Vállalkozásfejlesztés a XXI. században 2022/1. kötet
Az üzleti szervezetek túlélési esélyei napjaink legújabb kihívásainak idején

Innovatív irodalomkutatás R programnyelv alapú szövegelemzéssel

Prof. Dr. Takács István

Egyetemi tanár, Óbudai Egyetem, Keleti Károly Gazdasági Kar
takacs.istvan@uni-obuda.hu

Absztrakt: Az irodalomkutatás a tudományos publikációk elengedhetetlen kelléke: hozzájárul az adott kutatás helye, jelentősége, újszerűsége azonosításához. A szakirodalmi források mélyebb összefüggéseinek feltárásának támogatására a szövegbányászat nyújt újszerű eszközt. A cikk két kutatás: egy szakfolyóirat 10 évfolyamában közreadott cikkek elemzése során, illetve egy nemzetközi kutatás egyik alprojektjében alkalmazott módszertanról számol be. A kutatásokban a szövegbányászat módszereinek felhasználásával egyrészt 2009 és 2018 között az Annals of the PAAAE folyóiratban (393 angol nyelvű teljes cikk), másrészt a CAB adatbázisban 2010-2014, illetve 2015-2019 közötti időszakra szereplő, 1055 kertészeti folyóiratban (25 gyümölcs és zöldség faj kapcsán) megjelent 9246 cikk absztraktjainak vizsgálata történt, mindkét esetben arra keresve a választ, hogy a kutatások fókusza hogyan változott. A kutatás során R programnyelven fejlesztett rutinokkal került vizsgáltra többek között a kifejezések előfordulási gyakoriságát, asszociációját, elkészült a kifejezéspárok hálója, a szerzők hálója, a kifejezések és cikkek klaszterezése is. A cikk számba veszi az eredmények vizuális interpretálásának lehetőségeit is, a kapott eredményekből választott példákkal illusztrálva.

Kulcsszavak: szövegbányászat, imitáció, kvalitatív módszerek

1 Bevezetés

A problémafelvetés és a konkrét kutatási kérdés megfogalmazása, az annak megválaszolását szolgáló célok kijelölése után a tudományos kutatás következő, alapozó fázisa a szakirodalmi kutatás. Annak során megtudható, hogy milyen ismeretek állnak már rendelkezésre, melyek a "trendi" témák, s nem utolsó sorban milyen kevéssé kutatott kérdések vannak, s a kutató kíváncsiságának tárgya melyik kategóriába tartozik. A szakirodalmi források szisztematikus értékelése, csoportosítása, a kutatási fókuszok változása hagyományos irodalomkutatási eszközökkel is elvégezhető, s gyakran találhatók strukturált (többnyire táblázatba is rendezett) értékelések, de az informatika, benne az adatbányászat eszközeinek (benne a szövegbányászat) fejlődése kiszélesítette a módszertani lehetőségeket, s az emelt szintű statisztikai eszközök használata lehetővé vált, sőt egyre inkább

követelmény lesz. És ez igaz lesz nem csak az alapozó szakirodalmi elemzéseknél, hanem a szövegeket kezelő kvalitatív kutatásoknál is (például mélyinterjúk kiértékelése).

Witten szerint a szövegbányászat vagy másként szövegadatbányászat elfogadott módszerré vált az írott anyag mélyebb tartalmának és összefüggéseinek feltárására (Witten Ian H., 2004). Az eszközt egyre gyakrabban alkalmazzák, egyre több tudományos cikke jelenik meg, amelyek a módszer alkalmazásával nyert eredményeket taglalják. Rendelkezésre állnak szövegelemző szoftverek (például az Atlas), az öntevékeny kutatók számára az R nyílt forráskódú programnyelv önkéntes fejlesztői egyre több megoldást kínálnak használatra (lásd például Witten Ian H., 2004, Williams G., 2016 és Zhao Y., 2013) a "laikus" használók számára.

Jelen tanulmány nem elsősorban a szövegbányászattal elért eredmények reprezentálását célozza, hanem az eredményekhez vezető útra, a kutatás módszertani kérdéseire fókuszál. Azt kívánja bemutatni, hogy a két kutatás során a szövegelemzés R alapú sajátfejlesztésű módszerei hogyan készültek a kutató kezei alatt, illetve a problémamegoldás fázisai során hogyan fejlődött az eszközkészlet. A kezdeti fázis – a szakirodalmi forrásokon alapuló – immitáció volt, majd az azt követte az adaptáció, valamint az eszközökben rejlő lehetőségek megismerése után az arra alapozott egyedi fejlesztések következtek. A tanulmány célja, hogy érzékeltesse, hogy a kutató hogyan tud saját fejlesztésű eszközökkel hozzájárulni a kutatása sikeréhez.

Az első kutatás egyfajta indoklását adta, hogy a lengyel agrárközgazdászok egyik, ha nem a legjelentősebb tudományos folyóirata, az Annals of the PAAAE (i.e. Polish Association of Agricultural and Agribusiness Ecconomists) (lengyel címe: Roczniki Naukowe), amelyben magyar kutatók is rendszeresen publikálnak. A folyóirat 2019-től csak angol nyelvű cikkek megjelenését tűzte ki célul, abból adódóan 2018. évvel a folyóirat egy korszaka lezárult. Az utolsó 15 évben, elsősorban magyar szerzők megjelenésével (de a lengyel és magyar szerzők mellett a vizsgált időszakban tíz további országból voltak szerzők), a folyóirat nemzetközi, de elsősorban is a CEE országok egy agrárközgazdasági folyóirata lett.

A kutatási kérdés az volt, hogy a folyóiratban leképeződnek-e a vizsgált időszak gazdasági-társadalmi változásai. A vizsgálat kezdőévének választott 2009-re öt év telt el a térség országainak Európai Unióhoz történt csatlakozástól. Ekkor az újonnan csatlakozott országokban az uniós támogatási programokkal kapcsolatos kutatások vannak napirenden, de ugyanakkor az éppen lezajlott pénzügyi válság is már megjelenik a cikkekben (Takács I., Baranyai Zs., 2009).

A régió sajátosságaiból is adódóan az egyik súlypont a vidéki térségek fejlesztése, annak részeként a turizmus és a vidék kapcsolata volt az agrárökonómusok kutatásaiban. Ezekkel kapcsolatos kutatások eredményei is megjelentek a folyóiratban: cikkek számolnak be például a turizmus foglalkoztatási kérdéseit (Balińska, A., 2009), a turizmus vidékfejlesztésben betöltött szerepét (Brelik A.,

2009), vagy például a vidéki területek és a városfejlesztés összefüggéseit (Staszewska, S., 2009) vizsgáló kutatásokról. Az Európai Unióban a regionális egyenlőtlenségek csökkentése alapvető célkitűzés, ugyanakkor a területek fejlettségében meglévő eltérések eltérő fejlesztéspolitikával mérsékelhetők. Ez a kérdékör is megjelenik 2009. év cikkeiben (Koreleski D., 2009) Természetesen 2018-ra nem tüntek el a korábban meghatározó témák, például a vidéki turizmus, de új kontextusok megjelenése tapasztalható (Wojcieszak, M., Jan, Z., 2018) ugyanakkor elindul a tudomány felkészülése az új költségvetési ciklus kérdéseire is (Wieliczko B., 2018).

A szövegelemzésen alapuló kutatás, a kifejezéseinek elemzésével, arra kereste a választ, hogy milyen változások azonosíthatók a vizsgált évtized során megjelent cikkekben.

A vizsgálat kutatási kérdése: a szövegbányászat eszközeivel kapott profil egyezik-e a folyóirat deklarált szakmai célkitűzéséből eredő profilképpel, illetve van-e kimutatható hatása annak, hogy a külföldi szerzők aránya is számottevő volt, mivel a különböző országokban eltérő publikációs kultúra, illetve a szakmai nyelv használatának lehetséges különbségei hatnak a cikk szókészletére, annak szűken vett szakmai tartalmán túl is. A cikkek teljes szövege képezte a vizsgálat tárgyát.

Az ebben a kutatásban kapott eredmények hozták az ötletet a Postharvest handling 4th edition szerkesztői számára, hogy a szövegbányászat módszereit alkalmazva kiválasztott zöltség és gyümölcs fajok betakarítás utáni kezelés és a zöldség-gyümölcs fogyasztás közötti ok-okozati hatások vizsgálati eredményeit közreadó publikácók megállapításait elemezze (Florkowski, W., J., Takács, I., 2022). A kiválasztott friss gyümölcsök és zöldségek egy főre jutó fogyasztása változásainak leírásával több ezer publikált tanulmány foglalkozik. A kutatás anyaga a tárgykörrel foglalkozó cikkek CAB adatbázisban található absztraktjai voltak. A szövegbányászati elemzés két időszakra vonatkozik, 2010-2014 és 2015-2019, amelyek megfelelnek a könyv második és harmadik kiadása közötti, illetve a harmadik kiadás óta (negyedik kiadás elkészítésének évéig terjedők) időszakoknak, amelyek jól körülhatárolható időszakokat hoznak létre, kezelhető, releváns számú, elemezhető tanulmányokkal.

A tanulmány célkitűzései:

1) a módszertan kialakítás folyamatának bemutatása, amley módszertan célja az Annals of the PAAAE folyóirat vizsgált tíz évének jellemzésére, a kutatási fókusz változásának azonosítására az időszakban megjelent 393 folyóiratcikkben a kifejezések gyakoriságának változásával;

2) a módszertan kialakítás folyamatának bemutatása, amley módszertan célja a CAB adatbázisban 2010-2014, illetve 2015-2019 közötti időszakra szereplő, 1055 kertészeti folyóiratban (25 gyümölcs és zöldség faj kapcsán) megjelent 9246 cikk absztraktjainak vizsgálata.

2 Anyag és módszer

A kutatás egyik adatforrását az Annals of the Polish Association of Agricultural and Agribusiness Economists (PAAAE) folyóirat számaiban 2009-2018 között (from Vol. 11 to Vol.20) angol nyelven közreadott cikkek adták. A cikkek a https://rnseria.com/resources/html/archives webcímen érhetők el pdf formátumban. A 10 évfolyamban összesen 393 angol nyelven közreadott cikk volt jelölve (1. táblázat), amelyből 2013-ban egy cikk, 2015-ben 24 cikk nem volt feltöltve a repozitóriumba, csak absztrakt szerepelt, így a továbbiakban 2013-ban 40, 2015-ben 8 cikk szövegelemzésére került sor, illetve a 2015. évi adatok egyes elemzésekben a későbbiekben figyelmen kívül lettek hagyva.

Cikkek eredete származási ország szerint	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Eredet szerint mind-összesen	
	db	db	db	db	db	db	db	db	db	db	db	%
Évenkénti összes	26	38	50	55	41	42	32	40	25	44	393	100.0
Lengyel	15	15	32	30	20	26	24	19	9	29	219	55.7
Lengyel és más				1		1	2	2		3	9	2.3
Magyar	9	17	16	22	18	14	4	17	13	10	140	35.6
Magyar és más									1		1	0.3
USA	1	1		1	2						5	1.3
USA és más			1				1	1	1		4	1.0
Egyéb országok	1	5	1	1	1	1	1	1	1	2	15	3.8

1. táblázat: Az angol nyelvű cikkek származási országainak gyakorisága a PAAE Annals 11-től 20. évfolyamáig

Forrás: saját szerkesztés

A második adatforrás a CAB (i.e. Commonwealth Agricultural Bureau) adatbázisban 2010-2014, illetve 2015-2019 közötti időszakra szereplő, 1055 kertészeti folyóiratban, 25 gyümölcs és zöldség faj kapcsán megjelent 9246 cikk absztraktja (2. táblázat) volt.

A szövegbányászat és a szövegek statisztikai elemzése R programnyelvvel történt (Williams, G., 2016) és (Zhao, Y., 2013): által közölt minta rutinok adaptálásával, valamint az R package-ok manualjainak (lásd https://cran.r-project.org/) felhasználásával RStudio integrált fejlesztőkörnyezetben.

A folyamat lépései a következők voltak:

1) Adat előkészítés – eredeti dokumentumok, adatállományok (.pdf, .csv, .doc, .docx, .rtf formátumú fájlok) beolvasása, corpus létrehozása;
2) Corpus előkészítése elemzéshez – nagy-kisbetű konverzió, szövegtisztítás (punctuation, numbers, whitespace, stopwords, stemming and completion);

3) Elemezhető szövegállományok: Text-Document-Matrix (Tdm) és Document-Text-Marix (Dtm) létrehozása;
4) Kifejezések gyakoriságának megállapítása és vizuális ábrázolása;
5) Választott kifejezések asszociációjának számítása;
6) Szófelhő készítés;
7) Kifejezések klaszterelemzése;
8) Cikkek klaszterelemzése;
9) Kifejezéspárok hálójának vizsgálata
10) Szerzők hálózatának vizsgálata
11) Kifejezések korrelációs ábráinak (Correlation plots) elkészítése.

Zöldség/gyümölcs	1975-2019	1985-2019	2010-2014	2015-2019	2010-2019
Apples	0	0	752	722	1474
Artichokes	0	0	22	9	31
Bananas	0	0	320	288	608
Blueberries	0	0	81	119	200
Broccoli	516	0	108	132	240
Cabbage	317	0	54	76	130
Cherries	0	0	149	272	421
Fresh-cut_vegetables	0	0	17	13	30
Grapefruit	0	0	41	55	96
Grapes	0	0	302	353	655
Iceberg_lettuce	0	106	27	21	48
Kale	0	0	13	19	32
Kiwi	0	0	129	158	287
Onions	0	0	119	111	230
Oranges	0	0	329	326	655
Peaches	0	0	334	332	666
Pears	0	0	283	353	636
Potatoes	0	0	287	337	624
Radishes	0	0	24	31	55
Romaine	0	51	19	20	39
Spinach	0	224	54	74	128
Strawberries	0	0	252	328	580
Sweet_onions	0	0	2	1	3
Tangerines	0	0	161	123	284
Tomatoes	0	0	495	599	1094
Mindösszeseb	833	381	4374	4872	9246

2. táblázat: A CAB adatbázisból letöltött cikkek megoszlása időszakok szerint
Forrás: saját szerkesztés

Megjegyzés: A folyamat gyakorlati kivitelezése első fázisában egy 283 tényleges utasítássort tartalmazó batch fájl felhasználásával, a második fázisban az újabb elemzések elvégzéséhez további mintegy 300 sort tartalmazó rutin létrehozásával történt. (A rutinok leginkább a "deszkamodell" fejlesztési fázisnak megfelelő állapotúak – részben az R programnyelv sajátosságaiból is következően –, de egy-egy elemzési fázis egybeszerkesztett utasítássorait teljes körűen tartalmazzák, további beavatkozás nélkül működőképesek, futtathatók.) Ezt a terjedelmi korlátok okán a tanulmány nem tartalmazza, ugyanakkor ez is egy újszerű eredmény, amely a későbbiekben hasonló célú vizsgálatokra felhasználható. A

cikk olvasásában eddig eljutók számára – az R programnyelv közösségi fejlesztésének megfelelően – megkeresésre rendelkezésre bocsátom.

A fejlesztéshez kapcsolódóan meg kell jegyezni, hogy minden R programnyelvre vonatkozó előtanulmány (és így ismeret) nélkül indult. Ebből adódóan meg kell említeni néhány alaptankönyvet, amelyek segítségre voltak az R programozási, adatstruktúrákra, adatkezelésre vonatkozó szabályrendszerének megértésében (lásd: Solymosi, N., 2005, Keresztúri, J. L., Antal, B., Illés, F., 2017).

Az első lépésekhez a szövegelemzési rutin minták Williams (2016) és Zhao (2013) cikkeiből származtak, de a második fázisban Silge, Robinson (2017) Text Mining with R című könyve mutatta meg (Jane Austin munkássága elemzése példáján) a strukturált adatbázisban végzett szövegelemzés módszertanát.

Az Eredmények fejezet további információkat tartalmaz az egyes vizsgálati fázisok, vizsgálati eszközök módszertani vonatkozásairól.

3 Eredmények

Az R programutasítások, rutinok hasznos futatásaira a leginkább alkalmas az RStudio integrált fejlesztőkörnyezet. A fejlesztők számára hasznos funkciók alkalmazásába a laikus fejlesztő is gyorsan beletanulhat, ugyanakkor a fejlesztés hatékonyságát, valamint a fejlesztés részeredményei (tesztelt, üzemképes rutinok), verziók megőrzését, a programok átláthatóságát a szövegszerkesztővel kombinált alkalmazás biztosította.

Az RStudio fejlesztőkörnyezetből adódóan lehetőség van arra, hogy az általános előkészítést – a futtatási környezet felépítését (szükséges könyvtárak telepítését) – általában az első futtatást megelőzően egyszer elégséges elvégezni, de mégis célszerű minden egybeszerkesztett, bemásolásra szánt programot ezen könyvtárak betöltésével kezdeni, hogy a tesztelt környezet mindenkor azonosan rendelkezésre álljon.

A következő fázis a munkakörnyezet kialakítása, az input és output munkakönytárak strukturálása. Az input maga az adatbázis, benne az eredeti szövegforrásokkal (nyers adatok, amelyek még nem alkalmasak szövegelemzésre), illetve az adattisztítás, és konverziók után létrejövő saját szövegadatbázis. Ha rendelkezésre áll a tisztított szöveg, akkor kezdhető a szövegbányászat maga.

3.1 Munkafájl létrehozása

A munkafájl létrehozásával kapcsolatos műveletek a következők: az eredeti dokumentumok (.pdf, .csv, .doc, .docx, .rtf formátumú fájlok) beolvasása,

amelyekre specifikus utasítások állnak rendelkezésre, és akár kötegelten is elvégezhető a művelet ciklusba szervezve a beolvasást; azt követi a szövegkonverzió és szövegtisztítás: a központozás, a számok, az üres szóközök, az úgynevezett 'stopwords' (pl. kötő- és töltelékszavak, elemzési szempontból önálló tartalmat nem hordozó kifejezések) eltávolítása, kifejezések egységesítése (ragozott, magyar nyelvű szövegek esetén igekötős kifejezések egységes szóalakra cserélése) elvégzése után egy már elemzésre alkalmas munkafájl készül, és a továbbiakban azzal készülnek a szövegelemzések. A felsorolt adatelőkészítési feladatok egy részére utasítások állnak rendelkezésre, de a 'stopwords'-ök esetén saját szólista használatára is van mód, illetve a kifejezések egységesítése két fázisban történik: elsőben – az úgynevezett korpusz (szövegtest) elkészülte után – a kifejezések gyakoriságát is tartalmazó lista készül (a szövegben lévő kifejezések fontosságának megítélésére), amelyet – a második fázis részeként – részletesen át kell nézni, s a szótövek szerinti konverziós táblát el kell készíteni. (3. táblázat) (Ez magyar nyelvű szövegek esetén eléggé munkaigényes.)

word	lexicon
amounts	amount
analysed	analyse
analyses	analyse
analyzed	analyse
antioxidants	antioxidant
apples	apple
applications	application
areas	area
atmospheres	atmosphere

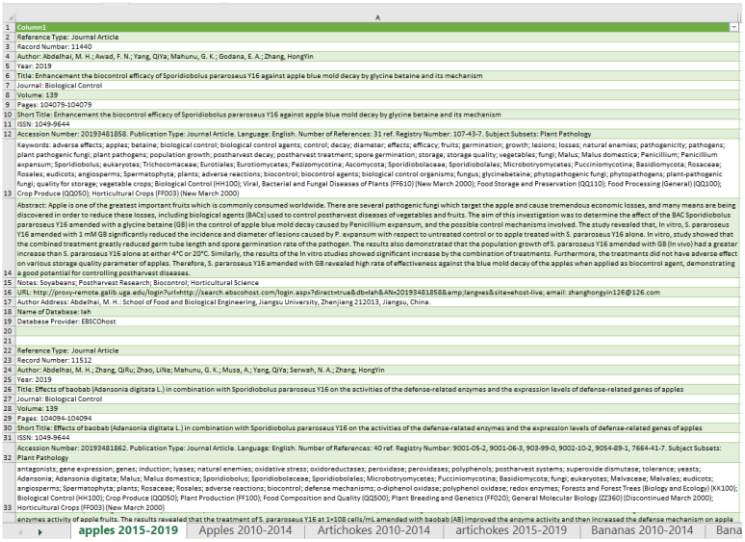
3. táblázat: Részlet a kifejezések egységesítéséhez készült segédtáblázatból
Forrás: saját szerkesztés

Az utasítás a szövegkorpuszban lévő szavak közül a 'word' oszlopban található kifejezéseket a 'lexicon' oszlopban lévő kifejezésre cseréli.

A következő műveletek ezt a forrásfájlt fogják használni a szövegelemzési műveletek elvégzéséhez.

Az előkészítő munka fontos része, hogy az elemzési igények figyelembe vevő adatstruktúra kerüljön kialakításra. Az összetettebb feladatot a második kutatási projekt adatállományának rendezése, a későbbi munkafájl előkészítése volt, így mintaként ez kerül bemutatásra. Az 1. ábrán látható a CAB adatbázisból letöltött adatok strukturája. Ezek egy része felesleges a későbbi elemzéshez, ezért kihagyásra kerültek. Az egyesített adatállományban ezért csak a gyümölcs/zöldség faj és a cikkek időszaka, a cikk azonosító, szerzők, megjelenés éve, cikk címe, folyóirat címe, a cikk sorszáma az adott faj esetén, valamint az absztrakt (az ábrán már konverziók és tisztítás utáni) szövege. (2. ábra) Az adatelőkészítést az Excelben viszonylag egyszerűen meg lehetett oldani (az adatok tömegessége okán

– a szűrést segítendő – függvények alkalmazásával a konvertált adatbázisba kerülő cellák értékeinek összerendezése, s új táblázatba mentése elérésére). A későbbi műveletekhez az R-ben kezelt input adatok .csv fájlokba kerültek mentésre.

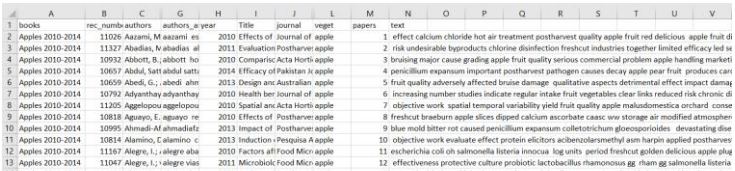


	A
1	Column1
2	Reference Type: Journal Article
3	Record Number: 11440
4	Author: Abdelhai, M. H.; Awad, F. N.; Yang, QiYa; Mahunu, G. K.; Godana, E. A.; Zhang, HongYin
5	Year: 2019
6	Title: Enhancement the biocontrol efficacy of Sporidiobolus pararoseus Y16 against apple blue mold decay by glycine betaine and its mechanism
7	Journal: Biological Control
8	Volume: 139
9	Pages: 104079-104079
10	Short Title: Enhancement the biocontrol efficacy of Sporidiobolus pararoseus Y16 against apple blue mold decay by glycine betaine and its mechanism
11	ISSN: 1049-9644
12	Accession Number: 20193481858. Publication Type: Journal Article. Language: English. Number of References: 31 ref. Registry Number: 107-43-7. Subject Subsets: Plant Pathology
	Keywords: adverse effects; apples; betaine; biological control; biological control agents; control; decay; diameter; effects; efficacy; fruits; germination; growth; lesions; losses; natural enemies; pathogenicity; pathogens; plant pathogenic fungi; plant pathogens; population growth; postharvest decay; postharvest treatment; spore germination; storage; storage quality; vegetables; fungi; Malus; Malus domestica; Penicillium; Penicillium expansum; Sporidiobolus; eukaryotes; Trichocomaceae; Eurotiales; Eurotiomycetes; Pezizomycotina; Ascomycota; Sporidiobolaceae; Sporidiobolales; Microbotryomycetes; Pucciniomycotina; Basidiomycota; Rosaceae; Rosales; eudicots; angiosperms; Spermatophyta; plants; adverse reactions; biocontrol; biocontrol agents; biological control organisms; fungus; glycinebetaine; phytopathogenic fungi; phytopathogens; plant-pathogenic fungi; quality for storage; vegetable crops; Biological Control (HH100); Viral, Bacterial and Fungal Diseases of Plants (FF610) (New March 2000); Food Storage and Preservation (QQ110); Food Processing (General) (QQ100);
13	Crop Produce (QQ050); Horticultural Crops (FF003) (New March 2000)
	Abstract: Apple is one of the greatest important fruits which is commonly consumed worldwide. There are several pathogenic fungi which target the apple and cause tremendous economic losses, and many means are being discovered in order to reduce these losses, including biological agents (BACs) used to control postharvest diseases of vegetables and fruits. The aim of this investigation was to determine the effect of the BAC Sporidiobolus pararoseus Y16 amended with a glycine betaine (GB) in the control of apple blue mold decay caused by Penicillium expansum, and the possible control mechanisms involved. The study revealed that, in vitro, S. pararoseus Y16 amended with 1 mM GB significantly reduced the incidence and diameter of lesions caused by P. expansum with respect to untreated control or to apple treated with S. pararoseus Y16 alone. In vitro, study showed that the combined treatment greatly reduced germ tube length and spore germination rate of the pathogen. The results also demonstrated that the population growth of S. pararoseus Y16 amended with GB (in vivo) had a greater increase than S. pararoseus Y16 alone at either 4°C or 20°C. Similarly, the results of the in vitro studies showed significant increase by the combination of treatments. Furthermore, the treatments did not have adverse effect on various storage quality parameter of apples. Therefore, S. pararoseus Y16 amended with GB revealed high rate of effectiveness against the blue mold decay of the apples when applied as biocontrol agent, demonstrating
14	a good potential for controlling postharvest diseases.
15	Notes: Soyabeans; Postharvest Research; Biocontrol; Horticultural Science
16	URL: http://proxy.remote.gatlib.uga.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=lah&AN=20193481858&lang=es&site=ehost-live; email: zhanghongyin126@126.com
17	Author Address: Abdelhai, M. H.: School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China.
18	Name of Database: lah
19	Database Provider: EBSCOhost
20	
21	
22	Reference Type: Journal Article
23	Record Number: 11512
24	Author: Abdelhai, M. H.; Zhang, QiRu; Zhao, LiNa; Mahunu, G. K.; Musa, A.; Yang, QiYa; Serwah, N. A.; Zhang, HongYin
25	Year: 2019
26	Title: Effects of baobab (Adansonia digitata L.) in combination with Sporidiobolus pararoseus Y16 on the activities of the defense-related enzymes and the expression levels of defense-related genes of apples
27	Journal: Biological Control
28	Volume: 139
29	Pages: 104094-104094
30	Short Title: Effects of baobab (Adansonia digitata L.) in combination with Sporidiobolus pararoseus Y16 on the activities of the defense-related enzymes and the expression levels of defense-related genes of apples
31	ISSN: 1049-9644
	Accession Number: 20193481862. Publication Type: Journal Article. Language: English. Number of References: 40 ref. Registry Number: 9001-05-2, 9001-06-3, 903-99-0, 9002-10-2, 9054-89-1, 7664-41-7. Subject Subsets:
32	Plant Pathology
	antagonists; gene expression; genes; induction; lyases; natural enemies; oxidative stress; oxidoreductases; peroxidase; peroxidases; polyphenols; postharvest systems; superoxide dismutase; tolerance; yeasts; Adansonia; Adansonia digitata; Malus; Malus domestica; Sporidiobolus; Sporidiobolaceae; Sporidiobolales; Microbotryomycetes; Pucciniomycotina; Basidiomycota; fungi; eukaryotes; Malvaceae; Malvales; eudicots; angiosperms; Spermatophyta; plants; Rosaceae; Rosales; adverse reactions; biocontrol; defense mechanisms; o-diphenol oxidase; polyphenol oxidase; redox enzymes; Forests and Forest Trees (Biology and Ecology) (KK100); Biological Control (HH100); Crop Produce (QQ050); Plant Production (FF100); Food Composition and Quality (QQ500); Plant Breeding and Genetics (FF020); General Molecular Biology (ZZ360) (Discontinued March 2000);
33	Horticultural Crops (FF003) (New March 2000)
	enzymes activity of apple fruits. The results revealed that the treatment of S. pararoseus Y16 at 1×108 cells/mL amended with baobab (AB) improved the enzyme activity and then increased the defense mechanism on apple

apples 2015-2019 | Apples 2010-2014 | Artichokes 2010-2014 | artichokes 2015-2019 | Bananas 2010-2014 | Bana

1. ábra: A CAB adatbázisból leszűrt adatállomány szerkezete (részlet)

Forrás: Cruickshank, 2020, Library of the University of Georgia, USA, kutatásra átvett állomány



	A	B	C	G	H	I	J	L	M	N	O	P	Q	R	S	T	U	V
1	books	rec_numb	authors	authors_a	year	Title	journal	veget	papers	text								
2	Apples 2010-2014	11026	Aazami, M	aazami_es	2010	Effects of	Journal of	apple	1	effect calcium chloride hot air treatment postharvest quality apple fruit red delicious apple fruit di								
3	Apples 2010-2014	11327	Abadias, M	abadias_al	2011	Evaluation	Postharver	apple	2	risk undesirable byproducts chlorine disinfection freshcut industries together limited efficacy led se								
4	Apples 2010-2014	10932	Abbott, B.	abbott_ho	2010	Comparisc	Acta Horti	apple	3	bruising major cause grading apple fruit quality serious commercial problem apple handling marketi								
5	Apples 2010-2014	10657	Abdul, Satt	abdul satt	2014	Efficacy of	Pakistan J	apple	4	penicillium expansum important postharvest pathogen causes decay apple pear fruit produces carc								
6	Apples 2010-2014	10659	Abedi, G.; ,	abedi_ahn	2013	Design anc	Australian	apple	5	fruit quality adversely affected bruise damage qualitative aspects detrimental effect impact damag								
7	Apples 2010-2014	10792	Adyanthay	adyanthay	2010	Health ber	Journal of	apple	6	increasing number studies indicate regular intake fruit vegetables clear links reduced risk chronic di								
8	Apples 2010-2014	11205	Aggelopou	aggelopou	2010	Spatial ana	Acta Horti	apple	7	objective work spatial temporal variability yield fruit quality apple malusdomestica orchard conse								
9	Apples 2010-2014	10818	Aguayo, E.	aguayo_re	2010	Effects of	Postharver	apple	8	freshcut braeburn apple slices dipped calcium ascorbate caasc ww storage air modified atmosphen								
10	Apples 2010-2014	10995	Ahmadi-Af	ahmadiafz	2013	Impact of	Postharver	apple	9	blue mold bitter rot caused penicillium expansum colletotrichum gloeosporioides devastating dise								
11	Apples 2010-2014	10814	Alamino, C	alamino_c	2013	Induction (	Pesquisa A	apple	10	objective work evaluate effect protein elicitors acibenzolarsmethyl asm harpin applied postharves								
12	Apples 2010-2014	11167	Alegre, I.; ,	alegre_aba	2010	Factors aff	Food Micr	apple	11	escherichia coli oh salmonella listeria innocua log units period freshcut golden delicious apple plug								
13	Apples 2010-2014	11047	Alegre, I.; ,	alegre_vias	2011	Microbiolc	Food Micr	apple	12	effectiveness protective culture probiotic lactobacillus rhamonosus gg rham gg salmonella listeria								
14	Apples 2010-2014	11104	Alegre, I.; ,	alegre_vias	2012	An Enterol	Postharver	apple	13	currently chlorine widely decontaminant minimally processed mp food industry achieve log reduct								
15	Apples 2010-2014	11100	Alegre, I.; ,	alegre_vias	2013	Antagonist	Food Micr	apple	14	recently reported application strain cpa pseudomonas graminis previously isolated apple reduce po								
16	Apples 2010-2014	11401	Alegre, I.; ,	alegre_vias	2013	Control of	Food Micr	apple	15	consumption freshcut fruit substantially risen last year leading number outbreaks associated fruit								
17	Apples 2010-2014	11191	Al-Hindi, R	alhindi_ali	2011	Isolation a	African Jot	apple	16	investigates current spoilage fruit fungi plant cell wall degrading enzyme various fresh postharvest								
18	Apples 2010-2014	10710	Al-Rawash	alrawashd	2014	Post-harve	American J	apple	17	control blue mold apple caused penicillium expansum harvest commercial cold storage microwave								
19	Apples 2010-2014	11199	Altisent, R	altisent_gr	2010	Aroma vol	Acta Horti	apple	18	work assess extra period time air storage ultra oxygen ulo atmosphere helpful regeneration emis								
20	Apples 2010-2014	11145	Altisent, R	altisent_gr	2011	Increased	European J	apple	19	golden cultivated apple group europe recent year mutants emerged improve aspects production se								
21	Apples 2010-2014	11286	Altisent, R	altisent_gr	2011	Comparisc	Journal of	apple	20	efforts improve fruit quality rewarded consumer satisfied tasting fruit apple often storage controll								
22	Apples 2010-2014	11280	Altisent, R	altisent_gr	2012	Eating qua	Journal of	apple	21	subsequent seasons emission volatile compounds consumer acceptability quality parameters anal								
23	Apples 2010-2014	10994	Altisent, R	altisent_ja	2013	Injection c	Journal of	apple	22	identification flavorimpact compounds potential flavors assessed incorporation naturally occurrin								
24	Apples 2010-2014	10954	Altisent, R	altisent_pl	2014	Comparati	LWT - Foo	apple	23	development freshcut apple products requires reconsideration cultivar selection characteristics re								
25	Apples 2010-2014	11073	Amarante,	amarante_	2012	Assessmer	Acta Horti	apple	24	bitter pit postharvest physiological disorder apple related ca deficiency aggravated high levels mg								
26	Apples 2010-2014	11160	Amarante,	amarante_	2012	Radiation,	Acta Horti	apple	25	damages caused hail storms apple trees avoided covering nets net change amount quality light su								
27	Apples 2010-2014	10986	Amarante,	amarante_	2013	Changes in	Acta Horti	apple	26	apple xylem becomes dysfunctional fruit develops change xylem functionality related nutritional im								
28	Apples 2010-2014	10992	Amarante,	amarante_	2013	Tissue san	Acta Horti	apple	27	apple susceptibility bitter pit bp associated ca concentrations high mgca kca nca ratios fruit metho								
29	Apples 2010-2014	11139	Amarante,	amarante_	2010	Preharvest	Revista Br	apple	28	bitter pit postharvest physiological disorder apple related ca deficiency aggravated high levels mg								

2. ábra: A munkaadatállomány szerkezete (részlet) (a fájl neve: 'mydatabase_cleaned.csv')

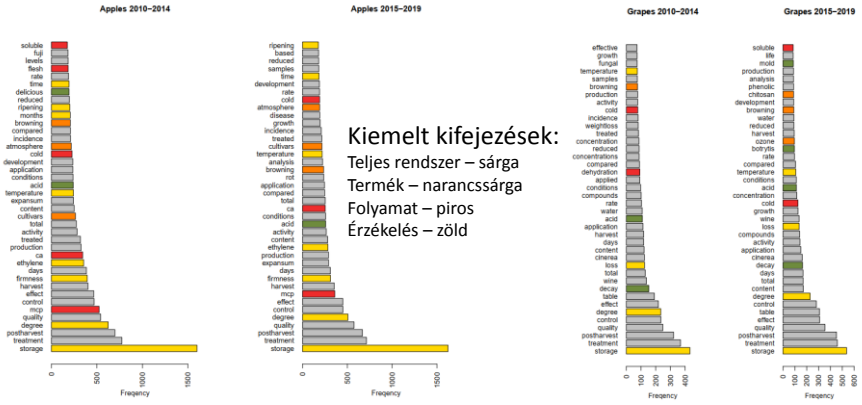
Forrás: Saját szerkesztés

A laikus felhasználó számára szolgáló információ, hogy a 'mydatabase_~.csv' nevet viselő adatbázis fejléce egyben a különböző R függvényekben a változók azonosítására is szolgál.

3.2 A szógyakoriságok vizsgálata

A szövegelemzés első fázisa a kifejezések előfordulási gyakoriságának vizsgálata. Ez a fázis akár manuálisan is elvégezhető lenne, de nagy szövegtestek (korpuszok) esetén ennek tényleges megvalósítása gyakorlatilag lehetetlen. Ebből következően ezt is célszerű program segítségével elvégezni.

Két formában is elvégezhető a művelet az outputját alapul véve: táblázatos, illetve oszlopdiagramos megjelenítéssel. Az előbbi a további adatfeldolgozásokhoz, elemzésekhez is nélkülözhetetlen, s a vizualizációhoz (oszlopdiagramhoz) is szükséges. Ez utóbbi kapcsán megoldásra került a saját színezés problémája, így a vizuális megjelenítés további információt nyújt az elemző számára.



2. ábra: A kifelezések gyakoriságának grafikus megjelemítése színhozzárendeléssel

Forrás: Saját szerkesztés

Az oszlopok színezéséhez segédtáblázatot kell elkészíteni, amely alapján a program ‘megszerkeszti’ a saját szín (mycolor) vektort, amely a oszlop-függvény paramétere.

term	color
applied	grey75
artichoke	darkorange1
ascorbic	darkolivegreen4
atmosphere	darkorange1
baby	grey75
bacterial	darkolivegreen4
bags	firebrick2
berry	darkorange1
blue	grey75
blueberry	darkorange1
botrytis	darkolivegreen4

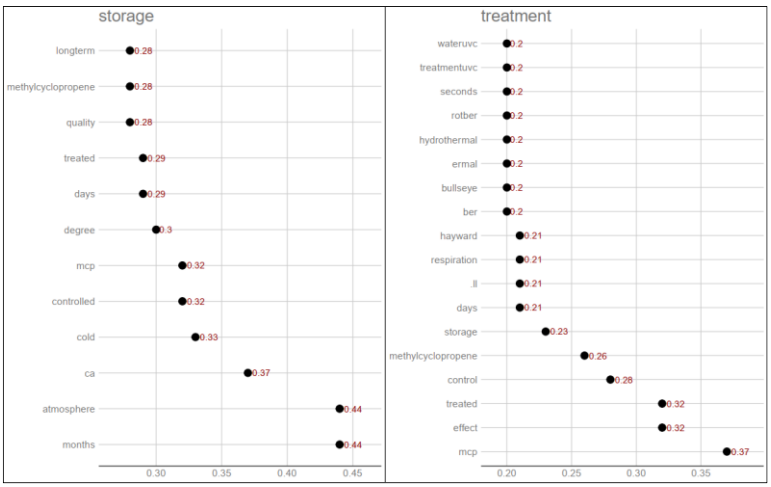
3. táblázat: A kifejezések egységesítéséhez készült segédtáblázat (részlet) (‘Mycolors_list.csv’

Forrás: Saját szerkesztés

A szógyakoriságok kiértékelésénél döntés előtt áll az elemző, hogy hol húzza meg a határt, amikor ábrázol. Túl nagy számú kifejezés értékelhetetlenné teszi az ábrát, ezért 15-20 kifejezésnél többet nem érdemes kirajzolni, amit a programutasítás paraméterezésénél lehet/kell beállítani.

3.3 Kifejezések asszociációja

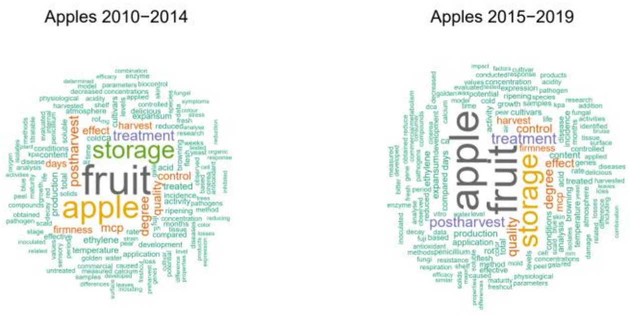
A szövegtest nem egy monolit halmaz, hanem az egyes cikkek önálló egységet képviselnek továbbra is, így a szógyakoriság szövegegységenként is értelmezhető, ezért a kifejezések kapcsolata szorosságának elemzésére alkalmas. Az elemző által kiválasztott kifejezések kapcsolatának szorosságát számolja a többi kifejezéshez. Táblázatos, illetve grafikus formában is előállítható az output. A kifejezések kapcsolata (asszociációja) vizuális megjelenítése megmutatja, hogy mely szavak vannak a legszorosabb kapcsolatban a kiválasztott kifejezéssel. (3. ábra) .



3. ábra: Kifejezések asszociációja
Forrás: saját szerkesztés

3.4 Szófelhő

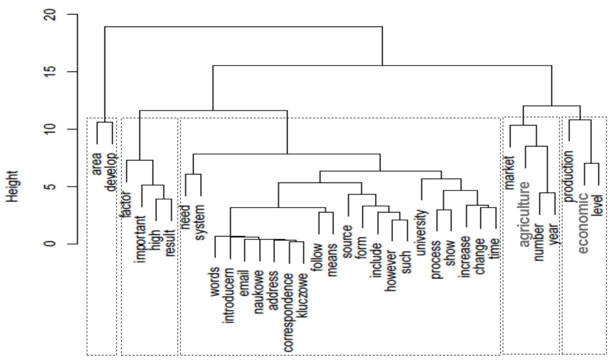
A kifejezések gyakoriságának megjelenítésében népszerű, szemléletes eszköz a szófelhő (4. ábra). Viszonylag egyszerűen paraméterezhető utasítás, ugyanakkor a kifejezések mélyebb (valós) kapcsolatrendszerének feltárására nem alkalmas, de igen látványos. Saját színek használatát nem sikerült alkalmazni, számtalan megoldási kísérlet után sem. Az output kép (.jpg) vagy .pdf formátumban elmenthető. Tapasztalat szerint utóbbit jobban lehetett a későbbiekben használni, s a nyomtatási lapméret paraméterek változtatásával (i.e. a lapméret növelésével) a felbontást is lehetett növelni.



4. ábra: A szófelhő
Forrás: Florkowski, Takács, 2022

3.5 Kifejezések és cikkek klaszterezése

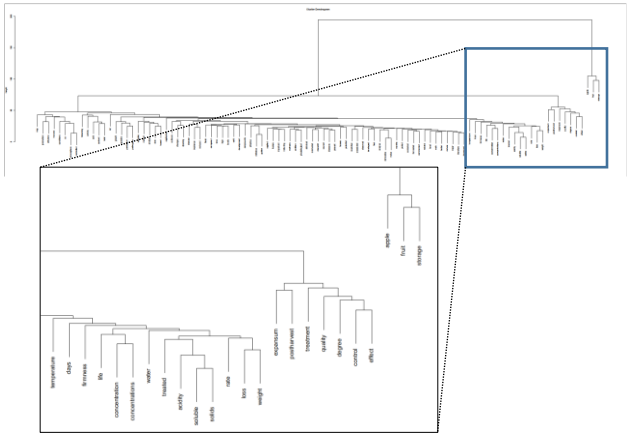
A kifejezések hierarchikus klaszterezése Ward.D módszerrel (Murtagh, F., Legendre, P., 2014). a Term-Document Matrix (Tdm) felhasználásával történt (a korpuszból egy R függvénnyel lehet előállítani). A hierarchikus klaszterezés eredményének vizuális megjelenítésére a dendrogram nagyon informatív (5., 6. és 7. ábra), ugyanakkor – magasabb hasonlósági szint megadása esetén a nagy kifejezésszám miatt – a továbbelemzéshez célszerű kisebb részeket kiemelni (lásd 6. ábra).



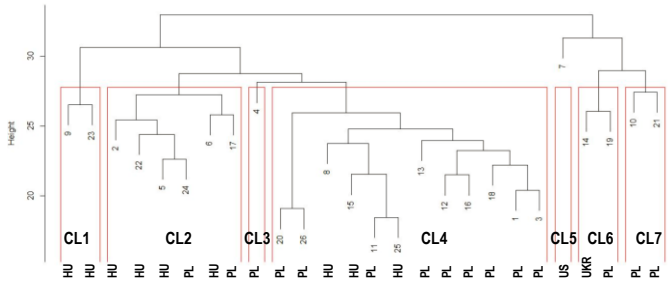
5. ábra: A 2009. évi cikkek hierarchikus kluster dendrogramja (Ward.D2, sparse 0,2)
Forrás: Takács-Takács-György, 2019

A kifejezések klaszterezése mellett értékes információval szolgál a dokumentumok csoportosítása is. Ehhez azonban transzponálni kell a szövegmátrixt, megkapva a Document-Term Matrix-t (Dtm), amely alkalmas a

dokumentumok hierarchikus klaszterezésére. (7. ábra) A csoportok további jellemzését szolgálja a klaszterek kulcs kifejezéseinek listája (4. táblázat)



6. ábra: Az Apples 2015-2019 állomány cikkeinek hierarchikus kluster dendrogramja (Ward.D2, sparse 0,2)
Forrás: Saját szerkesztés



7. ábra: A 2009. évi cikkek hierarchikus kluster dendrogramja (Ward.D2 módszer, sparse érték 0,65)
Forrás: Takács-Takács-György, 2019

Cluster	Term				
	1.	2.	3.	4.	5.
CL1.	market	agriculture	production	time	increase
CL2.	agriculture	number	area	develop	year
CL3.	factor	high	time	source	important
CL4.	develop	economic	area	factor	level
CL5.	need	system	area	production	level
CL6.	production	system	develop	economic	agriculture
CL7.	area	develop	result	change	form

4. táblázat: A klaszterek kulcs kifejezései (Annals of PAAAE, 2009. évfolyam)
Forrás: Takács-Takács-György, 2019

Az itt szemléltetésként bemutatott eredmények részletes interpretációja Takács és Takács-György (2019), valamint Florkowski és Takács (2022) forrásokban olvasható.

3.6 Kifejezések gyakoriságának továbbelemzése

A szövegelemzés során számos olyan output állítható elő (például .csv formátumban) amelyek táblázatkezelővel további elemzésekre ad módot.

A következők példában, abból a feltételezésből kiindulva, hogy a kifejezések gyakorisága összefügg a szerző által hangúlyozni kívánt mondanivalóval, a tíz évfolyamban megjelent cikkekben az egy cikkre vetített előfordulási (említési) gyakoriság vizsgálta történt. (5. táblázat) Az értelmezést segítő információ: az adattisztítás után megmaradt szókészletből évenként mintegy 100-250 kifejezésre teljesült a kritérium, hogy az adott évfolyamban publikált cikkekben legalább 50 alkalommal kerültek említésre. A jelentéstartalom alapján elvégzett további összevonás eredményeként kapott kifejezések (például 'agriculture' tartalmazva az 'agricultural' kifejezés előfordulásait is, 'farming', 'rural') jellemzően tükrözik a folyóirat szakmai orientációját is.

Kifejezések előfordulások gyakorisága sorrendjében	Átlagos előfor-dulás	Válto-zás mere-dek-sége	Átla-gos helye zés	Helye-zés módu-sza	Helye-zés szórá-sa	Leg-jobb helye zés	Leg-rosszabb helyezés	Év		Helye-zés változás mere-deksége
								Leg-jobb helye-zés	Leg-rosszabb helyezés	
agriculture	12.15	0.35	1.3	1	0.7	1	3	Több	2011	-0.08
farm/farming	9.93	-0.18	3.1	3	2.0	1	8	2011	2016	0.13
develop	9.12	-0.20	4.6	2	3.9	1	14	2012	2013	0.15
production	8.84	0.31	4.4	2	2.2	2	8	Több	2011	-0.33
area	6.92	-0.20	8.0	7	3.4	4	15	2009	2013	0.26
Poland/Polish	6.81	0.06	7.9	8	3.4	4	15	Több	2012	-0.38
market	6.14	-0.10	10.8	NA	4.3	5	18	2010	2012	0.16
economic/economy	6.04	-0.35	12.3	5	7.3	5	27	Több	2017	1.73
rural	5.67	-0.18	14.2	9	8.5	4	28	2016	2018	1.42
product	5.18	0.10	14.2	11	4.5	9	23	2018	2012	-0.72

5. táblázat: A leggyakoribb kifejezések időbeli változásának vizsgálata (Annals of PAAAE, Vol. 11-től Vol. 20-ig.) (részlet)
Forrás: Takács-Takács-György, 2019

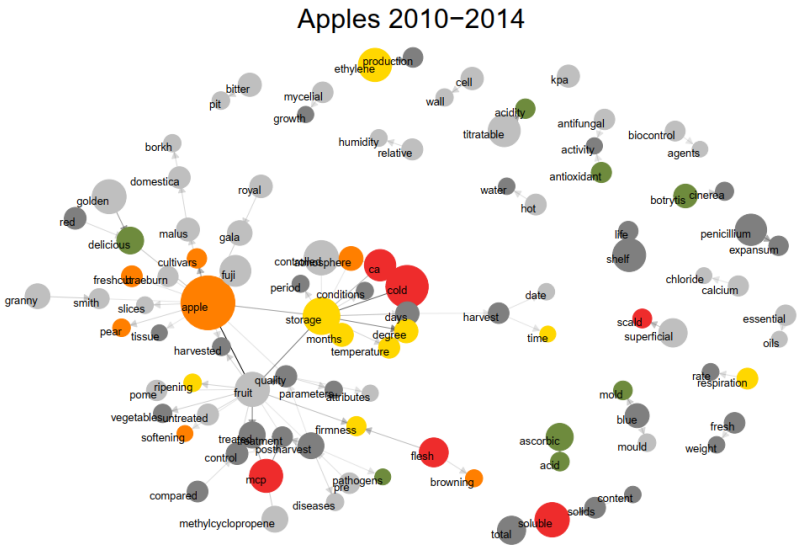
3.7 A kifejezéspárok, kifejezés láncok elemzése

A szópárok előfordulása (lásd állandó jelző az ókori irodalom óta, példaként említve 'Hókarú Nauszika'-t) meghatározó jelentést hordoz, ezért szakmai szövegekben, de akár mélyinterjúkban is fontos lehet feltárni ezeknek a szókapcsolatoknak az előfordulási gyakoriságát. Az elsődleges információtartalom

(az előfordulási gyakoriság számszerű azonosítása) mellett további lehetőség is rendelkezésre áll: megjeleníthetők a vizuális kapcsolatok (lásd 8. ábra), a kifejezések előfordulási gyakorisága (a kifejezéshez rendelt pontok arányos méretezésével), a kifejezések csoportba sorolása (a saját szín alkalmazásával), illetve a nyilak megmutatják a szókapcsolat irányát (sorban egymás után következésüket.

Szópárok				Szóhármasok			
word1	word2	n		word1	word2	word3	n
fruit	quality	124		controlled	atmosphere	ca	53
fruit	storage	122		total	soluble	solids	52
fruit	treated	104		apple	malus	domestica	50
fruit	firmness	77		malus	domestica	borkh	43
fruit	vegetables	55		golden	delicious	apple	38
fruit	harvested	51		soluble	solids	content	33
fruit	ripening	50		granny	smith	apple	26
fruit	apple	28		storage	controlled	atmosphere	25
fruit	softening	28		controlled	atmosphere	storage	24
fruit	weight	27		fresh	weight	stage	24

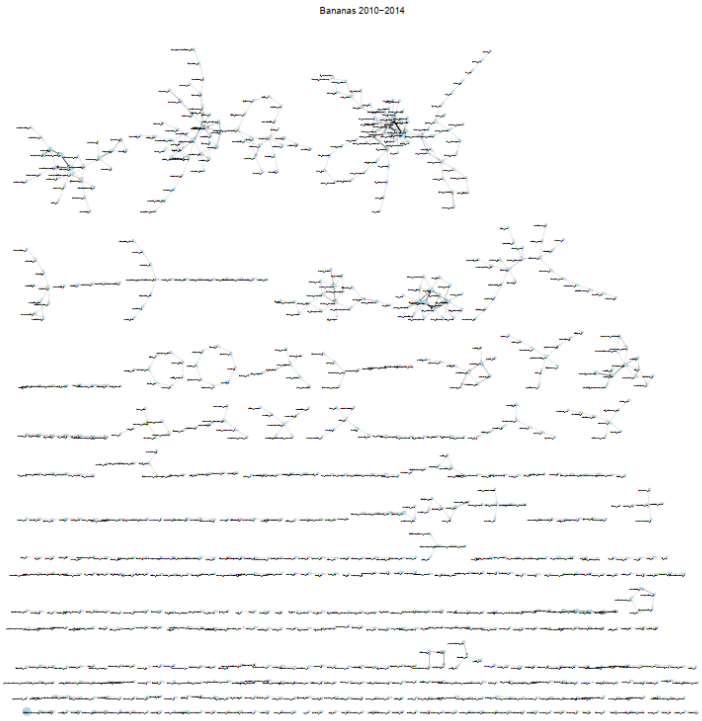
6. táblázat: A leggyakoribb szópárok és szóhármasok előfordulási gyakorisága (Apples 2010-2014, részlet)
Forrás: saját szerkesztés



8. ábra: Kifejezések hálózata: a szókapcsolatok vizuális megjelenítése (Apples 2010-2014)
Forrás: Florkowski, Takács, 2022

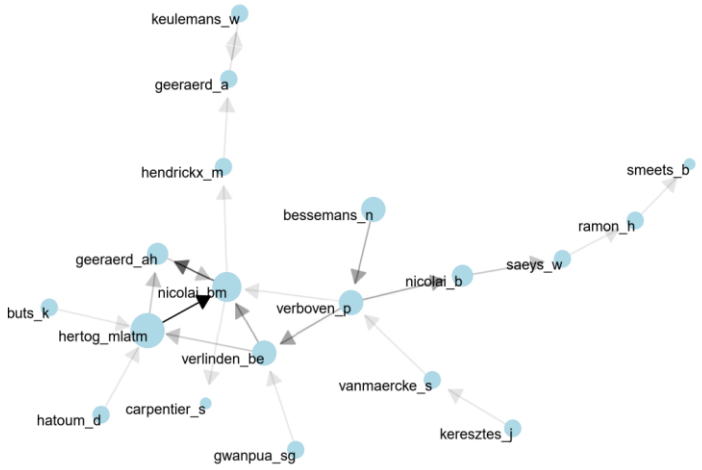
3.7.1 Kifejezéspárok felhasználása a cikkek szerzői kapcsolati hálójának feltérképezésére

A kifejezések hálózatosodása feltérképezésének sikeres megvalósítása után felvetődött, hogy a cikkek szerzői által alkotott szerzői hálózatok is feltárhatók-e ugyanezzel az eszközzel. Már az első kísérletek bíztatók voltak, de nyilvánvalóvá vált, hogy a 8. ábrán is látható ábrázolásmód esetén nehezségek árán azonosíthatók a csoportok. Ennek egyik oka, hogy annak a feltétel teljesítése, hogy minden szerző megjelenjen, akkor is, ha egyszezős a cikk, megnövelte az ábrázolandó egységek (szavak, azaz szerzők neve) számát, másrészt a program a tér egyenletes kitöltésre a csak egy-egy cikket jegyző szerzői teameket behelyezte a hálózatok közötti "lyukakra", aminek a következtében első látásra nem volt lehatárolhatók egymástól a szerzői hálózatok. Ilyenkor van segítségre, ha a programutasítások részletes leírását, lehetséges paraméterezését tanulmányozzuk. A 8. ábránál a layout = "fr" opció helyett a "stress" opció bizonyult használhatónak (9. ábra), amelynél az egy-két cikkes szerzők alul helyezkednek el sorokban, s részletesebben tanulmányozva jól azonosíthatóak a szerzői teamek. Az érdekes a felső rész, ahol viszont kirajzolódnak a hálózatok. A példaként hozott minában igen kiterjed kapcsolati hálók jelennek meg. A 2. táblázatból tudható, hogy a 2010 és 2014 közötti öt évben banánnal kapcsolatosan megjelent cikkekből 320 szerepel a CAB adatbázisban. A 8. ábra átfogó kép nyújtására alkalmas, de részletes elemzés csak az abból kivágott részleteken végezhetők.



9. ábra: A szerzők hálózatosodása (Bananas 2010-2014)
Forrás: saját szerkesztés

A 10. ábrán szereplő kimetszésen ugyanúgy azonosíthatók az előfordulási gyakoriságok (pontok mérete), illetve követhetők a sorrendek (nyilak iránya), láthatók a kölcsönösségek (mindkét irányba mutató nyíl).



10. ábra: Azonosított szerzői hálózat (részlet a szerzők hálózatosodása ábrából) (Apples 2015-2019)
Forrás: Florkowski, Takács, 2022

3.8 A szövegek érzelmi tartalmának elemzése

A mindennapi életben is érzelmi tartamat, jelentést is rendelünk egyes kifejezéskhez, s azokat vagy pozitív vagy negatív töltöttségűnek gondoljuk. Magának az érzelmi hanulatát a pozitív és negatív töltésű kifejezések aránya határozza meg, s egy szépirodalmi szövegnél ez hatással is van az olvasóra. Szakcikkek esetén ez az érzelmi hatás nem feltétlen érvényesül, ugyanakkor a problémára való figyelemfelhívás sikeressége, a tárgyalt szakmai kérdés jellegének érzékeltetése a kifejezések érzelmi töltségével is erősíthető.

A 11. ábra egy kísérlet eredményét mutatja, amelyben egy általános érzelmi szótár (‘lexicon’) felhasználásával került vizsgálatra a leggyakoribb kifejezések érzelmi töltöttsége. A kísértlet használható eredményt hozott. Természetesen a szakmai sajátosságokat egy saját szótár megalkotásával és meghivatkozásával érvényesíteni lehet, azonban az jelentős munkaráfordítást igényel.

3.9 A kifejezések fontossága a vizsgált dokumentumokban

A kifejezések dokumentumokban betöltött szerepéről ad információt az úgynevezett tf-idf analízis. Az információ-lekérdezésben a tf–idf (term frequency – inverse document frequency) a gyakoriság – inverz dokumentumgyakoriság, tükrözi, hogy mennyire fontos egy szó a korpuszban található dokumentum számára. Két hányados (pontosabban az egyik logaritmusával való) szorzatként számolható (1), amelyben az egy-egy kifejezés előfordulásának aránya az összes

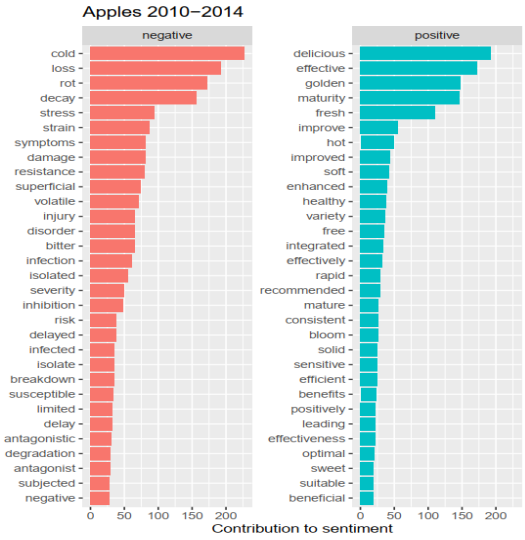
szóból (2) kerül megszorzásra annak logaritmusával, hogy az összes dokumentum hányszorosa a kifejezést tartalmazó dokumentumok számának (3).

tf: kifejezés előfordulása/összes szó előfordulása (1)

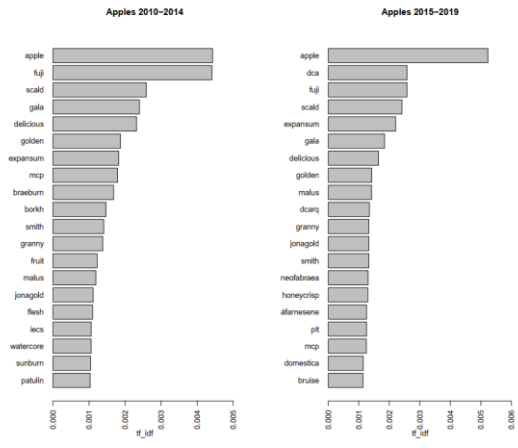
idf: log(összes dokumentum száma/kifejezést tartalmazó dokumentum száma) (2)

tf-idf = tf * idf (3)

A 12. ábra, illetve 6. táblázat mutat be példát az elemzési módszerre.



11. ábra: Kifejezések érzelmi töltöttsége (Apples 2010-2014)
Forrás: saját szerkesztés



12. ábra: A kifejezések tf-idf elemzése (Apples 2010-2014 és Apples 2015-2019)
Forrás: saját szerkesztés

Dokumentumcsoport	Szó	n	Összes szó	tf	idf	tf_idf
Apples 2010-2014	apple	2139	89855	0.02381	0.18610	0.00443
Apples 2010-2014	fuji	182	89855	0.00203	2.17853	0.00441
Apples 2010-2014	scald	148	89855	0.00165	1.57240	0.00259
Apples 2010-2014	gala	137	89855	0.00152	1.57240	0.00240
Apples 2010-2014	delicious	193	89855	0.00215	1.07992	0.00232
Apples 2010-2014	golden	148	89855	0.00165	1.13708	0.00187
Apples 2010-2014	expansum	243	89855	0.00270	0.67446	0.00182
Apples 2010-2014	mcp	525	89855	0.00584	0.30673	0.00179
Apples 2010-2014	braeburn	64	89855	0.00071	2.36085	0.00168

6. táblázat: Kifejezések tf-idf elemzés (Apples 2010-2014, részlet)
Forrás: saját szerkesztés

A 6. táblázatból tf-idf oszlop értékeiből látható, hogy nem feltétlenül a nagyobb előfordulási gyakoriságú (n) kifejezés tf-idf értéke lesz magasabb.

Következtetések

A szakirodalmi források elemzése fontos alapozó fázisa minden kutatásnak. A kutató végső célja új eredmények létrehozása, így a már meglévő tudás "újrafelfedezése" haszontalan időráfordítás. Az okszerű forráselemzés nem csak segít elkerülni ezt, hanem inspirálólag hathat a kutatóra, illetve segít kijelölni azokat a réseket, amelyekre behatolva új vagy újszerű eredmények létrehozására van esély.

A korszerű információfeldolgozási eszközök, az adatbányászat, azon belül is a szövegbányászat (ha van hozzáférés) nagy adataállományok hatékony kezelésére alkalmasak.

A cikkben bemutatott eljárások a nyílt hozzáférésű R programozási nyelv utasításkészletét használva, saját fejlesztésű rutinokkal (amelyek terjedelmi okokból nem kerültek be a cikkbe, de igény esetén megkaphatók) bizonyították, hogy a szakirodalmi források elemzése, de adott esetben nagyobb számú mélyinterjú kiértékelése is történhet emelt szinten (statisztikai eszközökkel) is.

Összefoglalás

A tanulmány az irodalomkutatás tudományos színvonalának emeléséhez mintákat mutatott be két korábbi kutatás kapcsán alkalmazott, többnyire továbbfejlesztett R programnyelv alapú szövegbányászati eszközök alkalmazására. A két korábbi kutatás: egy szakfolyóirat 10 évfolyamában közreadott cikkek elemzése, illetve egy nemzetközi projekt a Postharvest handling 4. kiadásának az elkészítése, amelyben az egyik alprojektjében kerültek alkalmazásra a szövegbányászati módszerek. A kutatásokban egyrészt a 2009 és 2018 között az Annals of the PAAAE folyóiratban megjelent 393 angol nyelvű teljes cikk, másrészt a CAB adatbázisban lévő, 2010-2014, illetve 2015-2019 közötti időszakban 1055 kertészeti folyóiratban (25 gyümölcs és zöldség faj kapcsán) megjelent 9246 cikk absztraktjainak vizsgálata történt. Mindkét kutatás arra keresve a választ, hogy a

kutatások fókusza hogyan változott a vizsgált időszakokban. A kutatás során R programnyelven fejlesztett rutinokkal került vizsgáltra többek között a kifejezések előfordulási gyakoriságát, asszociációját, elkészült a kifejezéspárok hálója, a szerzők hálója, a kifejezések és cikkek klaszterezése is. A cikk számba veszi az eredmények vizuális interpretálásának lehetőségeit is, a kapott eredményekből választott példákkal illusztrálva.

Hivatkozások

[1] Balińska Agata (2009): Tourism as a Form of Non-Agricultural Activity in the Opinion of the Selected Rural Communities Inhabitants. Annals of the PAAAE; 11 (6): 5-10.

[2] Brelik Agnieszka (2009): Rural Tourism Development in Poland. Annals of the PAAAE; 11 (6): 17-20.

[3] Florkowski, Wojciech, J., Takács István (2022): What mining the text tells about minding the consumer: the changing fruit and vegetable consumption patterns and shifting research focus. In: WOJCIECH, J. FLORKOWSKI; NIGEL, H. BANKS; ROBERT, L. SHEWFELT; STANLEY, E. PRUSSIA (szerk.): Postharvest handling: A Systems Approach. London, Egyesült Királyság: Academic Press as an imprint of Elsevier. 684 p. pp. 517-564.

[4] Keresztúri Judit Lilla, Antal Beáta, Illés Ferenc (2017): Bevezetés az R programozásba. Egyetemi jegyzet Vállalati pénzügyi információs rendszerek című tantárgyhoz. Budapesti Corvinus Egyetem. 64 p. Online: http://unipub.lib.uni-corvinus.hu/2707/1/Bev_R_prog.pdf. Letöltés: 2022.10.01.

[5] Koreleski Dariusz (2009): Interregional Differentiation of the Development Level in Poland With Regard to the Chosen Parameters. Annals of the PAAAE; 11 (6): 66-71.

[6] Murtagh, Fionn, Legendre, Pierre (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification, 31, 274–295.

[7] Silge, Julia, Robinson, David (2017): Text Mining with R. A Tidy Approach. O'Reilly Media. 184 p. Online: https://www.pdfdrive.com/text-mining-with-r-a-tidy-approach-d158192175.html. Letöltés: 2022.10.01.

[8] Solymosi Norbert (2005): R <-…erre, erre …! Bevezetés az R-nyelv és környezet használatába. SZIE Állatorvostudományi Kar. 123 p. Online: https://cran.r-project.org/doc/contrib/Solymosi-Rjegyzet.pdf. Letöltés: 2022.10.01.

[9] Staszewska Sylwia (2009): Rural Areas – a New Space for Urban Development. Annals of the PAAAE; 11 (6): 115-120.

[10] Takács István, Baranyai Zsolt (2009): Agricultural Products Noted on Commodities Exchange and Global Financial Crisis. Annals of the PAAAE; 11 (6): 121-127.

[11] Takács, István; Takács-György, Katalin (2019): Main focuses of English papers of annals (PAAAE) during the last ten years. annals of the Polish Association of Agricultural and Agribusiness Economists 21: 3 pp. 470-480.

[12] Wieliczko Barbara (2018): Financial Instruments in Cap 2020+. Annals of the PAAAE; 20 (4): 205-209.

[13] Williams Graham (2016): Hands-On Data Science with R Text Mining. https://onepager.togaware.com/TextMiningO.pdf. Access: 30.05.2019

[14] Witten Ian H. (2004): Text mining. https://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf. Access: 30.05.2019

[15] Wojcieszak Monika, Jan Zawadka (2018): Cultural Values as a Determinant of the Development of Tourism in Rural Areas and their Popularity Among Poles Based on the Example of Folk Culture Museums. Annals of the PAAAE; 20 (1): 149-155.

[16] Zhao Yanchang (2013): R and Data Mining: Examples and Case Studies. http://www2.rdatamining.com/uploads/5/7/1/3/57136767/rdatamining-book.pdf. Access: 30.05.2019