoribb kulcsszó együttes előfordulását mutató Google (Bing, Yahoo) találatok száma az „egyetem” kulcsszóval kötelezően együttes előfordulásban;
* a szövegből véletlenszerű mintázással választott pl. 20%-nyi, egyenként legalább 10 karakteres szavakra kapott Google (Bing, Yahoo) találatok átlagos száma;
* a szövegből véletlenszerű mintázással választott pl. 20%-nyi, egyenként legalább 15 karakteres szavakra kapott Google (Bing, Yahoo) találatok átlagos száma;
* a szövegből véletlenszerű mintázással választott pl. 20%-nyi, egyenként legalább 20 karakter hosszúságú szövegrészletekre kapott Google (Bing, Yahoo) találatok átlagos száma;
* az átlagos találati szám maximumához tartozó véletlenszerűen kiválasztott karakterláncok hosszúsága (pl. 20%-os mintázás);
* a maximális találati számhoz tartozó karakterlánc hosszúsága;
* a szövegben található hivatkozások (idézőjelek közötti karakterek) aránya.

A fenti mutatószámok tekintetében a „méretfüggőség” kérdésével mindenképpen foglalkozni kell. Nem mindegy ugyanis, hogy a vizsgálandó munka mennyire népszerű témában íródott, a benne fellelhető kulcsszavak összes és átlagos előfordulása az interneten a téma népszerűségétől egyértelműen függ. Más üzenete van annak az elemzésnek, ahol az elemző eltekint[[27]](#footnote-27) a mutatószámok méretfüggetlenítésétől, illetve ahol ez kikényszerítésre, esetleg együttes alkalmazásra kerül.

Az objektumok ezután csökkenő sorrendben rangsorolásra kerülnek minden egyes mutatószám alapján, s így a nyers inputtáblából előáll a rangsorszám-tábla. A rangsorolás módszertanilag indokolt, az esetleg eltérő nagyságrendben mozgó mutatószámokat egymással teszi összehasonlíthatóvá (vö. standardizálás).

## COCO hasonlóságelemzési modul választása

Az OAM ezzel minden független változó tekintetében készen áll a modul kiválasztása utáni közvetlen elemzésre. A mutatószámok megválasztásához és a méretfüggetlenség kérdéséhez hasonlóan egyszeri szubjektív döntés szükséges az objektív elemzés kereteinek végső beállításához.

### Y0 elemzés (anti-diszkriminációs modell)

Az előkészített adatvagyont első körben COCO Y0 elemzésnek vetjük alá. A hasonlóságelemzés során egy becslés keletkezik, amelynek lényege, hogy optimalizálás keretében minden tulajdonság (független változók, Xi) esetén minden rangsorszámhoz hozzárendelünk egy konkrét értéket (lépcsőt), amennyivel az adott tulajdonság hozzájárul a függő változó (Y) értékéhez. Y0 modell esetén a függő változó egy konkrét mutatószám helyett tulajdonképpen 0 (gyakorlati megfontolásból valójában egy, a feladat méretéhez választott kellően nagy konstans érték). Ezzel mintegy feltételezve, hogy „mindenki másképp egyforma” (vö. Mérő L.). A függő változó a plágiumkereső esetében tulajdonképpen *hibapont*: minél „előkelőbb” rangsorszámot kapott egy munka a szövegkorpusszal történő összevetések során, annál nagyobb értékű lépcsők összege adja a hibapontjait. A becsült hibapontok összege (additív modell esetén) és a valós Y‑értékek közötti különbségek (Delta) négyzetösszege (négyzetes hiba) a hasonlóságelemzési célfüggvény, amelynek (egy vagy több) minimuma (ideális esetben zérushelye[[28]](#footnote-28)) adja a végeredményt. A konkrét Y0 modell esetében az eredmény a következőképpen értelmezendő:

* a konstans feletti hibapont: az objektum *gyanús,*
* a konstans körüli hibapont: az objektum *semleges,*
* a konstans alatti hibapont: az objektum *nem gyanús,*

a többi objektummal összehasonlítva. A három kategória pontos határvonalai a vizsgált objektumok számától, illetve a maximális és minimális becsült Y‑érték közötti különbségtől függően jelölhetők ki, a felső kvartilisba tartozó értékek azonban már általában megalapozott gyanúra adhatnak okot.

A gyanú mértékét a hasonlóságelemzés belső logikája (vö. modellszimmetria-vizsgálatok, illetőleg inverziók[[29]](#footnote-29)) belső minőségbiztosítás keretében képes hitelesíteni vagy elvetni. A hiteles gyanúmértéken túl tehát a robotszakértő egy negyedik szakvélemény megalkotására is képes: kimondhatja, ha egy objektum gyanússága a vizsgált kontextusban nem ítélhető meg hitelesen.

A belső minőségbiztosítás által hitelesített eredménnyel rendelkező objektumok kiadják a referencia-adatbázist, az hasonlóságelemzés során feltárt lépcsőértékek segítségével további, az elemzésben részt nem vett dokumentumok hibapontjai is egyszerű szakértői rendszer segítségével kiszámolhatók: az új dokumentum esetében is meg kell állapítani az elemzés során alkalmazott mutatószámok értékét, majd ezen értékeket a már rendelkezésre álló rangsorban elhelyezve a megfelelő lépcsőértékeket összegezve gyors eredmény adható.

Mivel az Y0 elemzésben alkalmazott konstans minden referencia-adatbázis felállítására irányuló futtatás esetében más-más érték lehet, a szerzők e helyen kívánnak élni egy javaslattal a *standard originalitás-index[[30]](#footnote-30)* (továbbiakban: SOI) használatának bevezetését illetően, amely mutató a vizsgált dokumentum hibapontjainak és a referencia-adatbázis hasonlóságelemzési konstansának hányadosaként áll elő.

### Exploratív modell

Lehetőség van egy másik megközelítésből is megközelíteni a dokumentumok originalitásának fogalmát, ez esetben a már feltárt mutatószámokon kívül új, függő változó bevezetésére nincs szükség. Az egymás után sorban futtatott elemzések mindegyike egy már meglévő mutatószám értékét teszi meg Y-nak, a fennmaradó mutatószámokkal pedig megkísérli annak magyarázatát irányítás nélkül (vö. COCO MCM). Amennyiben az elemzés kezdetén „n” darab mutatószám került bevonásra, úgy az exploratív modellezés során mindösszesen „2n” futtatásra van szükség a megfelelő inverz alakokat is figyelembe véve, amelyek a már korábban említett belső minőségbiztosítást garantálják.

Exploratív modellezés esetén azon objektumok tekinthetők első értelmezés szerint originálisnak, amelyek esetén a futtatás „n” rétegében a különböző mutatószámok a legkevésbé magyarázzák egymást.

## Értelmező szövegpanelek

A szerzők által javasolt standard originalitás-index automatikus számítása és a mindenkori felhasználó elé tárása egyszerűen automatizálható folyamat, s az így kapott érték a vizsgált dokumentum egészére általánosságban érvényes egyszerű, szemléletes és könnyen kezelhető értékelést ad[[31]](#footnote-31). Ahhoz azonban, hogy valódi, hasznos, a dokumentum készítése során tulajdonképpen valós idejű támogatást kaphasson a mindenkori szerző, szükséges még megfelelő szöveges támogatás nyújtása is az elemzési eredmények értelmezéséhez. Ezen szöveges támogatás is automatizáltan valósítandó meg (vö. sablonszöveg-generáló szakértői rendszer) a mesterséges intelligenciára építő eljárás jegyében.

Megfelelő statisztikai eljárásokkal létrehozhatók azok a szöveges (igény szerint akár vizuális) jelentések, amelyek mutatószámonként, akár a szöveg egyes részleteire külön-külön is feltárják a vizsgált dokumentum eredményeinek elhelyezkedését a referencia-adatbázisban szereplő többi szöveghez képest, s így kiemelik azokat a tulajdonságokat, fejezeteket, amelyek átírása, sűrítése a dokumentum összesített megítélését a legnagyobb mértékben befolyásolhatják kedvező irányba.

# Kockázatok

A projekt jelenleg tervezési fázisban van, folyamatosan történik a különböző mutatószámok, és azok alkalmazhatóságának vizsgálata, a korábban már jelzett méretfüggőségi kérdésben jelenleg nincs, s nem is feltétlenül kell, hogy létezzen „végleges” álláspont. Amennyiben a hasonlóságelemzési becslés során – különösen a méretfüggőek közül – valamely mutatószám irrelevánssá („zajjá”) válik, úgy a fejlesztésnek a szigorú méretfüggetlenítés irányába kell mozdulnia.

Mindeközben azonban a már korábban emberek által létrehozott plágiumgyanú-jelentések eredményeire támaszkodva folyamatos kontroll alatt tartható az elemzések eredménye, s jelenlegi best practise – tehát intuitív – döntéshozatal érzékenységével, sebességével és erőforrásigényével összevethető az új módszer aktuális állapota, ami elősegíti hatékony fejlesztést.

További kockázat, hogy a hasonlóságelemzésre alapuló plágiumgyanú feltárása sem képes a szöveg emberi agy által történő értelmezésére (vö. context-free alkalmazás). Ez más oldalról akár előnyként is felfogható: míg számos ember nem képes számára ismeretlen szakterület szövegeit érdemben kezelni, addig a hasonlóságelemzés a szöveg valódi értelme nélkül is racionális gyanúmomentumok feltárására képes.

Speciális kockázat az itt is látható alkotási folyamat plágium-szempontú értékelése: vagyis egy korrektúrázás mikor számít értékteremtőnek, s mikor nem? A korrektúra-verziók az idézés alternatívái, hiszen félreérthetetlenül jelzik, mi volt az alapmű, s mi a szerző saját üzenetrétege. A kettő (de különösen a második) azonban egymás nélkül kevesebb ér.

# Kitekintés

Amennyiben a mutatószámok véglegesítése és a demóként funkcionáló első referencia-adatbázis felállítása – várhatóan legfeljebb egy hónapon belül – megtörténik, a Kutatócsoportnak lehetősége nyílik már a 2013-as év őszi félévében a béta-tesztelésre valós hallgatói feladatok esetében folyamatos szakértői támogatás és a dokumentumokra vonatkozó intuitív értékelések felhasználása mellett. A SZIE GTK jelenleg ugyanis is olyan tehetséggondozási koncepció mellett tette le a voksát az eddigi normál tantervi alkotási kényszerek helyett (ahol az ezen szint elérésére nehezen képes hallgatók számára az alkotási vágy felébresztése, vagyis a személyiségfejlődés adott szintjének elérése komoly nehézséget jelentett még időkorlátok nélkül is), melynek keretében az innovatív és minőségbiztosított gondolkodásmód csak az erre jelentkezők számára válik lehetőséggé és egyben kötelezettséggé. Az így kialakuló szakkörök, szakkollégiumok, PhD-szemináriumok, akciócsoportok tagjai az alkotásokat bemutató dokumentációk értékelését részben a plágium-robottól kapják majd.

Az ötlet jelenleg kockázati tőke bevonásáról szóló tárgyalások alapját képezi, sikeres véglegesítés esetén megkezdődhet a széleskörű, a módszer nyelvfüggetlenségéből és a nyelvi támogatás minimális szövegigényéből következő nemzetközi alkalmazás kialakítása és bevezetése is.

Végső soron pedig az itt felvázolt eljárás – a hasonlóságelemzés, mint módszer alapvetősége, illetve az internet, mint szövegkorpusz várhatóan dinamikus fejlődése folytán – teljes context-free alkalmazás keretében az írott, vagy élőszóban elhangzó bármely közlés egyediségének, originalitásának (valós idejű) mérése is lehetővé válhat a megfelelő infrastruktúra (számítási kapacitás, hangfelismerés) mellett. Ilyen körülmények között már – ellentétben azzal, ahogyan talán első hangzásra hat – egyáltalán nem a tudományos fantasztikum része kell, hogy legyen akár parttalanná váló (pl. parlamenti) viták esetén a szónokok mikrofonjának automatikus némítása, amikor azok érdemi közlésről demagóg, közhelyes stílusra váltanak át. Ezzel a bevezetőben elérendő célként megfogalmazott „tényalapú szakpolitizálás” gyakorlata a társadalmi hasznosság szempontjából talán az elvárható legmagasabb szintekre emelkedne…

# Összefoglalás

A plágium fogalma és súlya a teljes szövegegyezéstől a parafrázisig sok rétegben megfogalmazható és értelmezhető. A teljes szövegegyezés keresése hagyományos algoritmusokat alkalmazó nagy erőforrás-igényű best practise jelenleg is ismert (pl. SZTAKI KOPI). A mesterséges intelligenciára épülő finom, többrétegű és önkalibráló; parafrázisokat is feltárni képes módszer fejlesztése alapkutatási jelentőséggel bír. A COCO hasonlóságelemzés egyszerű, mély matematikai tudás nélkül is könnyen kezelhető lehetőséget nyújt különböző objektumok (adott esetben írásművek) összevetésére, amennyiben a megfelelő mutatószámokkal képesek vagyunk leírni az egyes objektumok jellemzőit. A szükséges erőforrásigény, mivel a módszer képes a publikus internetes keresőkkel való együttműködésre, szövegkorpusznak pedig mindenkor az interneten fellelhető tartalmat tekinti, rendkívül alacsony. A dokumentumokat megfelelően leíró, a plágiumkeresésben sikeresen felhasználható mutatószámok és alkalmazásuk kereteinek kutatása jelenleg folyamatban van pl. a SZIE My-X Kutatócsoportján belül. Párhuzamos kutatás folyik a létrejövő manuális megoldások lehető legmagasabb fokú automatizálására is.

A COCO Y0 hasonlóságelemzési modell képes feltárni az összehasonlított írásművak közül azokat, amelyek a többivel összevetve relatíve plágiumgyanúsabbak, mivel összességében nagyobb egyezést mutatnak az internetes szövegkorpusszal. Az így létrehozott referencia-adatbázis elemeivel egyszerű szakértői rendszerek segítségével vethető össze bármely további dokumentum. Az objektumok egyediségét a standard originalitás‑index (SOI) segítségével jellemezhetjük, mely egy időtől és a figyelembe vett referencia-adatbázistól függő relatív érték, a dokumentum Y0 modell által számított hibapontjainak és a referencia-adatbázis felállításakor alkalmazott hasonlóságelemzési konstansnak a hányadosa.

Exploratív modellezéssel a dokumentum mutatószámainak egymásból való levezethetősége vizsgálható meg. Ez esetben egyedibb az a dolgozat, amely esetén a futtatott modellek a legkevésbé magyarázzák egymást. A COCO hasonlóságelemzés beépített minőségbiztosítási rendszerrel dolgozik (inverz modellezés), ezáltal képes kimondani, ha egy kérdésben a rendelkezésre álló adatok alapján nem nyilvánítható hitelesnek minősülő vélemény.

Mind a hasonlóságelemzés, mind a szövegbányászati eszközök nyelv-függetlenek, emellett minimális, szövegpanelekből álló strukturált nyelvi támogatás kialakítása szükséges csak az eredményes használathoz, ezért a módszer széleskörű nemzetközi, ráadásul univerzális (context-free) felhasználásra alkalmas.

A projekt kutatási fázisban van, a 2013 őszi félév során a béta-tesztelés megindulhat pl. a SZIE GTK-n. Jelenleg kockázati tőke bevonásáról szóló tárgyalások zajlanak, sikeres véglegesítés esetén megkezdődhet egy objektív, mesterséges intelligenciát alkalmazó plágiumkeresési modul nemzetközi alkalmazás kialakítása és bevezetése, mely felett folyamatos kontrollját és továbbfejlesztési lehetőségeit továbbra megőrzi az emberi intuíció.

1. <http://hvg.hu/itthon/20120111_plagium_botrany>

   <http://www.atv.hu/kulfold/20120306_plagiumbotrany_ujabb_nemet_politikust_fosztottak_meg_doktori_cimetol>

   <http://www.spiegel.de/thema/karl_theodor_zu_guttenberg/>

   <http://www.spiegel.de/thema/bernd_althusmann/> [↑](#footnote-ref-1)
2. <http://hvg.hu/itthon/20120111_Schmitt_doktori_disszertacio_plagium>

   <http://hu.wikipedia.org/wiki/Schmitt_P%C3%A1l_pl%C3%A1gium%C3%BCgye>

   <http://hu.wikipedia.org/wiki/Gyurcs%C3%A1ny_Ferenc#Elt.C5.B1nt_dolgozatok_.C3.A9s_pl.C3.A1giumv.C3.A1d> [↑](#footnote-ref-2)
3. <http://www.metazin.hu/node/788> [↑](#footnote-ref-3)
4. http://miau.gau.hu/miau/181/tehetseggondozas\_2013osz.doc [↑](#footnote-ref-4)
5. http://miau.gau.hu/oktatas/2013osz/KOPI\_Plagiumkereso\_v15\_public.pdf [↑](#footnote-ref-5)
6. vö. http://miau.gau.hu/miau/181/etdk\_2013\_v4.doc [↑](#footnote-ref-6)
7. vö. <http://miau.gau.hu/miau2009/index.php3?x=e080> [↑](#footnote-ref-7)
8. <http://en.wikipedia.org/wiki/Plagiarism#Etymology> [↑](#footnote-ref-8)
9. <http://www.plagiarism.org/plagiarism-101/what-is-plagiarism>:

   inter alia „changing words but copying the sentence structure of a source without giving credit” [↑](#footnote-ref-9)
10. <http://www.metazin.hu/node/1641>

    <http://turnitin.com/en_us/home>

    <http://safeassign.com/>

    <http://www.blackboard.com/Sites/International/EMEA/index.html> [↑](#footnote-ref-10)
11. <http://kopi.sztaki.hu/> [↑](#footnote-ref-11)
12. <http://miau.gau.hu> [↑](#footnote-ref-12)
13. <https://miau.gau.hu/mediawiki/> [↑](#footnote-ref-13)
14. <http://miau.gau.hu/oktatas/2013tavasz/plagium/> [↑](#footnote-ref-14)
15. <http://miau.gau.hu/my-x-factor-y> [↑](#footnote-ref-15)
16. <http://miau.gau.hu/miau2009/index_2u.php3?x=e07wiki&string=zene> [↑](#footnote-ref-16)
17. <http://miau.gau.hu/miau2009/index.php3?x=e0&string=roark> [↑](#footnote-ref-17)
18. COCO = component‑based object comparison for objectivity; magyarul: objektivitásra törekvő komponensalapú objektum-összehasonlítás. Az összehasonlítandó objektumok a mindenkori dolgozatok, közülük keresi a rendszer a relatíve legegyedibb és legkevésbé egyedi munkákat. [↑](#footnote-ref-18)
19. <https://miau.gau.hu/mediawiki/> [↑](#footnote-ref-19)
20. A feldolgozási előzményekbe mindenkor valós időben kerülnek be a statikus elemzésnek alávetett dokumentumok mutatószámai is, azonban ezekből már tetszőleges időközönként lehetséges a referencia-adatbázis újbóli hasonlóságelemzéssel történő frissítése. A feldolgozási előzményekben tárolt kellően nagyszámú minta esetén – amennyiben például a dokumentumoknak és készítőiknek statisztikai adatai (a dokumentum nyelve, a készítésének helye, az érintett tudományterület megnevezése, a készítő neme, életkora etc.) is rendelkezésre állnak – lehetséges akár a feldolgozási előzményekből reprezentatív mintavétellel történő referencia-adatbázis képzése; akár az aktuálisan vizsgálandó dokumentummal földrajzi, tartalmi etc. kapcsolatban álló korábbi munkákkal történő összehasonlítás. [↑](#footnote-ref-20)
21. A szövegek képként való megadása kapcsán rá kell mutatni arra is, hogy ez egy fajta kerülő út lehet a kisebb-nagyobb plágiumok feltárásának ellehetetlenítésére, ahol a szerző lényegében minden kritikus részt képként illeszt a szövegbe, mely kapcsán immár ennek tudatában sajnos előbb azt kell vizsgálni karakterfelismerés keretében, vajon milyen betűmennyiség kötődik a grafikákkal kitöltött helyekhez az oldalképekben… [↑](#footnote-ref-21)
22. <http://miau.gau.hu/miau/138/wordlenet.pdf> [↑](#footnote-ref-22)
23. <http://miau.gau.hu/miau/140/la140.doc> [↑](#footnote-ref-23)
24. <http://office.microsoft.com/hu-hu/word-help/dokumentum-automatikus-osszefoglalasa-HP005233452.aspx> [↑](#footnote-ref-24)
25. <http://technet.microsoft.com/en-us/library/cc179199(office.14).aspx> [↑](#footnote-ref-25)
26. OAM = objektum–attribútum mátrix [↑](#footnote-ref-26)
27. Nyilvánvaló esetekben (ld. [országok összehasonlítása](http://miau.gau.hu/miau/173/gdp/GDP-elemzes_final.doc) gazdasági mutatók alapján) a méretfüggetlenítés is trivialitásnak minősül: az ország abszolút mutatói helyett az egy főre vetített értékek már megfelelnek a méretfüggetlenség kritériumának. Jelen esetben azonban nincs a lakossággal analóg mennyiség, amely a szakirodalmi kánon szerint elfogadottan befolyásolná a dokumentum kulcsszavainak internetes gyakoriságát, így ebben a tekintetben legfeljebb a szűkítetlen keresés találati számával lehetne leosztani a szűkített keresések eredményeit. Ez azonban nem elsődleges (bár a továbbiakban szintén vizsgálandó) aspektusa a kérdésnek, ugyanis az originalitását az is befolyásolja egy munkának, hogy mennyire közkedvelt témában keletkezik, gyakran feldolgozott témák esetében nehezebb meggyőzően újat alkotni, s a Kutatócsoport prekoncepciója az, hogy ez a jelenség, ha nem is döntő erővel, de szükségszerűen meg fog jelenni az originalitás-index számításában. (Itt jegyzendő meg az is, hogy ahol ez racionálisan megtehető volt, mint a véletlenszerű mintázás esetén, ott a mutatószámok meghatározása a méretfüggetlenség elvének szem előtt tartásával, százalékos alapon történt.) [↑](#footnote-ref-27)
28. A nem hibátlan modell azt jelzi, hogy a figyelembe vett független változók alapján a függő változó értéke az objektumokat egymáshoz hasonlítva nem magyarázható teljes mértékben. (A végeredmény pontosítható további változóknak az elemzésbe való bevonásával) Más jellegű problémáknál az emberi manipulációra, mint speciális független változóra (pl. szándékosan felülárazott termékek) hívhatja fel a figyelmet. [↑](#footnote-ref-28)
29. Inverz modell: az irányvektorok felcserélésével történő futtatás; hitelesnek akkor tekinthető az eredeti futtatás, amennyiben az ott jól szereplő objektumok az inverz alakban rendre rosszul teljesítenek (vö. a szépségverseny győztese el kellene, hogy veszítse a „csúnyaságversenyt”, amennyiben persze ilyen létezne…) [↑](#footnote-ref-29)
30. A SOI mindenkor a vizsgált dokumentum egy adott referencia-adatbázison, egy adott időpontban értelmezett relatív egyediségét kifejező érték. Használatakor nyilvánvalóan mindkét keretfeltétel említése szükséges, ugyanis (amint az részben már a referencia-adatbázis képzésének általános elveinél (ld. Kutatási problémák, utolsó bekezdés) kifejtésre került) egy adott dokumentum egyedisége nem lehet azonos, ha kifejezetten témájában hasonló, vagy a valaha feldolgozott összes publikáció között kerül értékelésre. Szükséges a SOI meghatározásának időpontját is feltüntetni, amennyiben arra valaki hivatkozni kíván, ugyanis a referencia-adatbázis mindenkor bővül, így az idő múlásával vélelmezhetően kerülnek az adatbázisba hasonló művek, az egyediséget mérő módszer pedig alapértelmezés szerint nem tesz különbséget a rendszerbe kerülő elemek beérkezési dátuma között. Egy adott pillanatban minden, már meglévő elemhez viszonyít, így a korábbi munkák esetében a SOI értékének folyamatos csökkenése várható, mintegy matematikailag demonstrálva a szólást: „Minden csoda három napig tart” [↑](#footnote-ref-30)
31. A definícióból következik, hogy az 1-nél magasabb SOI esetén tekinthető a vizsgált dokumentum az aktuálisan figyelembe referencia-adatbázishoz képest relatíve egyedinek, míg a mélyen 1 alatti SOI-érték a közhelyesség/plágium gyanúját kelti fel. [↑](#footnote-ref-31)