Lengyelné Molnár Tünde

Emberi vagy gépi kivonatolás?*

Lehet-e olyan jó egy szöveg automatikusan létrehozott kivonata, mint az ember által készített referátum? De melyik emberé? Van-e szabályszerűség, hasonlóság a különböző területeken tanuló, dolgozó személyek releváns mondatkiválasztásában? A cikk egy nagy létszámú felmérés alapján keresi ezekre a kérdésekre a választ.

A tanulmány alapjául szolgáló kutatómunka célja a referátumkészítés automatizálásának elősegítése magyar nyelven. A kutatás része volt egy magyar nyelvű offline kivonatoló program fejlesztése is. A kvantitatív tartalomelemzés kategóriájába tartozó program egységeit a szöveg szavai képezik, az output pedig a szöveg mondataiból áll elő. Elkészült a program első verziója, melynek tesztelése nem állhat meg a működésbeli hibák kiküszöbölésénél, a sajátosságok elemzésénél; hatékonyságát az emberi kivonatokkal való összevetéssel kell tesztelni. Készült egy felmérés, amelyben a résztvevőket különböző témájú cikkek kivonatának elkészítésére kértem fel, mivel arra keresem a választ, hogy az emberek által előállított kivonat hasonló-e a gépihez, illetve ha nem, akkor mi az oka az eltérésnek.

Ehhez előbb nézzük meg, milyen módszereket alkalmaz a kivonatoló program!

A program működési elve

A kivonat előállításához vezető út első lépése a szótövek meghatározása, melyet a *Morphologic* cég „Helyeslem” szoftvere szolgáltat. A szótövek gyakorisági értékének meghatározása után történik a szignifikáns szavak megállapítása. Az outputként szolgáló mondatokat súlyozás határozza meg, ahol a mondatban szereplő szignifikáns szavakat pontozzák. A szignifikáns szavak meghatározási módját a felhasználó állítja be, aki jelenleg két módszer közül választhat:

- A *Luhn* módszerével történő meghatározás során akkor szignifikáns a szó, ha az aktuális szótőhöz tartozó előfordulási szám a szövegben háromnál több.
- A szótáralapú feldolgozásban akkor tekintjük szignifikánsnak a szót, ha az megtalálható a szótárban. Ebben a módszerben a felhasználónak lehetősége van saját szótár megadására, így ha rendelkezik bármilyen gyakorisági szótárral, szakterületre jellemző specifikus kifejezések figyelembe vételével tudja elemezni a szöveget. Saját szótár hiányában a program jelen formája a köznyelvi *Szószablya Gyakorisági Szótár* első 10 000 szavából készített adatbázist használja.

A mondat súlyozott pontszámát a szignifikáns szavakon túl a benne szereplő szópárok, szóhármasok és szónégyesek is növelik.[1] Az összpontszám meghatározása után átlagszámításra kerül sor, kiküszöbölve ezzel az eltérő mondathosszúságokból következő egyenlőtlenségeket.

A kutatómunka fontos része volt az emberi kivonatolás hatékonyságának vizsgálata. Törekedtem az emberek kivonatolási technikájában feltárni a szabályszerűségeket és a kapott tapasztalatokat beépíteni a programba.[2] A tartalomelemző elméletek szerint a mondatok cikken belüli elhelyezkedése befolyásolja a mondat szerepét, azaz kivonatkészítéskor súlyozottabban kell figyelembe venni az első, illetve az utolsó bekezdésben található mondatokat. Saját felméréseim is igazolták ezen elméletek egy részét: a vizsgálat alapjául szolgáló különböző témájú cikkek esetében a 340 kitöltőnek több mint fele az első bekezdés mondatait tartotta a leglényegesebbnek, viszont elenyésző részük tulajdonított kitüntetett figyelmet az utolsó bekezdés mondatainak. Ennek eredményét be is építettem a programba, és a szöveg első bekezdésében található mondatokat dupla súllyal vettem figyelembe. A szöveg utolsó bekezdésének mondatai a szakirodalom elméletének megfelelően magasabb súlyt kapnak, de mivel saját kutatásom ezt nem támasztotta alá, ezért csak 20%-kal emeltem a pontértékeiket.

---

* A 2009. április 15-17. között megrendezett Networkshop konferencián elhangzott előadás alapján.

Utolsó lépésként a mondatok gyakorisági értékük alapján kerülnek rendezésre, a megjelenítendő kivonat terjedelmét pedig a felhasználó állíthatja be egy százalékos érték megadásával.

A folyamat végére előáll egy kivonat, amely a szöveg program szerinti leglényegesebb mondatait tartalmazza. *De tényleg ezek a leglényegesebb mondatok?*

A felmérés

Készítettem egy felmérést is, amely a program hatékonyságának vizsgálatát szolgálja. A felmérésben gyakorló könyvtárosok, referátumkészítő szakemberek és informatikus könyvtáros-hallgatók által elkészített kivonatokat hasonlítok össze egymással és a program output-állományával.

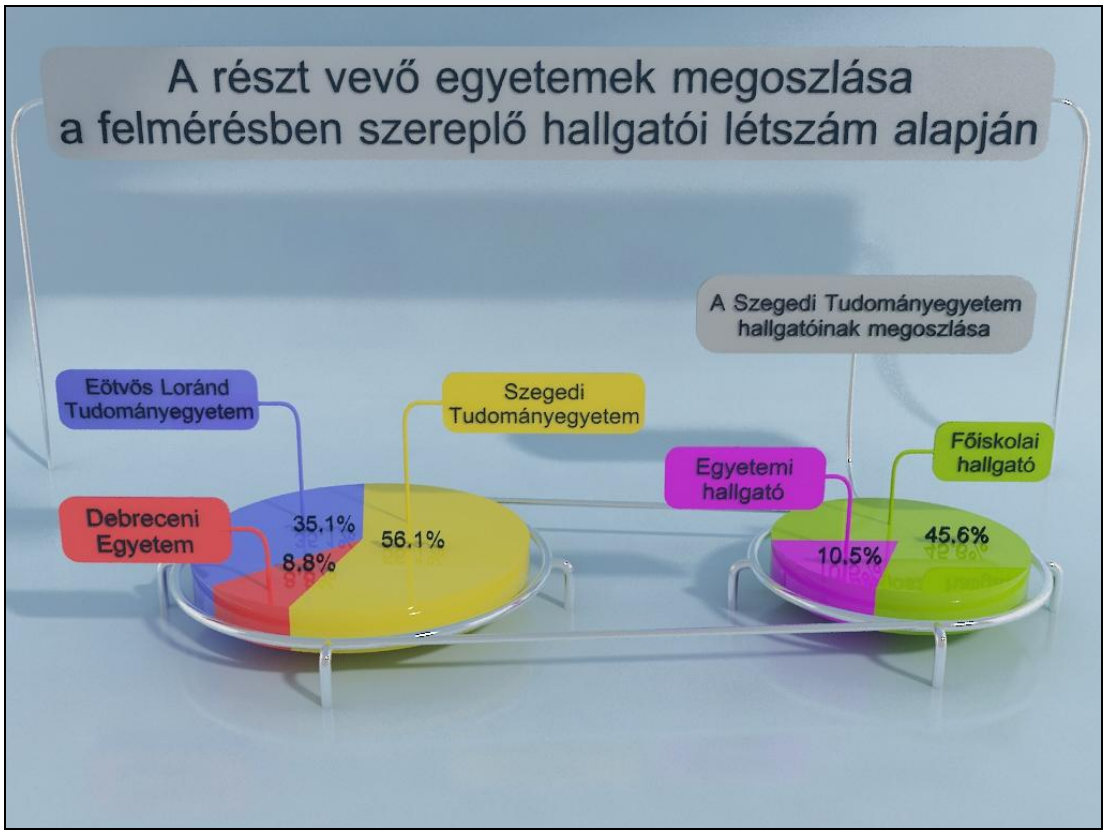
A felmérés alapja

Az empirikus mérés során különböző témájú szakmai cikkek kivonatának elkészítésére kértem fel több felsőoktatási intézmény informatikus könyvtáros-hallgatóit.
Az alapul szolgáló cikkek kiválasztásakor két vezető szakfolyóirat aktuális számaiból választottam egy-egy cikket, így a *Könyvtári Figyelő*ből[3], illetve a *Tudományos és Műszaki Tájékoztatás*ból[4].

A felmérésben résztvevőktől azt kértem, hogy készítsék el mindkét cikk kivonatát a leglényegesebb mondataik megjelölésével és rangsorolásával.[5] A felmérésben szereplő szakemberektől a kivonat mondatainak rangsorolását számítógépen, egy online kérdőív kitöltésével egybekötve vártam, a hallgatókkal a kivonatkészítést – megkönnyítve számukra az áttekinthetőséget – papíron végeztettem el. Számítógépes feldolgozásuk utólagos adatbevitellel történt.

A felmérésben résztvevők

A mintát egyrészt egyetemi, illetve főiskolai képzésben részt vevő informatikus könyvtáros szakos hallgatók képezik az *1. ábrán* látható összetételben:



1. ábra **A részt vevő egyetemek megoszlása a felmérésben szereplő hallgatói létszám alapján**

Nemcsak hallgatókat kértem fel a referátum elkészítésére, hanem szerettem volna adatokat gyűjteni a szakemberek általi kivonatkészítés jellemzőiről is, ezért felkértem a KATALIST levelezőlista olvasóit és több könyvtáros szakembert az online kérdőív kitöltésére.

Mivel választ kerestem többek között arra is, hogy mennyire fontos a szaktudás a kivonatok készítésekor, egy kontrollcsoportot is bevontam a felmérésbe: az egri *Eszterházy Károly Főiskola* magyar szakos hallgatóit. Azért esett rájuk a választás, mert nekik van jártasságuk a különböző hosszúságú és témájú szövegek tömörítésében, a művek lényegének kiemelésében, azonban nincs könyvtártudományi szaktudásuk, így alapot adnak a szaktudás és a kivonatkészítésben való jártasság értékének elemzésére.

Végezetül a felmérésben a *2. ábrán* látható összetétellel vettek részt a kitöltők:

2. ábra **A felmérésben résztvevők megoszlása**

A felmérés célja

A felmérés célja az volt, hogy megvizsgáljam, mennyire hatékony az általam készített kivonatoló program. Mielőtt hozzákezdtem a felméréshez, sejtettem, hogy rengeteg technikai-nyelvészeti problémával kell szembenéznem a program megírása során, de a legnagyobb kérdés az volt, hogy vajon jó lesz-e a program outputja. Ehhez ugyanis tudni kellene, hogy mi tekinthető „jó" kivonatnak. A felmérés eredményétől azt reméltem, hogy a kitöltők sok hasonló kivonatot fognak előállítani, és ez esetben lesz viszonyítási alapom.

Bemutatom a program által előállított mondatok és a felmérésben részt vevő személyek által előállított kivonatok mondatai közötti hasonlóságokat és eltéréseket, választ keresve arra, hogy létezik-e globális kivonat, továbbá, hogy van-e különbség az automatizálás alapjául választott két módszer adta output között: Luhn módszere és a szótár alapján történő kivonatolás eredménye között.

Elemzés

Az összehasonlítás alapjául a kitöltők által megadott mondatok súlyozva, mintacsoportonként kerülnek összevetésre. A súlyozásra azért volt szükség, hogy kiküszöbölhető legyen, mi számít relevánsabb mondatnak: a 10 ember által első helyre tett, vagy a 25 ember által a 10. helyre rangsorolt? Ezen felül természetesen készültek elemzések, amikor az eredeti kivonatok lettek összevetve egymással.

Mielőtt összevetnénk a program outputjával a felmérés eredményeként előállt, emberek által készített kivonatokat, nézzük meg röviden a mintacsoportok véleménye közötti hasonlóságokat és eltéréseket![6]

Kezdjük az interdiszciplinárisabb témával foglalkozó *Koltay Tibor* cikk-kivonatainak elemzésével!

A hallgatói mintacsoportok kivonata nagyfokú hasonlóságot mutat. Az informatikus könyvtáros szakos főiskolás és egyetemista hallgatói mintacsoport súlyozással előállított kivonata csak két mondatban tér el egymástól, illetve a kontrollcsoportot képző magyar szakosok kivonatától is. Ez 88%-os egyezőséget jelent.

A szakemberek kivonata ezzel szemben lényegesen eltér, az azonos mondatok legjobb esetben is csupán a kivonat 47%-át teszik ki.[7]

Érdemes megvizsgálni, hogy ha nemcsak azt a 17 mondatot vesszük figyelembe, amely bekerült a kivonatba, hanem minden mintacsoportnál megnézzük, hogy az eredeti cikk összes mondata milyen súlyozott pontokat kapott, és ezt elemezzük egy korrelációs mátrixszal, akkor milyen eredményt kapunk (*1. táblázat*)!

1. táblázat
Súlyozott pontszámokból képzett korrelációs mátrix

Súlyozott pontszámokból képzett korrelációs mátrix	Informatikus könyvtáros egyetemista hallgatók	Informatikus könyvtáros főiskolás hallgatók	Szakemberek	Magyar szakos hallgatók
Informatikus könyvtáros egyetemista hallgatók	1	0,917	0,486	0,872
Informatikus könyvtáros főiskolás hallgatók	0,917	1	0,445	0,856
Szakemberek	0,486	0,445	1	0,441
Magyar szakos hallgatók	0,872	0,856	0,441	1

A táblázat alapján jól látható, hogy a hallgatói mintacsoportok szaktól, tanulmány szintjétől függetlenül nagyon hasonló módon értékelik az egyes mondatok relevanciáját. Bár kiemelhető, hogy az egyetemista és főiskolai szintű képzésben részt vevő hallgatók esetében 0,9 fölötti nagyon szoros korrelációs kapcsolat áll fenn, azaz, ha az egyik mintacsoportba tartozó személyek előkelő helyre rangsoroltak egy mondatot, akkor a másik mintacsoport tagjai is relevánsnak tartják azt.

A szakemberek esetén már nem ekkora az egyetértés. Az ő véleményükhöz az egyetemista informatikus könyvtárosok mondatkiválasztási technikája áll a legközelebb, de ez is csak 0,486-os korrelációs értékkel. A szakemberek eltérő gondolkodásmódját jól tükrözi, hogy az 50 szakember 32 különböző mondatot rangsorolt az első helyre, és amit a legtöbben megjelöltek, az is csupán a kitöltők 12,5%-ának jelölését tudhatja magáénak, míg ez az érték a hallgatóknál 50% fölötti.

A másik, a könyvtártudományhoz sokkal jobban kötődő *Prokné Palik Mária* által írt cikk esetében előfordul olyan mondat is, amelyet az egyetemista informatikus könyvtáros kitöltők több mint 85%-a bevett a kivonat mondatai közé. Náluk és a magyar szakos kontrollcsoportnál hat mondat van, amelyet több mint a kitöltők felének kivonatában megtalálhatunk. Ez az érték a főiskolás informatikus könyvtáros-kitöltőknél hét mondat.

A szakemberek véleménye ennél a cikknél is teljesen eltérő struktúrát alkot, nincs olyan mondat, amelyben a kitöltők harmada egyetértett volna a kivonatba való beválasztás során.

A súlyozás útján előállt kivonatok hasonló képet mutatnak, mint az előző cikk esetén: az egyetemista és a főiskolás informatikus könyvtáros-hallgatók súlyozott kivonata csupán egy-egy mondatban tér el egymástól, és 0,936-os korrelációs értékkel az összes mondat pontszáma hasonló gondolkodásmódra utal.

A szakemberek kivonata is összhangban van az előző cikknél kapott értékekkel: a 17 mondatos kivonatból csak hét-hét mondatban egyezik meg a szakhallgatók kivonatával. Ez 41%-os egyezőséget jelent.

Viszont meglepő képet kapunk, ha a kontrollcsoport kivonatát elemezzük! Míg *Koltay Tibor* cikkénél két-két mondatban tért el az informatikus könyvtáros-mintacsoportoktól a súlyozott kivonat, a diszciplinárisabb témájú cikknél már öt-, illetve hatmondatos az eltérés, miközben a könyvtár szakosok között nagy az egyetértés. Az igazán váratlan fordulatot az okozza, hogy a könyvtártudomány szakembereinek kivonata a magyar szakos kontrollcsoport kivonatával mutat a legnagyobb hasonlóságot (53%-os egyezőség, a szakhallgatókéval pedig 41%-os az egyezés).

További érdekesség, hogy a felmérésben részt vevő
- 51 szakember 37 különböző mondatot választott ki legrelevánsabb mondatnak,
- míg a 241 szakos hallgató 30 mondat közül választott,
- a 48 fős magyar szakosokból álló kontrollcsoportnál 10 mondat került az első helyre valamely résztvevőnél.

Összefoglalva: vegyes eredményekkel kezdhetjük el a program hatékonyságának tesztelését. *Míg a hallgatók között interdiszciplináris témánál magas fokú egyezőséget tapasztalhatunk, addig a szakemberek eltérő módon látják a releváns mondatokat. Ha a cikk a könyvtártudomány szakkifejezése-*

iben bővelkedik, akkor a magyar szakos hallgatók véleménye leszakad az informatikus könyvtáros hallgatókétól, a szakemberek pedig továbbra is egyéni véleményeket tükröznek.

A továbbiakban az adatokat cikkenként párhuzamosan láthatjuk a két módszernek megfelelően, továbbá a két cikk szerzője is segítette a felmérést, és elkészítette saját cikkének kivonatát, rangsorolva a kivonatba került mondataikat, így lehetőséget adtak arra, hogy az ő véleményükhöz is viszonyíthassunk.

Program kontra ember

Koltay Tibor cikkének elemzése

A program eredményeként Koltay Tibor cikkénél a 86 mondat közül Lunh módszerével történő kivonatkészítés során 84 mondat kapott pontszámot, míg a Szószablya szótárának alkalmazása során valamivel kevesebb, a mondatok 91%-a, 78 mondat kapott 0-tól eltérő értéket. A 84 mondatból arra lehet következtetni, hogy a szerző a mondataiban olyan szavakat használ, amelyek a szövegben többször is előfordulnak. 47 olyan különböző szó van a cikkben – a tiltott szavakon kívül –, amely 3-nál többször fordul elő. A szövegben található 526 szótő közül 313 szerepel a Szószablya szótárban. A szópárokat és szóhármasokat vizsgálva a szövegben egyetlen szónégyes, 5 db szóhármas, és 38 db szópár található. (Ezek meghatározása a szótövek alapján történik.)

A két módszerrel történő elemzés során kapott pontszámok alapján a *2. táblázat* mondatait ítéli a program a legrelevánsabbnak. (Mint az eddigi elemzések során, itt is tekintsük meg az első 17 mondatot, a 20%-os kivonatot, de természetesen a program használata során lehetőség van más terjedelem listázására is.)

3. táblázat
A program által generált kivonat mondatai

Sorrend	A kivonat mondatai Luhn módszerével történő elemzés alapján	A kivonat mondatai a Szószablya szótár alapján
1.	1. mondat	1. mondat
2.	81. mondat	50. mondat
3.	83. mondat	21. mondat
4.	17. mondat	77. mondat
5.	38. mondat	81. mondat
6.	51. mondat	78. mondat
7.	69. mondat	38. mondat
8.	21. mondat	76. mondat
9.	76. mondat	5. mondat
10.	23. mondat	3. mondat
11.	86. mondat	82. mondat
12.	3. mondat	6. mondat
13.	5. mondat	17. mondat
14.	37. mondat	32. mondat
15.	77. mondat	51. mondat
16.	84. mondat	55. mondat
17.	22. mondat	80. mondat

A két módszerrel kapott kivonat 58,82%-ban tartalmaz azonos mondatokat: a 17 mondat közül 10-et. Az első helyre került mondat megegyezik a két módszernél, mely eredményt magyarázza a magasabb súly, ugyanis az első bekezdés mondatait magasabb súllyal veszik figyelembe.

Nézzük meg, hogy a program outputjaként előállt mondatok mennyire vannak összhangban a mintacsoportok kivonatával (*3. táblázat*)!

2. táblázat
Számítógépes output elemzése Koltay Tibor cikke esetében

Mintacsoportok	Luhn módszere alapján		Szószablya szótár alapján	
	Egyező mondatok száma	Egyezés aránya	Egyező mondatok száma	Egyezés aránya
Informatikus könyvtáros egyetemista hallgatók	3	17,65%	3	17,65%
Informatikus könyvtáros főiskolás hallgatók	4	23,53%	3	17,65%
Szakemberek	4	23,53%	6	35,29%
Magyar szakos hallgatók	3	17,65%	3	17,65%
Szerző kivonata	5	29,41%	3	17,65%

Mindkét módszert tekintve alacsony az egyező mondatok száma, a legmagasabb egyezés is csak 35%-os.

Luhn módszerét alkalmazva – ahol a gyakrabban előforduló szavak számítanak relevánsnak – a szerző válaszai alapján előállt kivonattal a legmagasabb az egyezés, a hallgatói csoportoknál pedig szinte teljesen azonos eredményt kapunk. Az egyező mondatok közül kettő minden mintacsoportnál megtalálható, de ez nem meglepő, hiszen az elemzés során a mintacsoportok súlyozott kivonatai alig tértek el egymástól. Érdekes, hogy a szerző saját kivonatával a legmagasabb az egyezés, pedig ha a szerző kivonatát vetjük össze a mintacsoportok súlyozott mondataival, akkor maximum nyolcmondatos egyezést találunk.

Koltay Tibor cikkének vizsgálatakor a Szószablya szótár alapján létrejött kivonat hasonló eredményt mutat, mint a Luhn módszerével kapott. A legmagasabb egyezés itt a szakemberek kivonatával áll fenn. A három hallgatói csoporttal ugyanazon mondatokban egyezik meg a program kivonata.

Összegezve: megállapítható, hogy Koltay Tibor cikke esetén – amely általánosabb témával foglalkozik – az egyezés nem túl magas számú a program kivonata és a mintacsoportok kivonata között.

Nézzük meg, hogy Prokné Palik Mária cikke hasonló eredményt mutat-e?

Prokné Palik Mária cikkének elemzése

Ennél a cikknél is Luhn módszerének alkalmazásával kapott több mondat pontszámot, bár a különbség kisebb, mint Koltay Tibor cikkénél. A szöveg 83%-a, 69 mondat ért el valamilyen pontértéket, ugyanis a szövegben 37 releváns szó található (amely háromnál többször előforduló nem tiltott szó). A Szószablya szótár alapján három mondattal kevesebb, a szöveg 85,5%-a kapott pontot, amelynek alapja, hogy a szöveg 487 szava közül a Szószablya szótárban megtalálható 297 szó. A szövegben 1 szónégyes, 6 szóhármas, és 40 darab szópár található.

A pontszámok alapján kialakult sorrendben az első 17 mondat eredményét a *4. táblázat* mutatja.

4. táblázat
A program által generált kivonat mondatai

Sorrend	A kivonat mondatai Luhn módszerével történő elemzés alapján	A kivonat mondatai a Szószablya szótára alapján
1.	6. mondat	6. mondat
2.	22. mondat	63. mondat
3.	3. mondat	3. mondat
4.	63. mondat	36. mondat
5.	21. mondat	46. mondat
6.	7. mondat	5. mondat
7.	10. mondat	48. mondat
8.	36. mondat	11. mondat
9.	50. mondat	47. mondat
10.	19. mondat	32. mondat
11.	42. mondat	21. mondat
12.	61. mondat	22. mondat
13.	64. mondat	24. mondat
14.	5. mondat	53. mondat
15.	48. mondat	74. mondat
16.	39. mondat	34. mondat
17.	68. mondat	1. mondat

A két módszerrel előállított kivonat 11 mondatban egyezik meg, amely érték Koltay Tibor cikkénél 10 mondat volt. Az eredményt azonban befolyásolja az első bekezdés magasabb súlyozása. A cikk első bekezdése hat mondatot tartalmaz, amelyből a Szószablya szótár alapján létrejött kivonatba három mondat került be, míg Luhn elvével hozva létre a kivonatot, az első bekezdés mondatai közül öt mondat található meg a kivonatban.

Prokné Palik Mária cikke esetében a Luhn módszerével kapott kivonat mondatai szinte minden mintacsoportnál[8] legalább kétszer annyi közös mondatot tartalmaznak, mint amennyi a szótár alapján készített kivonatban előáll.

Nézzük meg az egyező mondatok számát és arányát bemutató táblázatot (*5. táblázat*)!

5. táblázat
Számítógépes output elemzése Prokné Palik Mária cikke esetében

Mintacsoportok	Luhn módszere alapján		Szószablya szótár alapján	
	Egyező mondatok száma	Egyezés aránya	Egyező mondatok száma	Egyezés aránya
Informatikus könyvtáros egyetemista hallgatók	6	35,29%	2	11,76%
Informatikus könyvtáros főiskolás hallgatók	6	35,29%	3	17,75%
Szakemberek	8	47,06%	4	23,53%
Magyar szakos hallgatók	8	47,06%	5	29,41%
A szerző kivonata	4	23,53%	2	11,76%

A Luhn módszerével kapott kivonat mondatai nagyobb egyezést mutatnak a mintacsoportok kivonatával, mint ha a szerző kivonatát viszonyítjuk a mintacsoportokéhoz, illetve ha a szakemberek kivonatát viszonyítjuk a hallgatói csoportok eredményeihez.[9]

A szótár alapján készült kivonat mondatai alacsony egyezést mutatnak. A mondatok is változatosabbak, nincs olyan mondat, amely az összes mintacsoportnál megtalálható.

Véleményem szerint *a kapott eredmények oka a cikkek tartalmában, szövegében fedezhető fel*. Prokné Palik Mária cikke témáját és szövegezését tekintve is szakcikk, amelyben sok könyvtári szakkifejezést használ. Mivel a Szószablya szótár általános témájú, ezek közül kevesebb található meg benne. A Luhn módszerével 3-nál többször előforduló szavak 75%-a van meg a Szószablya szótárban. Ennek következtében nem a szakszavakat tartalmazó mondatok lesznek súlyozva, és ez magyarázza a mintacsoportok kivonatával való alacsonyabb egyezést. Luhn módszerével a szakmailag lényeges mondatok jobban előtérbe kerülnek, hiszen ha a szerző egy szakkifejezést többször használ, akkor az azt tartalmazó mondatok meg is kapják érte a pontértéket. Ellentmondásnak tűnhet a két output közötti magas átfedés. Ne feledjük azonban, hogy a szópárokért, szóhármasokért, az esetleges szónégyesekért mindkét módszernél pluszpontok járnak. Ez, illetve az első bekezdés súlyozása is oka a több közös mondatnak a két kivonat között.

Mivel a Koltay Tibor cikkében található szakszavak nemcsak a könyvtártudományhoz köthetők, hanem interdiszciplinárisak, *elmondható, hogy az automatikus kivonatolás szakcikkek esetében hatékonyabban használható*. Itt a leggyakrabban előforduló kifejezések 99%-a megtalálható a Szószablya szótár szavai között is, ezért nem tapasztaltunk lényeges eltérést a két módszer kivonata között.

Következtetések a felmérés eredményei alapján

A hallgatói mintacsoportok kivonataiban nagy az átfedés, a szakemberek és a szerzők kivonatával több közös mondatot találunk, de már szerkezetükben, összetételükben is eltérnek a hallgatói csoportok eredményétől.

A mintacsoportok kivonatainak elemzésénél kimutatható az első mondatok hangsúlya. Koltay Tibor cikke esetén az első három mondat mind a négy mintacsoportnál megtalálható,[10] míg Prokné Palik Máriánál az első két mondat található meg minden mintacsoportnál, és mellette még mindegyiknél van más mondat is az első bekezdésből. Ennek hatására a programba is beépítésre került, hogy az első bekezdés mondatai nagyobb súlyt kapjanak. Azonban a két alapul szolgáló cikk közül az elsőnek csak egyetlen mondatból áll az első bekezdése, így ezért a súlyozással csak az első mondat került be a számítógépes outputba, míg a másik cikknél hat mondatból áll az első bekezdés, ezért a súlyozás hatására három-öt mondat bekerült a kivonatba.[11] A mintacsoportoknál mindkét esetben három-három mondat származik a szöveg elejéről. Ezzel a módszerrel csak közelíteni lehetett a mintacsoportok eredményéhez, teljes mértékben nem sikerült elérni.

Kérdés, hogy szükséges-e egyáltalán ugyanolyan kivonat elérése, mint a mintacsoportoknál kapott eredmény? Beszélhetünk-e globális kivonatról, az

„abszolút lényegről”? Ennek megválaszolása nem könnyű, ezért nézzük meg a következő elemzést, amely talán közelebb visz minket a megoldáshoz!

Az automatizálás szempontjából a legfontosabb azoknak az elveknek a feltárása, amelyek az összes mintacsoportra teljesülnek. A kivonatok részletes elemzése már megtörtént, bemutattuk a csoportok kivonatainak szerkezetét, a közös mondatok számát, elhelyezkedését, csoportonkénti átfedését. Annak a lényeges eredménynek a kimutatására azonban még nem került sor, amelyben megvizsgáljuk, hogy hány olyan mondat van, amely az összes mintacsoportnál, illetve a szerzőnél is megtalálható a végső kivonatban.

Koltay Tibor cikke esetén három olyan mondat van, amely az összes mintacsoport és a szerző kivonatában is megtalálható.[12]

Prokné Palik Mária cikke esetén *nincs egyetlen olyan mondat sem, amely mind a négy mintacsoport és a szerző kivonatában is megtalálható lenne*. Hat olyan mondat van,[13] amelyek a mintacsoportok kivonatában közösek, ebből három található meg a kivonatoló program Luhn módszerével történő előállítás során, szótáralapú előállításnál pedig egyik sem szerepel a 20%-os kivonatban.

Azt tapasztaljuk, hogy a négy mintacsoportot és a szerzőt alapul véve a különböző személyek eltérően látják a cikkek lényegét. Bár *hasonlóságról beszélhetünk, de azonosságról semmi esetre sem*.

Megvizsgáltam az egyéni eredeti kivonatokat. A 340 ember kivonata között *nincs két olyan személy, aki pontosan ugyanazt a kivonatot* – azonos mondatok, azonos sorrenddel – *hozta volna létre bármely cikket is véve alapul*. Majd a követelményekből engedve az is vizsgálat tárgya volt, hogy hány olyan személy van, akinek a kivonatába ugyanazok a mondatok kerültek, eltekintve a sorrendtől. Ennek eredménye a *6. táblázat*ban látható.

Koltay Tibor cikkénél van 7 olyan személy, akik ugyanazon mondatokat választották be a kivonatukba, míg a másik cikknél 6 azonos kivonatot találunk. Ez nagyon alacsony érték! A felmérés eredménye alapján *nem beszélhetünk egyedüli, tökéletes KIVONAT-ról*. Láthatjuk, hogy a vizsgálati alanyok között alig találunk olyan személyeket, akik pontosan ugyanazt a kivonatot készítették volna el, így a számítógépes kivonatolástól sem várhatjuk, hogy megfeleljen egy adott eredménynek, mert nincs ilyen mintakivonat. A súlyozással előállított kivonat jó képet ad az egyes mintacsoportok tagjainak véleményéről, alkalmas a szerkezeti összetétel elemzésére, szabályszerűségek levonására, illetve a súlyozás jól tükrözi, hogy egy mondatot a többség előkelőbb helyre tesz, vagy kevésbé fontosnak tart. Az így előállt kivonat azonban nem tekinthető globális kivonatnak, hiszen amint kiderült, ezt mindenki másként látja.

6. táblázat
Egyező kivonatok alakulása

	Esetek száma	
	Koltay Tibor cikke	Prokné Palik Mária cikke
7 azonos kivonat	1	-
6 azonos kivonat	-	1
5 azonos kivonat	-	-
4 azonos kivonat	2	4
3 azonos kivonat	8	14
2 azonos kivonat	44	41
Teljesen különböző kivonat	174	149
Nem készített kivonatot	39	45

A két cikk eltérő témája szintén jól tükrözi az automatizálhatóság nehézségeit! Míg egy interdiszciplináris témánál a mintacsoportok között nagyobb egyetértéssel találkozunk,[14] addig az egy szaktudományhoz kapcsolódó cikk már jobban megosztja az olvasóközönséget. A számítógépes kivonatoló program két módszerrel történő szignifikáns szómeghatározása is eltérő képet mutat az eltérő tartalmú cikkek esetén. *Szakszöveg kivonatolása során Luhn módszere hatékonyabbnak bizonyul*,[15] nagyobb átfedést mutat a mintacsoportok súlyozással létrejött kivonatainak mondataival, mint a Szószablya szótár alapján történő kivonatkészítés. *Az interdiszciplináris szövegnél nincs lényeges eltérés az általános szótáron alapuló és a Luhn módszerével történő kivonatkészítés között.*

Összefoglalva: megéri esélyt adni a kivonatolás automatizálásának. Valószínű, hogy programmal még sokáig nem lehet a művekben lévő gondolatokat visszaadni; sőt még a mondatkiválasztás eredményessége is kritizálható, de mint láthattuk, az emberi kivonatkészítés sem a sémákon alapuló egységes gondolatok tükröződése. Minden ember

másként választja ki a releváns mondatokat, és a különböző szakterületeken dolgozó szakemberek látják a legeltérőbben a cikkek lényegét. Ezért úgy vélem, megéri energiát fektetni egy magyar nyelven működő kivonatoló programba, és lesznek területek, ahol majd ennek hasznát veszik.

Irodalom és jegyzetek

[1] A tiltott szavak nincsenek figyelembe véve.

[2] Az eredményeket részletesen az alábbi cikkben ismertettem: LENGYELNÉ Molnár Tünde: A kivonatkészítés sajátosságai egy felmérés adatainak a tükrében. = Könyvtári Figyelő, 53. köt. 2. sz. 2007. p. 475–495. <*http://www.ki.oszk.hu/kf/e107_plugins/content/content.php?content.56*>

[3] KOLTAY Tibor: Szöveg, információ, relevancia: néhány adalék a témakörhöz. = Könyvtári Figyelő, 51. köt. 3. sz. 2005. p. 514–518.

[4] PROKNÉ Palik Mária: A tartalmi feltárás problémái online könyvtári katalógusokban. = TMT, 52. köt. 11–12. sz. 2005. p. 525–527.

[5] Mivel a cikkeknek 20%-os tömörítését vártam el a felmérésben résztvevőktől, ezért mindkét cikk esetén 17 mondatot kellett megjelölniük.

[6] Az adatok részletesen megtekinthetők az Irodalom 2. hivatkozásában.

[7] Az egyetemista informatikus könyvtáros mintacsoport kivonatával az egyezés 47%; a főiskolai informatikus könyvtáros mintacsoport kivonatával az egyezés 35%; a magyar szakos mintacsoport kivonatával az egyezés 47%.

[8] A magyar szakosoknál 1,6-szoros.

[9] A szerző kivonata a mintacsoportok kivonatával 6–6–2–4 mondatban egyezik meg (a táblázatban található sorrendnek megfelelően); a szakemberek kivonata a hallgatói csoportokkal 7–7–9 mondatban egyezik meg.

[10] Az egyetlen kivétel: a főiskolás informatikus könyvtáros hallgatóknál a második mondat nem szerepel a kivonatban, de az első és a harmadik náluk is.

[11] Szószablya, illetve Luhn elve alapján történő kivonatolás során.

[12] Az 1., a 3., illetve a 26. mondat.

[13] 1,. 2., 7., 19., 30., 41. mondatok.

[14] A súlyozás után létrejött kivonatokra alapozva.

[15] Ha a hatékonyság mértékének mintacsoportokként az egyéni kivonatokból létrehozott kivonatokkal való egyezést tekintjük.

Beérkezett: 2009. IX. 14.-én.

Lengyelné Molnár Tünde
az Eszterházy Károly Főiskola Médiainformatika Intézetén főiskolai docens, tanszékvezető-helyettes.
Email: *mtunde@ektf.hu*

Népszerűbb e-könyvek, több kalózmásolat

Az e-könyvek növekvő népszerűségével párhuzamosan egyre inkább attól rettegnek a kiadók és a szerzők, hogy ezek piacán is megjelennek a kalózkópiák.

A fájlcserélés, illetve a kalózmásolatok megjelenésének lehetősége már most komoly aggodalmat kelt a könyvkereskedők és -kiadók körében. Abban ugyanakkor a piaci szereplők egyetértenek, hogy nem akarják ugyanazokat a hibákat elkövetni, mint amiket a zeneipar korábban.

A kiadók, a kereskedők és a szerzők félelme nem alaptalan. Néhány fájlcserélő portálon már felbukkantak az első e-könyv-kalózkópiák. Egyelőre még nem lehet komoly problémáról és bevételkiesésről beszélni, de a könyviparnak már most megfelelő stratégiát kell kidolgoznia, ha nem akar a zeneipar sorsára jutni.

Az elmúlt évek nem igazán voltak sikeresek a könyvkereskedők számára. A komoly érdeklődésre számot tartó művek eladása ugyanis az Egyesült Államokban tavaly 13 százalékkal csökkent. 2008-ban az e-kiadványok csak az eladott művek 1,6 százalékát tették ki. Idén nyáron viszont már rekordokat döntött a forgalmuk az USA-ban, így ez az arány a közeljövőben jelentősen megváltozhat. Ennek oka, hogy egyrészt a korábbinál több elektronikus könyvet adnak el világszerte, másrészt a gyártók is folyamatosan újabb olvasókat dobnak piacra. A szakemberek e két dolog miatt szinte biztosra veszik, hogy a kalózpéldányok száma is nőni fog.

Ed McCoyd, az *Association of American Publishers* ügyvezető igazgatója szerint már most megfigyelhető, hogy jelentősen emelkedett a fájlcserélő hálózatokban elérhető illegális e-könyvkópiák száma.

Tény, hogy alig néhány órával a rég várt alkotás megjelenése után a kalózok is lecsaptak *Dan Brown* új könyvére. A hangoskönyvet és az e-könyvet több tízezren töltötték le illegálisan.

/SG.hu Hírlevél, 2009. október 6., *http://www.sg.hu/*

(SzP)