VÉLEMÉNYEK

Elemezd a szakirodalmat szövegbányászattal! Példák a gyermekkori túlsúly és elhízás témakörében

Analyze the literature with text mining! Examples of checking the literature of childhood overweight and obesity

Ismertetők: Balkányi László, Vitrai József

Doi: 10.58701/mej.11195

Kulcsszavak: szövegbányászat; népegészségügy; túlsúly; elhízás; gyermekkor
Keywords: *text mining; public health; overweight; obesity; childhood*

Összefoglaló

A cikk szövegbányászati eszközöket mutat be nagy mennyiségű szakirodalom feldolgozására. Megállapítható, hogy jelentős fejlesztések történtek a témában az elmúlt években, és számos, a kutatók által közvetlenül használható eszköz és módszer áll már rendelkezésre. A szerzők a gyermekkori túlsúly és elhízás szakterületének irodalmán mutatnak be 7 módszert, 2 eszköz segítségével (VOYANT TOOLS és VOSVIEWER), az egyszerűbbtől a bonyolultabbak felé haladva: (1) a vizsgált ún. *szövegtest* kialakítása, (2) a közlemények számossága, időbeli trendje, (3) a szakkifejezés-gyakoriság elemzése, (4) a kollokáció (együttes előfordulás, együttjárás) elemzése, (5) korrelációelemzés, (6) kontextuselemzés, (7) fogalomhálózati elemzés, fogalomtérképezés – klaszterezés (csoportosítás). A módszerek és eszközök használatát példákon mutatják be. A módszerek együtt használva feltárják a legfontosabb szakterületi fejleményeket, trendeket. Például azonosítható volt az intervenciók változása az elmúlt tíz évben. A vizsgálat a medikalizáció (orvosi szemlélet) súlyának növekedését mutatta ki. Lényeges, hogy az egyszerűbb szövegbányászati programok már a szokásos eszközeinken is futtathatók, és nem igényelnek speciális szakértelemet sem. A megbeszélésben említett további, szofisztikált technikák, különösen a mesterséges intelligencia algoritmusok alkalmazása azonban speciális szaktudás és megfelelő kapacitású számítástechnikai eszközpark szükséges. Ez még nehezen érhető el a népegészségügyi kutatók számára. Összefoglalásul a szerzők megállapították, hogy a fejlett szövegbányászati eszközök használata lehetőséget ad a szakirodalom gyors, kvantitatív jellemzőkkel leírt, transzparens áttekintésére, amelyre akkor lehet szükség, ha egy-egy szakterület közleményeink száma meghaladja a vizsgált időszakban az ezres–tízezres nagyságrendet.

Summary
The article presents text mining tools for processing the vast amount of literature. It is noted that significant developments have taken place in recent years and that several tools and methods are available for direct use by researchers. The authors present 7 methods in the field of childhood overweight and obesity, using 2 tools (VOYANT TOOLS and VOSVIEWER), progressing from the simplest to the more complex analysis: (1) the design of the corpus, (2) trending of publication numbers, (3) the analysis of the frequency of terms, (4) collocation, (5) correlation, (6) context analysis, (7) concept network analysis, concept mapping - clustering. The use of methods and tools is illustrated with examples. These methods, used together can identify the most important developments and trends in the field, e.g. changes in interventions over the last ten years under study. Medicalization was also demonstrated during the study period. Crucially, the simpler text mining tools can be run on our usual IT tools and do not require any special expertise. However, additional sophisticated techniques mentioned in the discussion, in particular the use of artificial intelligence algorithms, require specific expertise and a computing toolkit of sufficient capacity. This is still difficult to access for public health researchers. In conclusion, the use of advanced text mining tools provides the opportunity for a rapid, quantitative, and transparent review of the literature, which becomes necessary when the number of publications in a given field exceeds the thousands to tens of thousands in the period under review.

BEVEZETÉS

Az alábbi munka arra tesz kísérletet, hogy az érdeklődő kutatók számára bemutassa az egyszerű szövegbányászati eszközök használatát a szakirodalom áttekintésében. Ahhoz, hogy megértsük, miért is van erre szükség, szükség van egy rövid háttérmagyarázatra és visszatekintésre. Általában a tudományos kutatások hitelességét a kutatás adatainak és az adatfeldolgozás módszereinek megismerésével tudjuk megítélni. Számos közlemény áttekintésének klasszikus formája a szisztematikus áttekintés (SZÁ, *systematic literature review*), amely átlátható és reprodukálható módon szintetizálja a kutatási eredményeket. James Lind klasszikus skorbut tanulmánya már 1753-ban megelőlegezte az SZÁ módszertanát (Bartholomew, 2002). Majd az 1970-es és '80-as évektől kifejlődött egy szigorú módszertan, ami az 1993-ban megalakult a *Cochrane Collaboration*[1] munkájának nyomán teljesedett ki, és vált általánosan elfogadott módszerré (Sur & Dahm, 2011). A 21. század azonban jelentős változást hozott. A kutatás volumene, a publikált eredmények száma egy-egy szűkebb szakterületen is már-már áttekinthetetlenné vált, meghaladva a tízezer cikk/év értéket (Balkányi et al., 2021). Segítséget – többek között – az információtudomány eszközeitől kaphatunk. A lehetőségek egyike az a megközelítés, amikor magukat a kutatási szövegeket értelmezzük és vizsgáljuk nagy mennyiségű adatként. Az ekkor alkalmazott eszköztárat nevezzük szövegbányászatnak (*text mining*). Ebben az írásban erre mutatunk példát, az egyszerűbb eszközöktől haladva a bonyolultabbak felé.

[1] https://www.cochrane.org/

A vizsgálathoz olyan szövegbányászati eszközöket (programokat) használunk, amelyek ingyenesen hozzáférhetők, és amelyeknél a használt algoritmusok leírását közzétették. Így a végzett kutatás teljesíti a reprodukálhatóság feltételeit. Az eszközök saját, intuitív grafikus felhasználói felülettel elérhetők, nem szükséges programozási tudás az elemzések elvégzéséhez. Annak érdekében, hogy az eszközök és módszerek használhatóságát bemutathassunk, egy szűkebb, viszonylag jól ismert, és kutatott szakterületet választottunk, amelyiken a népegészségügyi szakemberek saját háttértudása segíthet megítélni az eszköz használhatóságát.

MÓDSZERTAN ÉS EREDMÉNYEK

Az alábbiakban a különböző szövegbányászati eszközök, módszerek ismertetése után közvetlenül az azokkal nyert eredmények is bemutatásra kerülnek.

1. Szövegtest kialakítása

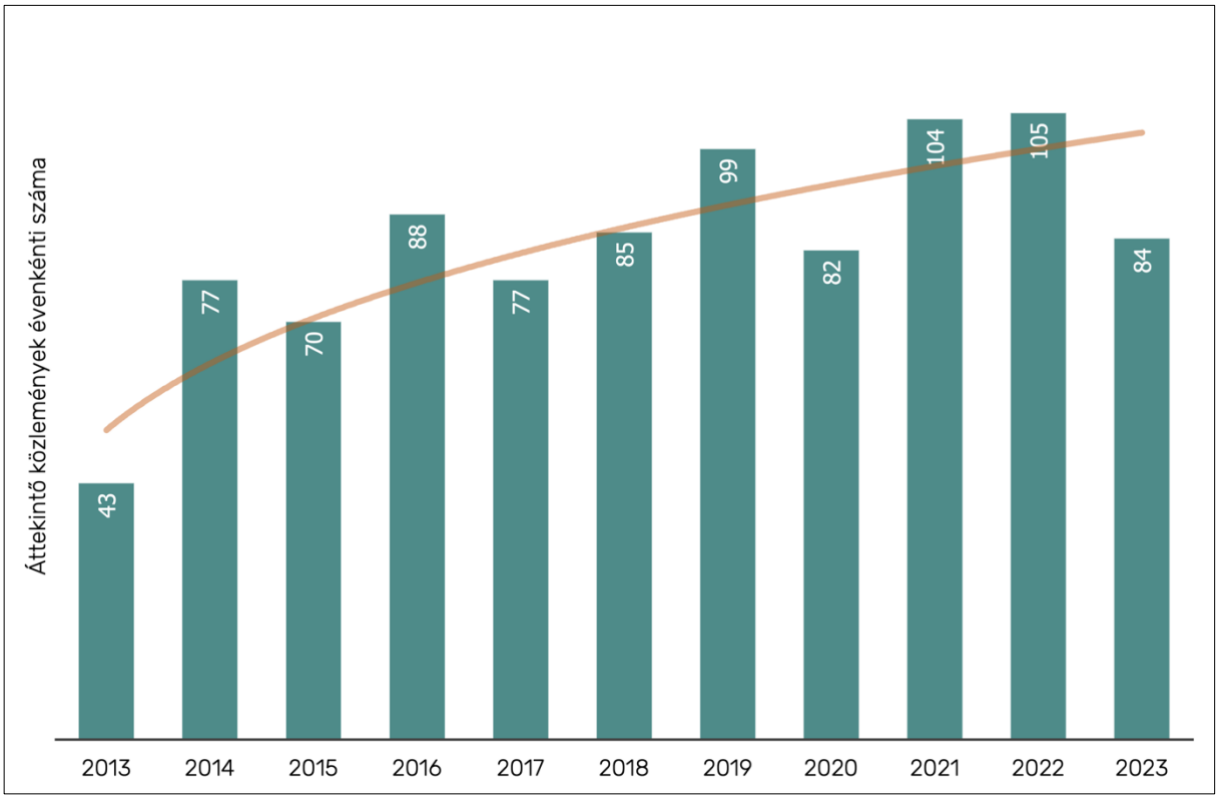
Az elemzés első lépése a vizsgálandó közlemények adatbázisának, az ún. szövegtest (*corpus*) összeállítása. Ehhez a kiválasztott témának megfelelő keresőkifejezéssel (*search string*) kiválogatjuk a szakirodalmi adatbázis(ok)ból a minket érdeklő közleményeket. Ebben az írásban témaként a gyermekkori túlsúlyt és elhízást választottuk, szakirodalmi forrásként a bárki számára hozzáférhető, legnagyobb orvosi és egészségügyi bibliográfiai adatbázist, a PubMed[2]-et (Medline) használtuk. A témához keresőkifejezésként a következőt állítottuk össze: child[Ti] OR children[Ti] AND obesity[Ti] OR overweight[Ti] AND "public health"[MeSH Terms]. A logikai műveleti jelek ("OR" és "AND") mellett használt "[Ti]" alkalmazása a keresést a címekre korlátozta. A "public health"[MeSH Terms]" a népegészségügy kifejezéssel kapcsolatos kulcsszavakkal jellemzett cikkek kiválasztását jelenti. A keresés tehát a gyermekkori túlsúly és elhízás szavakat a címükben tartalmazó, illetve a népegészségügy témaköréhez tartozó közlemények kiválasztását tette lehetővé. A kereső kifejezésünk 18 147 cikket azonosított 2023. április 7-én. A közlemények száma 858-ra csökkent, amikor a keresést az áttekintésekre (*review*) és a megjelenést a 2013 elejétől 2023 végéig terjedő időszakra korlátoztuk. A keresés eredményét a PubMed-ből több formában is exportáltuk (PubMed saját bibliográfiai formátum, csak PMID tartalmazó kivonat, absztraktokat is tartalmazó txt file, illetve Excel importot könnyítő csv állományok). A különböző változatok a különböző szövegbányászati eszközök vagy módszerek számára szolgálnak ideális bementként (*input file*). Az így létrehozott szövegtest szolgált kiindulási pontként az elemzésekhez.[3]

2. A szövegtestbe került cikkek számának időbeli elemzése

A kereső kifejezéssel kinyert közlemények évenkénti számának időbeli alakulása arról tájékoztatja a kutatót, hogyan változik évről-évre a téma iránti érdeklődés a tudományos közösségben. A PubMed keresési eredményeket tartalmazó oldalának ábrájáról rögtön leolvasható éves bontásban[4] a publikációk száma. Az 1. ábrán megjelenített görbe ezt reprodukálja. A gyermekkori túlsúlyt és elhízást áttekintő cikkek száma az elmúlt tíz évben folyamatosan — bár csillapodó ütemben —

[2] https://pubmed.ncbi.nlm.nih.gov/
[3] Megjegyezzük, hogy a vizsgált szövegtestek mindegyike időrendi sorrendben tartalmazza a megjelent cikkek adatait. Lehetséges olyan corpust is készíteni, ami a bibliográfiai adatokat cikkenként külön-külön állományként kezeli. Ebben az esetben a corpus maga több dokumentumból áll, ezek lehetnek akár a teljes cikkek is. Jelen vizsgálatunk során az eszközeink kapacitása miatt az egyszerűbb megoldást választottuk, egyetlen file-t tartalmazó, egyesített corpusokat használtuk.
[4] Finomabb bontást is megtehetünk, ismételt, például havi vagy heti időszakok szűkített, ismételt keresésekkel, ennek azonban a trend megítélésében vélhetően nincs jelentősége.

növekszik, így megállapítható, hogy az elmúlt évtizedben a téma élénken foglalkoztatta a kutatói közösséget, és az érdeklődés jelenleg is erősödik. Az áttekintő (review) cikkek évente száz körüli száma összesen 10 000/éves számosságú, ebben a témában megjelenő publikációra utal.



1. ábra: A szövegtestbe került cikkek számának változása a vizsgált időszakban (Forrás: saját szerkesztés; a 2023-as érték az első negyedév alapján előrevetített cikkszám; a trendgörbe az EXCEL hatványfüggvény illesztéssel készült)

3. A leggyakoribb szakkifejezések elemzése

A szövegtestben előforduló leggyakoribb szakkifejezések feltalálása a kutatási fókusz(ok) azonosítására szolgál. A szövegtest természetesen tartalmaz nem szakkifejezéseket is, ezért az elemzéshez célszerű azok számát minimálisra csökkenteni. Bár az általunk használt VOYANT TOOLS[5] program a szövegtestből automatikusan kitörli a nem szakkifejezésnek vagy értelmezhetetlenek tartott szavakat, betűsorokat (ezeket tartalmazza az ún. *stopwords* lista), de mégis érdemes az alaphelyzeti *stopword*-listát kiegészíteni a szövegtestben még gyakran előforduló, de nem releváns szavakkal, betűsorokkal (Hendrigan, 2019). A példaként használt szövegtestből a VOYANT TOOLS program segítségével kinyert, "nyers" szakkifejezéslistából először készítettünk egy gyakorisági elemzést (*terms frequency*). [1. táblázat]. Ennek áttekintése után a *stopword*-listát kiegészítettük többek között az „â" és a „ci" betűsorokkal, illetve a *criteria*, *model* szavakkal.

[5] https://voyant-tools.org/

első 25			utolsó 25		
rangsor	szakkkifejezés	gyakoriság	rangsor	szakkkifejezés	gyakoriság
1	intervention(s)	1521	61	countries	125
2	risk	1058	62	reduce	124
3	obese	875	63	composition	123
4	body	813	64	population	123
5	health	619	65	mortality	121
6	prevalence	586	66	cardiovascular	120
7	adolescents	543	67	insulin	120
8	loss	484	68	circumference	119
9	physical	484	69	strategies	119
10	childhood	452	70	resistance	115
11	effect	418	71	results	115
12	activity	380	72	cholesterol	114
13	adults	357	73	supplementation	110
14	exercise	357	74	moderate	106
15	women	341	75	support	104
16	factors	339	76	electronic	103
17	diet	319	77	sleep	103
18	control	313	78	metformin	100
19	treatment	303	79	improve	98
20	fat	255	80	consumption	97
21	diabetes	245	81	nutrition	97
22	lifestyle	243	82	aerobic	96
23	individuals	236	83	cancer	96
24	criteria	231	84	family	95
25	dietary	225	85	behaviour	94

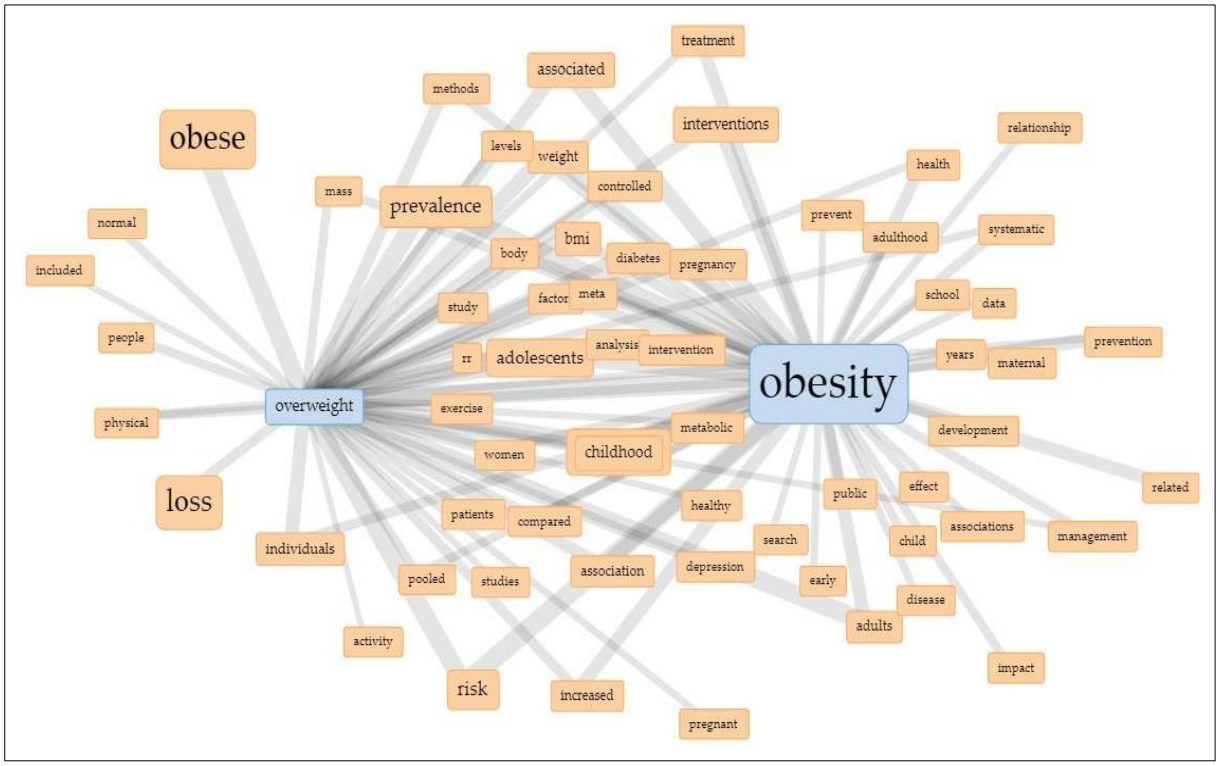
1. táblázat: A szövegtestben legalább 85 feletti, azaz 10%-os gyakorisággal előforduló szakszavak listájának első és utolsó 25 eleme (Forrás: saját szerkesztés; a teljes lista megtalálható a mellékletben)

A leggyakoribb szakkifejezések gyakorisága listája az általános szakmai gondolkodást tükrözi a gyermekkori túlsúly és elhízás területén: a vizsgálatunkba bevont áttekintések az állapot jelentette kockázatok visszaszorítására irányuló, a gyermekek, a serdülők és a nők körében megvalósítandó intervenciókkal, azaz a fizikai aktivitás és az étrend ellenőrzésével foglalkoznak leggyakrabban. A túlsúly és elhízás következményeként fellépő betegségek közül a gyakorisági listára került még a cukorbetegség, a szív- és érrendszeri betegségek, illetve a kapcsolódó tünetek mint a vérnyomásérték, vagy a koleszterin, valamint két gyógyszer is, az inzulin és a metformin.

4. Kollokációelemzés

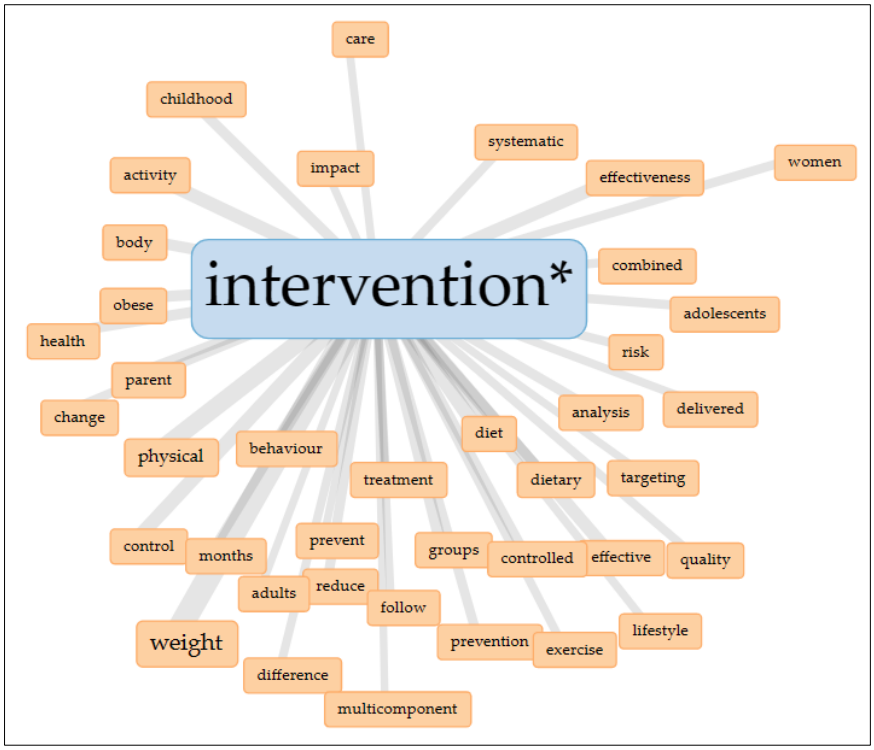
A kollokáció a szakkifejezések egymás közelében való előfordulását jellemzi, ezáltal a közleményekben előforduló lényeges fogalmi kapcsolatok kiemelésének egyik módja. Egy szövegtérben (ami akár több cikk alkotta szövegtest is lehet) a fogalmak közelségének mérésére számos módszer elérhető, az eredmények numerikusan és vizuális megjelenítéssel is megmutathatók. A VOYANT TOOLS kolloká-

ciós gráf olyan szakkifejezéshálót rajzol fel, amelynek egyes elemei az áttekintő közlemények szövegében egymás közelében fordulnak elő. A közelség értékeinek (*proximity*) kiszámítása ún. erő-irányított hálózati gráf (*force directed netwok*) technikával történik. Az erő-irányított hálózatban a gráf csomópontjai szakkifejezések, amelyek 2 (vagy több) dimenziós térbeli pontokként, az élek pedig e pontokat összekötő vonalakként jelennek meg. A VOYANT TOOLS a kollokációs hálózat két dimenzióban való optimális megjelenítéséhez ismétlődő számításokat végez úgy, hogy a csomópontok (az egyes szakkifejezések) és az élek hossza (a szakkifejezések egymás közelében való előfordulási gyakorisága) mintegy vonzó fizikai erőként szimulálódik. A szimuláció iteratív módon halad előre, a csomópontok és élek a rájuk ható (virtuális) erők hatására mozognak, amíg a rendszer el nem ér egy stabil állapotot. Ekkor a csomópontok és élek úgy helyezkednek el, hogy a hálózati struktúra megjelenítése optimalizált: a szoros kapcsolatban álló csomópontok csoportjai közel helyezkednek el egymáshoz, a kevesebb kapcsolattal rendelkező csomópontok pedig távolabb jelennek meg. [2. ábra]



2. ábra: A túlsúly és elhízás kifejezések kollokációs hálózata. (Forrás: saját szerkesztés; a kulcsszavak kék, a kapcsolódó szakkifejezések téglaszínű háttéren, az összekötő vonalak rövidsége az együttes előfordulással arányos, a címke mérete a szakkifejezés előfordulásának gyakoriságára utal)

A 2. ábrán szembetűnő, hogy a túlsúllyal és elhízással foglalkozó áttekintő közlemények a legtöbb használt szakkifejezés terén jól átfedi egymást, vagyis kevés olyan szakkifejezés van, amit inkább csak az egyik, vagy a másik terület kutatásában lenne gyakori. Érdekes továbbá, hogy a például a szociális kapcsolatok (*relationship*) kifejezés az elhízásnál gyakran felmerül, de a túlsúlynál kevésbé kutatott téma, míg a testmozgás (*physical*) például inkább a túlsúlynál gyakran kutatott terület. A következőkben azt vizsgáljuk, hogy az egyik leggyakoribb szakkifejezéshez, az intervencióhoz milyen kollokációs gráf tartozik, vagyis milyen fogalmak fordulnak elő az áttekintésekben az *intervention* szakkifejezés környezetében:



3. ábra: Az intervenció szakkifejezés kollokációs hálózata. (Forrás: saját szerkesztés; a kapcsolódó szakkifejezések téglaszínű háttéren, az összekötő vonalak rövidsége az együttes előfordulással arányos, a címke mérete a szakkifejezés előfordulásának gyakoriságára utal)

A 3. ábrán jól látszik, hogy a vizsgált szakkifejezéshez közel találjuk a jól ismert intervenciós (beavatkozási) területeket: viselkedés, terápia, étrend, testmozgás, a szülők szerepe stb., míg távolabb kerültek a beavatkozások egyes jellemzői, mint például a prevenció, életstílus, vagy a többkomponensű és a minőség.

5. Korrelációelemzés

Ez az elemzési technika a kutató által kiválasztott szakkifejezés-pároknak a szövegtestekben mért, egymáshoz viszonyított gyakoriságváltozását mutatja. Esetünkben, mivel az általunk vizsgált szövegtest egyetlen állomány, annak tíz egymást követő azonos méretű szegmensében Pearsons-korreláció kiszámításával történt meg az összehasonlítás. Egy szakkifejezés-pár közötti korreláció +1 és -1 közötti értéke abszolút értékben akkor magas, ha a szakkifejezések gyakorisága párhuzamosan, vagy éppen ellentétesen változik. A korrelációs együttható értéke +1, ha a gyakoriságok azonosan, és -1, ha teljesen ellentétesen változnak, és 0, ha a két gyakoriság egymástól függetlenül változik. Az elemzési technika bemutatásához a túlsúly és az elhízás szakkifejezésre végeztünk korrelációelemzést. A 2. táblázatban érdemes megfigyelni, hogy a két központi szakkifejezésünkkel (*obesity* és *overweight*) más és más szakkifejezések korrelálnak a legerősebb mértékben. Az is megfigyelhető, hogy az *obesity* szakkifejezéssel kapcsolatos korrelációs értékek trendszerűen nagyobbak, mint az *overweight*-tel kapcsolatos értékek. Ez arra utal, hogy súlyosabb állapotot (*obesity*) leíró tanulmányok jobban "összeszedettek", nagyobb a belső tartalmi konzisztencia a vizsgált szövegtestben. Érdekes megfigyelni a különös szakkifejezések felbukkanását a magas korrelációs értékek között, mint például a *minerals* az *obesity* oszlopban. Ilyen esetben a "furcsa" eredmény értelmezéséhez érdemes kikeresni azokat a cikkeket, amelyekben ez a kifejezés előfordult.

célszakkifejezés	szakkifejezések	korreláció
obesity	causing	0,89
	disabilities	0,86
	disproportionately	0,83
	children	0,81
	likelihood	0,80
	airways	0,80
	enormous	0,80
	minerals	0,80
	children's	0,79
	developmental	0,79
	heterogeneous	0,79
	disparity	0,79
	need	0,78
	lumbar	0,77
	force	0,77
	contents	0,77
	intestinal	0,77
	migrant	0,77
	extended	0,76
	decades	0,76
	culturally	0,73
	environment	0,73
	chosen	0,72
	intellectual	0,72
	institutions	0,71

célszakkifejezés	szakkifejezések	korreláció
overweight	obese	0,84
	combination	0,77
	associations	0,71
	million	0,68
	exposure	0,67
	gestational	0,67
	mass	0,67
	animal	0,67
	advanced	0,65
	intention	0,63
	drinks	0,63
	geographical	0,62
	intrauterine	0,62
	lipoprotein	0,62
	model	0,61
	mechanism	0,61
	habits	0,61
	lipid	0,60
	cases	0,60
	january	0,60
	ldl	0,60
	hdl	0,60
	excessive	0,60
	effectively	0,59
	medicus	0,59

2. táblázat: A 25 legmagasabb korrelációt mutató szakkifejezés-pár az obesity és az overweight szakkifejezésekre (Forrás: saját szerkesztés; a vonatkozó ökölszabály szerint, ha a korrelációs együttható akár pozitív, akár negatív tartományban eléri a 0,3 tizedet, akkor gyenge, ha a 0,5-öt, akkor közepes, míg 0,7 felett erős az összefüggésre utal)

6. Kontextuselemzés

A szakkifejezések kontextusának vizsgálata egyes szakkifejezések szerepének megértését segíti, emiatt a "mélyfúrással" feltárt kontextus a szövegbányászat nagy kincse. A VOYANT TOLS Contexts eszköz a kiválasztott szakkifejezés minden egyes előfordulását mutatja a környező (megelőző, illetve követő) szöveggel (a kontextussal) együtt. Hasznos lehet annak alaposabb tanulmányozásához, hogy a kifejezéseket hogyan használják a vizsgált szövegtestben. A következő elemzésben az intervenció szakkifejezésnek a szövegtestben előforduló fogalmi környezetét vizsgáltuk, elemezve a szakkifejezést megelőző, illetve a követő szavakból álló, a közlési sorrendben a vizsgált tíz év alatt az első száz és utolsó száz kontextust. A 3. táblázat első oszlopa a kontextusnak a szövegtestben elfoglalt pozícióját mutatja (tokenindex). Esetünkben ez, mivel az absztraktok időben egymást követően kerültek a szövegtestbe, ez egyúttal időbeni pozíciót is jelent. A 4. táblázatban a vizsgált kontextusok színkódolásának magyarázatát mutatjuk be. Színkódokkal láttuk el ugyanis a kontextustípusokat, vagyis a hasonló intervenciókat. Az összesen közel 900 absztraktból azonosított első száz kontextus a szemlézett tíz év első évére, míg az utolsó száz kontextus az utolsó évre esik. Így az is feltárható, hogy milyen újabb intervenciók jelentek meg a vizsgált tízéves periódus utolsó éveiben.

Pozíció	megelőző szavak	vizsgált szakkifejezés	követő szavak
	az időben első száz kontextus		
275	of life should lead to	interventions	focused on these groups. 2013
559	of 5-8.5 was after	intervention	. After months, a of 3
624	a valuable adjunct to lifestyle	intervention	, which often achieves only limited
1429	them in maintaining healthy lifestyles.	Intervention	options, both for caregivers and
1572	There is a need for	interventions	that consider common mechanisms for
2218	105 194 Systematic of nutritional	interventions	indicate limited efficacy in reducing
2244	effective components within nutrition-related	interventions	involving children to 11 years
2843	the effect of screen time	interventions	on obesity in children and
2905	this meta-analysis. The that	interventions	targeting screen time are effective
2953	was no difference between the	intervention	and control groups in body
2974	is no that screen time	interventions	alone can obesity risk in
3291	assessment and provide opportunities for	interventions	. To systematically the existing literature
4026	modest effect sizes for obesity	interventions	, the aim of the present
4595	recent meta-analysis that exercise	interventions	are effective for promoting weight
4726	controlled trial; (ii) structured exercise	intervention	, alone or combined with other
4732	alone or combined with other	intervention	components; (iii) control no structured
5933	OBJECTIVES: To the effectiveness of	interventions	to support the initiation or
5985	and quasi-RCTs that compared	interventions	to support the initiation and
6001	who are overweight or obese.	interventions	social support, education, physical support
6012	or any combination of these.	interventions	were compared either with each
	az időben utolsó száz kontextus		
187547	of wearable and smartphone-based	interventions	to promote physical activity and
187613	SMD) for the comparison between	intervention	control in steps per (SMD
187661	There were no between the	intervention	and control groups for systolic
187684	that wearable and smartphone-based	interventions	are effective strategies in promoting
187729	To the efficacy of drug	interventions	for the treatment of obesity
187782	We controlled (RCTs) of pharmacological	interventions	for treating obesity (licensed and
187813	minimum of three months' pharmacological	intervention	and six months' follow-up
187823	up from baseline. We excluded	interventions	that specifically dealt with the
187922	the trials, were to drug	intervention	and to comparator groups (91
187960	RCTs. The length of the	intervention	ranged from to weeks, and
188070	trials), participants' views of the	intervention	(not reported), morbidity associated with
188077	reported), morbidity associated with the	intervention	(measured in one orlistat trial
188089	reporting more gallstones following the	intervention	; very certainty evidence), all-cause
188101	one suicide in the orlistat	intervention	group; certainty evidence) and socioeconomic
188109	evidence) and socioeconomic (not reported).	Intervention	comparator for difference (MD) in
188151	BMI in favour of the	intervention	comparator for change in weight
188194	weight in favour of the	intervention	serious adverse events: 24/878
188203	24/878 (2.7%) in the	intervention	groups 8/469 (1.7%) in
188228	52/1043 (5.0%) in the	intervention	groups 17/621 (2.7%) in
188312	a series of associated on	interventions	for obese children and adolescents

3. táblázat: Az intervention szakkifejezés első és utolsó száz kontextusából 20 példaként bemutatott kontextus (Forrás: saját szerkesztés; a színezés magyarázata a 4. táblázatban található)

a kontextustípusok színkódjai			
az időben első száz kontextus		**az időben utolsó száz kontextus**	
	combination more interventions		personal devices (wearables etc)
	lifestyle, behaviour		family-based
	nutrition, diet		nurse, nurse-led
	screen time		
	support		
	breastfeeding, infant feeding		
	school		
	environment		
	medical		
	parent - child		
	policy		
	training, PA		
	pregnancy		

4. táblázat: Az intervention szakkifejezés kontextuselemzéséhez felhasznált színkódok (Forrás: saját szerkesztés; az eltérő színek a kontextusok különböző típusait jelölik; az összes színkód nem jelenik meg a 3. táblázatban, mert ott csak az összesen vizsgált 100 első és 100 utolsó kontextusból csak az első 20-20 szerepel)

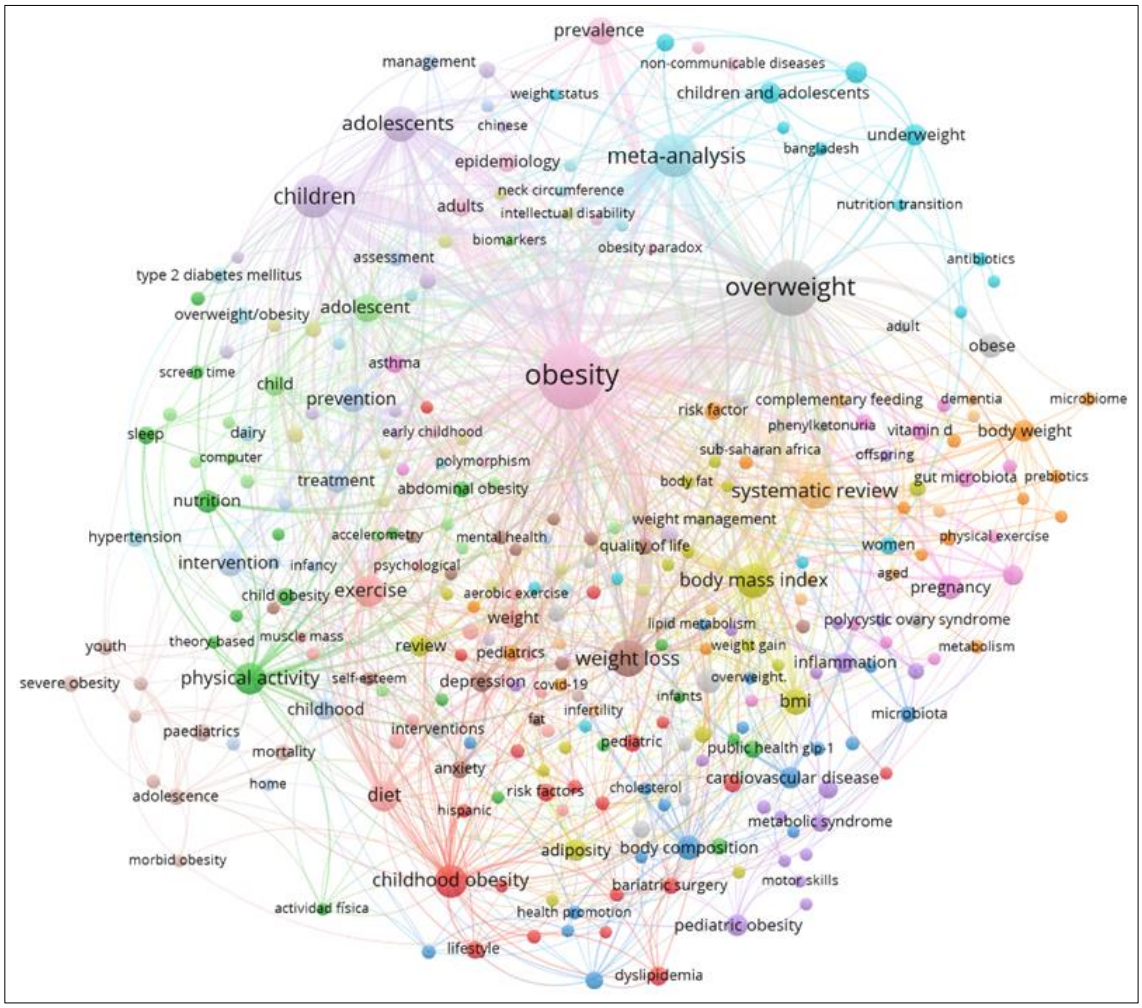
A színkódolás könnyen felismerhetővé teszi a kontextus-típusokat, sőt azok időbeli változása is nyomon követhető. Az első százból 53, míg a második százból 55 esetben az intervenció típusára találunk utalást a közvetlen (legfeljebb 5 szavas) kontextusokban. Valószínűsíthető azonban, hogy ha ugyanezt 10 szavas beállítással megismételnénk, akkor várhatóan több típus azonosítása is sikerülne. Az elemzés eredménye rávilágít a „medikalizáció" súlyának növekedésére: 6 versus 12 előfordulás az idő előrehaladtával, és azonosított új vagy újabban megjelenő intervenciókat is: *personal devices (wearables etc), family-based, parent child, nurse, nurse-led*.

7. Szakkifejezések hálózatának elemzése

Az előzőekben ismertetett szakkifejezés-előfordulási gyakoriság és fogalmi környezet kollokációs és korrelációs elemzései mellett egy-egy szakterület irodalmának hálózati vizsgálata további, más jellegű megértést is lehetővé tesz. A következőben ismertetésre kerülő szakkifejezések hálózati térképének csomópontjai (*nodes*) a szakkifejezések, míg az élek (*edges*) a köztük lévő előfordulási-közelségi kapcsolatot jelenítik meg, ahol a közelséget az együttes előfordulási gyakorisággal jellemezzük. Ezek a hálózati elemek egy fogalomteret feszítenek ki. A VOSviewer[6] eszköz segítségével a bibliográfiai adatbázisoknak a közleményekre vonatkozó, ún. meta-adatai alapján az ilyen hálózatok felrajzolhatók (van Eck & Waltman, 2022). Igy valójában egy adott terület tudományos publikációinak egyfajta hálózata konstruálható, a publikációkat jellemző meta-adatok, azaz a (1) folyóiratok, (2) kutatók, (3) kutatási szervezetek, (4) országok, (5) kulcsszavak, illetve szakkifejezések mentén. (A VosViewerben beállítható, hogy csak a szerzők által megadott kulcsszavak, vagy a kiadói/szerkesztői indexszavak, vagy mindez együtt képezze a vizsgálat tárgyát, a továbbiakban ezeket együtt kulcsszavaknak nevezzük). Hangsúlyozni kell, hogy

[6] https://www.vosviewer.com/

ez az eszköz nem a közlemények szövegeit, összefoglalóit, hanem a közlemények meta-adatait elemzi. Ezek között a szakkifejezések (amelyek egy vagy több kulcsszóból állhatnak) úgynevezett szemantikai, azaz tartalmi jelentéssel bíró meta-adatok. Amikor a VOSVIEWER ezekkel képez és elemez hálózatot, tulajdonképpen meta-szövegbányászat történik. A 4. ábrán látható szakkifejezés hálózati térkép elkészítéséhez ugyanazt a kiinduló PubMed keresési eredményt használtuk, mint a korábban bemutatott VOYANT TOOLS segítségével készített elemzéseknél. A szövegbányászat tárgya azonban itt nem a címek vagy absztraktok szak-kifejezései, hanem a cikkek kulcsszavai voltak. A kulcsszavakból álló hálózati térképen az eltérő színek alhálózatokat, ún. klasztereket jelölnek.



4. ábra: Kulcsszavak hálózati térképe (Forrás: saját szerkesztés; különböző színek a klaszterezési algoritmussal talált szakkifejezés-alhálózatokat, klasztereket jelölik, a kulcsszavak betűmérete a gyakoriságukkal, az összekötő vonalak vastagsága az együttes előfordulással arányos)

A kulcsszavak hálózati térképén megfigyelhetjük például, hogy a *children* kulcsszóhoz közel található a *type 2 diabetes mellitus*, míg az *overweight* és az *obesity* között az *obesity paradox*. A gyakran együtt előforduló kulcsszavak könnyen beazonosíthatóvá teszik a közlemények témáját. Az 5. táblázatban az ábrázolt kilenc klaszter[7] 10 leggyakoribb kulcsszavai láthatók.

[7] Az alkalmazott VOSVIEWER algoritmus összesen 10 klasztert talált, azonban a legkisebb "klaszter" már csak egyetlen cikket tartalmazott, így azt nem láttuk érdemesnek bemutatni.

1. klaszter		2. klaszter		3. klaszter		4. klaszter		5. klaszter		6. klaszter		7. klaszter		8. klaszter		9. klaszter	
obese	45	exercise	125	body composition	41	prevention	77	pregnancy	48	adolescent	86	hypertension	24	prebiotics	18	nutrition	37
underweight	35	physical activity	122	pediatric obesity	33	treatment	49	probiotics	33	asthma	23	adolescence	19	metabolism	15	mental health	13
type 2 diabetes	36	diet	114	dyslipidemia	29	epidemiology	20	vitamin d	25	internet	17	severe obesity	18	microbiome	15	stress	11
inflammation	25	childhood obesity	89	cardiovascular disease	27	assessment	20	metformin	17	dietary intake	16	mortality	17	metabolic disease	13	self-esteem	10
adults	32	intervention	63	metabolic syndrome	26	management	20	lifestyle interventions	14	computer	14	youth	16	supplementation	11	public health	8
children and adolescents	26	adiposity	38	network meta-analysis	25	sleep duration	17	gestational diabetes mellitus	14	healthy lifestyle	12	type 2 diabetes mellitus	15	antibiotics	9	obesity prevention	8
polycystic ovary syndrome	17	childhood	33	blood pressure	23	waist circumference	15	physical exercise	11	web	12	paediatrics	14	weight change	9	iron deficiency	5
women	17	weight	33	randomized controlled trials	18	sugar-sweetened beverages	14	gestational weight gain	9	biomarkers	11	dairy	13	chronic diseases	8	rct randomised controlled trial	4
gut microbiota	18	lifestyle	22	pediatric	18	risk factor	13	c-reactive protein	8	randomised controlled trials	11	nafld	13	dementia	8	dementia	8
trends	17	interventions	18	weight management	17	eating disorders	12	cytokines	7	nurses	9	social determinants	13	older people	8	older people	8

5. táblázat: A 10 leggyakoribb kulcsszavak listái a VOSVIEWER eszközzel előállított klaszterekben (Forrás: saját szerkesztés; a számok a teljes kapcsolati erősség értékét mutatja, ami jelzi, hogy az adott kulcsszó hány publikációban hányszorosan hivatkozott)

A táblázat értelmezéséhez vizsgáljuk meg például az *exercise* kulcsszót a 2. klaszterben! A 125-ös magas érték egyfajta erős fogalmi beágyazottságot jelent, hiszen az megfelelhet annak, hogy 5 cikk további 25 olyan cikket idéz, amelyekben ugyanez a kulcsszó előfordul (5*25=125). A legtöbb klaszter lényege jól megragadható a legmagasabb kapcsolati erősségek mentén, így az 1. klaszter például a konkrét betegségekkel kapcsolatos túlsúlyt, illetve elhízást vizsgáló cikkek csoportja (pl. 2. típusú diabétesz, gyulladások, policisztás ovárium stb.). A 2. klaszter az étrendet, a testmozgást mint intervenciót említő cikkek halmaza, míg a 3. klaszter a kóros állapotok (magas vérnyomás, diszlipidémia stb.) és az elhízás, túlsúly vizsgálata. De nem minden klaszter azonosításához elég a 10 legmagasabb kapcsolati erősség vizsgálata, ezért a mellékletben közöljük a teljes klaszter szakkifejezés listát.

MEGBESZÉLÉS

A diszkusszió első kérdése az lehet, hogy a szövegbányászati módszer hogyan fejlődik, honnan, hová jutottunk? A fentiekben bemutattunk jónéhány szövegbányászati módszert, az egyszerűbbektől a bonyolultabbak felé, de még így is csak a lehetséges elemzési módszerek töredékét tudtuk felvillantani. Ezek a nagyon egyszerű módszerek a szövegtest, illetve a kulcsszó mint adat alapvető jellemzőit mutatják be. A szövegbányászat kezdeti évtizedét (2005–2015) jól mutatja be Alison O'Mara-Eves és munkatársai (2015) cikke, amelyből idézzük a számunkra releváns megállapítását: "Érdemes megjegyezni, hogy a szisztematikus áttekintés során a cikkszűrést támogató szövegbányászati módszerek a szakbírálók eredeti munkafolyamatát követték, azaz a címek, az összefoglalók és a kulcsszavak használatával végzik a szűrést. Vannak itt

... megfontolandó kérdések ... mennyire jól képes a cím, az összefoglaló és a kulcsszavak mint meta-adatok kielégíteni egy összetett információs igényt? Ami például az absztraktok használatát illeti annak megállapítására, hogy összességében milyen állításokról van szó, Blake azt találta, hogy az orvosbiológiában a teljes szövegű cikkekben szereplő összes tudományos állítás kevesebb, mint 8%-a volt megtalálható a cikkek kivonatában." Az eltelt közel két évtized alatt a szövegbányászat sokat fejlődött. Ennek megmutatására idézzük Dandan Tao és munkatársai (2020) cikkéből az 5. ábra, amely jól foglalja össze a 2020-as évek fejlődő technikáit.

5. ábra: A szövegbányászat alapvető technikái és azok fejlődése (Forrás: Tao et al., 2020; magyarra fordítva)

A diszkusszió második kérdése lehet, hogy kínál-e az információtudomány a szövegbányászaton túl más eszközöket is a nagy mennyiségű szakirodalom feldolgozására. Nem meglepő módon a válasz igen, a manapság nemcsak tudományos körökben sokat említett mesterséges intelligencia (MI) kutatások terén találunk ilyen eszközöket. Ebben a cikkben erre nincs mód részletesen kitérni, de talán érdemes idézni Chen és munkatársai (2019) felsorolását, amely az elmúlt 5–10 év fejleményeit jellemzi: "Számos ígéretes eszköz és módszer van, például a "véletlen erdő" (random forest), a "támogatási vektorgépek", (support vector machines, SVM), a konvencionális neurális hálózatok, melyek a betegségek felismerésében segítenek. A "szemantikus web" eszköztára, ontológiák és téma-modellezés (topic modelling) a klinikai vagy orvosbiológiai szövegbányászatot támogatja, a mesterséges intelligencia neurális hálózatai eszközként és a logisztikus regresszió módszerként az előrejelzésekhez, vagy a konvolúciós neurális hálózatok (CNN) és ugyancsak az SVM-ek az epidemiológiai monitorozáshoz és morbiditási, klasszifikációs vizsgálatokhoz alkalmasak." A cikk szisztematikusan összefoglalja a gépi és

kognitív intelligencia eszközök használatát az emberi egészség szolgálatában. Meg kell jegyeznünk, hogy a népegészségügyi kutató számára személyes eszközként ezek a technikák még nehezen elérhetők, részben speciális MI szaktudást, részben megfelelő kapacitású számítástechnikai eszközparkot is igényelnek. Ezzel szemben az egyszerű szövegbányászati eszközök a szokásos eszközeinken futtathatók és az egyszerűbb vizsgálatokhoz nem kell speciális szakértelem.

KÖVETKEZTETÉSEK

A szemléltetés céljából kiválasztott gyermekkori túlsúly és elhízás témakör szövegbányászattal végzett elemzéseinek eredményei természetesen további értelmezési feladatot jelentenek az ezzel foglalkozó szakemberek számára. Írásunkban csupán lehetőségeket, elemzési irányokat szándékoztunk megmutatni. Ami a felszínes, bemutató jellegű vizsgálatunk eredményeiből elsőként levonható, hogy a túlsúly és az elhízás nem egymás témaköre, csupán részben átfedik egymást, miközben jelentősebb eltérések is azonosíthatók. Levonható továbbá az a biztató következtetés is, hogy a szakirodalom középpontjában e két egészségprobléma megoldására irányuló, különféle beavatkozások állnak. Ezek az intervenciók időben változnak, részben kedvező módon gazdagodnak, másrészt kedvezőtlen módon haladnak a (talán túlzott) medikalizáció felé, azaz orvosi eszközökkel keresnek megoldást egy nem minden tekintetben orvosi problémára. A bemutatott eszközök (VOYANT TOOLS és VOSVIEWER) csupán példák, a nyilvános digitális tudományos térben (open science, public domain) számos olyan eszköz létezik, amelyek intuitív grafikus felhasználói felületei segítségével a nem szövegbányász kutató is el tudja végezni az egyszerűbb, de informatív vizsgálatokat. Különösen ígéretes terület a hálózatos medicina (network medicine) felé átjárást adó fogalmi hálózatok természetének vizsgálata. Ki kell emelni azt is, hogy közel vagyunk a mesterséges intelligencia rutinszerű használatához is, amikor a kutató még inkább a saját feladatára koncentrálhat és a rutinszerű szellemi feladatokat algoritmusokra bízhatja a jövőben (pl. áttekintő közlemények előszűrése).

Végül meg kell állapítsuk, hogy a szakcikkek összessége közvetlenül kutatható adattá vált. A továbbiakban érdemes figyelemmel követni e közvetlen kutathatóság lehetőségeinek gazdagodását, bővülését.

HIVATKOZÁSOK

Bartholomew, M. (2002) James Lind's Treatise of the Scurvy (1753). Postgraduate Medical Journal, 78, 695–696. http://dx.doi.org/10.1136/pmj.78.925.695

Sur, R. L., & Dahm, P. (2011). History of evidence-based medicine. Indian journal of urology : IJU : journal of the Urological Society of India, 27(4), 487–489. https://doi.org/10.4103/0970-1591.91438

Balkányi, L., Lukács, L., & Cornet, R. (2021). Investigating the Scientific 'Infodemic' Phenomenon Related to the COVID-19 Pandemic. Yearbook of medical informatics, 30(1), 245–256. https://doi.org/10.1055/s-0041-1726483

Hendrigan, H. (2019). Mixing Digital Humanities and Applied Science Librarianship: Using Voyant Tools to Reveal Word Patterns in Faculty Research. Issues in Science and Technology Librarianship, (91). https://doi.org/10.29173/istl3

van Eck, N., J. & Waltman L. (2022) VOSviewer Manual. Manual for VOSviewer version 1.6.18. Leiden Universitate, CWTS Meaningful Metrics. https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.18.pdf

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews, 4(1), 5. https://doi.org/10.1186/2046-4053-4-5

Tao, D., Yang, P., & Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. Comprehensive reviews in food science and food safety, 19(2), 875–894. https://doi.org/10.1111/1541-4337.12540

Chen, X., Cheng, G., Wang, F. L., Tao, X., Xie, H., & Xu, L. (2022). Machine and cognitive intelligence for human health: systematic review. Brain informatics, 9(1), 5. https://doi.org/10.1186/s40708-022-00153-9

INFORMÁCIÓK A SZERZŐKRŐL

Balkányi László
Egészségügyi Informatikai Kutató-Fejlesztő Központ (EIKFK), Pannon Egyetem, Veszprém
laszlo@balkanyi.com

Vitrai József
Széchenyi István Egyetem Egészség- és Sporttudományi Kar Preventív Egészségtudományi Tanszék, Győr
vitrai.jozsef@gmail.com

CIKKINFORMÁCIÓK

Beküldve: 2023. 04. 19.
Elfogadva: 2023. 04. 19.
Megjelentetve: 2023. 06. 09.