Mészáros Evelin

(Tudományos segédmunkatárs, MTA TK Politikatudományi Intézet)

Sebők Miklós

(Tudományos főmunkatárs, MTA TK Politikatudományi Intézet)

A szövegbányászati módszerek alkalmazásának lehetőségei a joggyakorlat-elemzésben

Bevezetés[1]

Az információs technológia fejlődésével a jogrendszerben rendelkezésre álló adatok mennyisége folyamatosan növekszik. Az új információk tárolása és elemzése egyre nagyobb emberi és technikai kapacitást igényel. Az adatállomány bővülésének és ennek kezelésének problematikájával az adatbányászat foglalkozik, amely az 1980-as évek végétől kezdve teljesedett ki, mint önálló tudományterület. Ekkortól került az információs technológia, de egyben társadalom- és jogtudományi elemzések homlokterébe a Big Data, a nagyméretű adatbázisok menedzsmentjének és a bennük «rejtőző» információk kinyerésének kérdése.

Ez utóbbival foglalkozik az adatbányászat, melynek célja a látens információk kinyerése nem tökéletesen strukturált adatbázisokból. Az adatbányászat a közigazgatási- és jogrendszer számára is egy fontos eszközt jelenthet működésének optimalizálására. Ahogy egy, a digitális társadalom kialakulásával foglalkozó hazai tanulmány fogalmazott „a nemzetközi tapasztalatok azt mutatják, hogyha az állam első számú vezetője érdektelen az információs korszak adta új állam-és társadalomszervezési lehetőségek kihasználása iránt, akkor lelassul az átalakulás, szigetszerű marad az elektronikus ügykezelés és közigazgatás, nem csökken, hanem újratermelődik a digitális írástudatlanság a társadalomban."[2] Nem csak az állampolgárok ügyfélközpontú kiszolgálása teremthető meg adatbányászati támogatással, de a közigazgatási és peres eljárások során keletkezett dokumentumok feldolgozása, tárolása és elemzése is meggyorsulhat, ami javítja az érintett szolgáltató rendszerek hatékonyságát.

Már a fentiekből is következik, hogy a Big Data adatbázisok ugyanúgy épülhetnek számokra, mint szövegekre. Az ezek hasznosításával foglalkozó adat- és szövegbányászat a gyakorlati kutatások során gyakran össze is mosódik, amit jelez a „text and data mining" tudományterületi elnevezés is.[3] Ezzel együtt a nagyméretű szöveges adatbázisok kezelése speciális kihívást jelent számos társadalmi alrendszerben, így a jogrendszerben is. Ezek eltérnek a hagyományos numerikus adatbázisoktól, amennyiben nem kvantitatív, hanem kvalitatív információkat (szavakat) tartalmaznak. Ahhoz, hogy ezeket gépi elemzésre alkalmas formátumra először tehát át kell alakítanunk őket adattá: ez a „szöveg, mint adat" (text as data) paradigmája.

Miközben az adat- és ezen belül a szövegbányászat a nemzetközi üzleti életben és jogrendszerekben egyre nagyobb hangsúlyt kap, ennek hazai oktatása és alkalmazott kutatási használata még gyerekcipőben jár. A tudományterület társadalomtudományi-közpolitikai célú fejlesztésének igénye nyomán alakult a Magyar Tudományos Akadémia Társadalomtudományi Kutatóközpontjában (MTA TK) az a POLTEXT elnevezésű kutatócsoport, melynek e sorok szerzői is tagjai.

A projekt keretében 2016-ban megjelent *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban* című tankönyv[4], illetve 2017 novemberében a kutatócsoport két napos intenzív szövegbányászati képzést szervezett, mely az erősen jelentkező kutatói érdeklődést volt hivatott kielégíteni. A POLTEXT projekt nemzetközi beágyazottságának növelésében és a területen kulcsfontosságú tudástranszfer biztosításában kulcsszerepet játszott a 2018-ban mintegy harminc külföldi és hazai előadó részvételével megrendezett szövegbányászati workshop.

Az alábbiakban a kutatócsoport fent említett eredményeire építve először bemutatjuk a szövegbányászat alapfogalmait és legfontosabb eljárásait egy a Kúria joggyakorlatára épülő mini esettanulmány keretében. A 3. és 4. fejezet ezen esettanulmány kapcsán alkalmazza az egyik bevett módszer, a csoportosítás két változatát: a hierarchikus illetve a k-közép klaszterelemzést. A záró szakasz röviden összefoglalja a szövegbányászát lehetséges felhasználását a joggyakorlat elemzésben.

[1] A tanulmány az NKFIH FK- 123907 témaszámú projektje illetve az MTA TK POLTEXT Inkubátor támogatásával készült.

[2] Csepeli, Gy., Síklaki, I., Rudas, T. és Nagyfi, R.: Üzleti és közpolitikai alkalmazások, marketing, adatbányászat, szociális intelligencia. Eötvös Loránd Tudományegyetem, Társadalomtudományi Kar, 2012 Digitális Tankönyvtár (forrás: http://tarsadalominformatika.elte.hu/tananyagok/uzletikozp/lecke7_lap1.html Letöltve: 2018. 01.26.)

[3] Holl, A. (2015). Szövegbányászat, adatbányászat, ismeretfeltárás. Új lehetőségek a tudományos kommunikációban. Magyar Tudomány, (6), 680–685. pp. 680.

[4] Sebők, M. (szerk.) (2016). Kvantitatív szövegelemzés és szövegbányászat a politikatudományban. L'Harmattan.

2. Szövegbányászati alapfogalmak és eljárások

2.1. Alapfogalmak

A továbbiakban egy, a Kúria joggyakorlatából vett mini esettanulmány keretében mutatjuk be a szövegbányászat legfontosabb fogalmait és eljárásait. Ehhez egy relatíve kisméretű, a Kúria fizetésképtelenséggel kapcsolatos határozataiból álló szöveges adatbázist használtunk fel, melyet a bíróság honlapjáról töltöttünk le.[5]

A szövegbányászat központi fogalma a **korpusz**, ami egy „meghatározott szempontok alapján kiválasztott szövegmennyiség, amelyen a nyelvész a vizsgálatát végzi"[6]. A korpusz tehát a szövegek egyfajta tárháza, a szó legtágabb értelmében, hiszen sem a tárolás módjára, sem az adatmennyiségre vonatkozóan nem tartalmaz megszorításokat. Joggyakorlat-elemzési esettanulmányunk szempontjából ez a rendelkezésre álló dokumentumok halmaza, jelen esetben 36 darab fizetésképtelenségi üggyel kapcsolatos határozat szöveges formátumban.

A szövegbányászati eljárások alkalmazása során a szövegre mint adatra tekintünk, és az egyes dokumentumokat a benne szereplő szavak és kifejezések gyakoriságával jellemezhetjük. Ehhez kapcsolódik egy másik k**özponti fogalom**, a **dokumentumkifejezés-mátrix**, amely egy olyan táblázat, amelynek soraiban a fájlnevek szerepelnek, oszlopaiban pedig a kifejezések, és a cellák tartalma arra utal, hogy az adott oszlopban szereplő kifejezés az adott dokumentumban hány alkalommal szerepel (ld. 1. táblázat)

	bírósági	bíróságnak	biztosított	címlap	csőd	csődegyezség	csődegyezségben	csődegyezséget	csődegyezségi	csődeljárá
Cspkf. VII.30.146_2013_3. számú határozat.txt	1	0	2	1	1	1	0	0	0	0
Gfv. VII.30.017_2014_9. számú határozat.txt	0	3	0	1	0	8	1	4	1	2
Gfv. VII.30.034_2013_10. számú határozat.txt	0	0	0	1	0	0	0	0	0	3
Gfv. VII.30.098_2014_6. számú határozat.txt	0	0	0	1	0	8	1	0	0	1
Gfv. VII.30.171_2013_13. számú határozat.txt	3	5	2	1	1	13	1	2	0	9

1. táblázat
Példa a dokumentumkifejezés-mátrixra

Az 1. táblázatból látszik, hogy a hatékony adatelemzéshez szükség van a korpusz tisztítására is – mivel egy adott szövegben egy konkrét szó többféle alakban is előfordulhat és vannak a feladat szempontjából nem lényeges szavak (kötőszavak, névelők stb.).

2.2. A szövegek előkészítésének fázisai

A szövegek elemzésre való előkészítése legalább három fontos lépésből áll: tokenizálás, a ragozás eltávolítása, és a tiltólistás szavak kiszűrése. A **tokenizálás** során a dokumentumokat egységekre – *token*-ekre – bontjuk. Ezek lehetnek szavak vagy kifejezések. A token tehát egy adott dokumentumban szereplő karaktersorozat, ami együttvéve jelentéstani elemzési egységként szolgál.

A tokenizálás már magában foglalja az írásjelek eltávolítását is, ami bizonyos nyelvek esetében önmagában is egy komplex feladat (például angol nyelv esetén az aposztrófok eltávolítása ellehetetleníti az adott token felismerését ld. *aren't*). Magyar nyelv esetében ez kevésbé hangsúlyos problémába, mely pl. a rövidítések esetében merülhet fel.

Ezt a lépést követi a **ragozás eltávolítása/szótövesítés** (*stemming*). Míg az előző folyamat az angol szövegek elemzésében okozhat némi fejtörést, a ragozás eltávolítása a magyar nyelvnél okozhat nehézségeket. Mivel nyelvi elemzésről van szó, figyelembe kell vennünk az adott nyelv nyelvtani sajátosságait is. A magyar nyelv az agglutináló nyelvek egyike, ami azt jelenti, hogy a szavak alaptövéhez kapcsolunk toldalékokat (képzőket, jeleket és ragokat).

2.3. Információ-visszakereső és leíró statisztikai módszerek

A dokumentumkifejezés-mátrix segítségével könnyen előállíthatóak az adott korpusz leíró statisztikái, olyan információk, melyek jól reprezentálják a korpuszt. Ez alapján könnyen vissza kereshetőek a korpuszos egyes elemei, illetve mélyebb, ún. információkinyerő elemzések is elvégezhetőek[7]. Ez utóbbiakat a kutatási kérdéstől függően deduktív vagy induktív módon is elvégezhetjük.

Deduktív módszertan alkalmazása esetén rendelkezésre áll egy előre definiált kategória-rendszer és a szövegbányászati feladat célja a korpusz elemeinek besorolása ezen kategóriákba. Jó példa erre a határozatok tematikus besorolása egy előre megadott klasszifikációs sémába. Induktív módszertan esetén nem állnak rendelkezésre ilyen előre megadott kategóriák, így ez a módszertan akkor alkalmazandó, ha az empirikus anyag mintázatait szeretnénk feltárni, melyeket ezek megismerése után lehet interpretálni és a kialakuló csoportokat címkékkel ellátni.

Rátérve a konkrét gyakorlati szövegbányászati eljárásokra, az egyik leggyakoribb feladat a korpusz jellemzőinek feltárása. A dokumentumok jellemzőit a szavak és kifejezések gyakoriságával azonosítjuk, és ennek a kinyerésére lehet példa a **szózsák** módszere, amely során az egyes szavak gyakoriságát vizsgálhatjuk egy korpuszon belül (ebben az esetben a szavak sorrendjével és kontextusával nem foglalkozunk).

Az így meghatározott leggyakoribb kifejezések grafikus megjelenítéséhez használhatunk **szófelhő**t, ami

[5] Honlap: http://www.kuria-birosag.hu/hu/fizkepugy

[6] Kugler, N. - Tolcsvai Nagy, G. (2000). Nyelvi fogalmak kisszótára. Korona Kiadó. pp. 132.

[7] Minderről bővebben ld. Sebők, M. (szerk.) (2016). Kvantitatív szövegelemzés és szövegbányászat a politikatudományban. L'Harmattan. pp. 39.

egy könnyen áttekinthető, de korlátozott információ-ótartalmú ábrában foglalja össze a korpusz bizonyos jellemzőit (ld. 1. ábra).

1. ábra

A Kúria-korpusz leggyakoribb elemei

Az *1. ábrán* a fizetésképtelenségi ügyek kapcsán összeállított korpuszunkban legalább 75-ször előforduló kifejezések szerepelnek. Az ábra visszaigazolja a dokumentumok összeállításának elvét, hiszen a leggyakoribb szavak valóban a fizetésképtelenséggel és az ezek kapcsán hozott bírósági határozatokkal kapcsolatosak. Ahhoz ugyanakkor, hogy jobban megértsük a korpusz tartalmát és az egyes elemek összefüggéseit további vizsgálatok szükségesek.

Egy további hasznos lépés lehet a korpuszban leggyakrabban előforduló kifejezések listázása. A 2. ábra egy részletet közöl arról a listáról, amely a korpuszban legalább 10-szer szereplő szavakat tartalmazza betűrendben. Ez a lista még viszonylag hosszú, de igény szerint szűkíthető (ha a legalább 150-szer előforduló kifejezéseket vesszük számításba, akkor már csak 3 kifejezés marad: „adó", „bíróság", „cstv").

```
> findFreqTerms(adtm, lowfreq=10)
  [1] "adó"                 "adósnak"             "adóssal"
  [4] "adóst"               "adott"               "alá"
  [7] "alapján"             "alapjául"            "alkalmazandó"
 [10] "áll"                 "állapította"          "álláspontja"
 [13] "állást"              "álló"                "amennyiben"
 [16] "attól"               "azaz"                "bank"
 [19] "befejező"            "bekezdé"             "bekezdés"
 [22] "bekezdésében"        "bekezdésének"        "benyújtott"
 [25] "betölté"             "bíróság"             "bírósághoz"
 [28] "bírósági"            "bíróságnak"          "biztosított"
 [31] "címlap"              "csőd"                "csődegyezség"
 [34] "csődegyezségben"     "csődegyezséget"      "csődegyezségi"
 [37] "csődeljárá"          "csődeljárás"         "csődeljárásában"
 [40] "csődeljárásban"      "csődeljárást"        "cstv"
 [43] "dátuma"              "dolog"               "ebből"
 [46] "egyet"               "egyetért"            "egyezség"
 [49] "egyezséget"          "egyezségi"           "együtt"
 [52] "eljárá"              "eljárásban"          "eljárási"
 [55] "eljárást"            "elleni"              "elsőfokú"
 [58] "elsőként"            "érdekében"           "érdemben"
```

2. ábra

Legalább 10-szer előforduló szavak

Egy fokkal összetettebb képet nyújt a korpusz kifejezéseinek együttes előfordulásának vizsgálata. A 3. ábra arról nyújt információt, hogy a korpuszban a „jog" kifejezéssel milyen szavak szerepelnek együtt a leggyakrabban. E szerint például a „bíróságok" (ld. az ábra bal felső sarka) kifejezés ennek előfordulásainak 72 százalékában együtt szerepel a jog kifejezéssel (azaz a „bíróságok" kifejezést tartalmazó 10 dokumentumból nagyjából 7-ben a „jog" kifejezés is megjelenik).

```
> findAssocs(adtm, "jog", .60)
$jog
  bíróságok  különböztet       döntést     szavazati        illeti
       0.72          0.72          0.69          0.69          0.67
     tartani     alpontjaiba       aszerint          bank    besorolnia
       0.66          0.61          0.61          0.61          0.61
   bizonyosan    biztosítékot    elbírálná        érint        hozhat
       0.61          0.61          0.61          0.61          0.61
    hozhatja      igénynek         illett    jogosítváni   jogosítványt
       0.61          0.61          0.61          0.61          0.61
   kaphasson   kategóriában      kérdéses     keresetét       kezelni
       0.61          0.61          0.61          0.61          0.61
 követelésér  követelésrészt  követelésüket  megalapozza meghatározhatja
       0.61          0.61          0.61          0.61          0.61
  megilleti     megtámadta        mennyi     nyújtania       összeget
       0.61          0.61          0.61          0.61          0.61
   pontjába      pontjainak       sorolja     szövegből        vehett
       0.61          0.61          0.61          0.61          0.61
     vétel visszafizetésér     vitatottan      banknak
       0.61          0.61          0.61          0.60
```

3. ábra

Jog kifejezéssel az esetek minimum 60%-ában együtt járó szavak

Egy további alapvető eljárás a **névelemfelismerés** (*Named Entity Recognition*). Az információ-visszakereséssel szemben ennek egyes alkalmazásai már a strukturálatlan adatokat is kezelni képes információ-kinyerés területére vezetnek (ld. alább). Névelemként definiálhatunk minden olyan szóalaksorozatot (*tokensorozatot*), amely valamely egyedi tulajdonságra utal. Ez a legtöbb esetben tulajdonnevek keresésére irányul, de a névelemfelismerés lehetőségei ennél sokkal tágabbak.[8] A joggyakorlat-elemzés során ez a szóalaksorozat a személynevek mellett megjelenhet határozatok címében vagy cégnevekben is. Fontos hangsúlyozni, hogy alkalmazásához a legtöbb esetben szükségünk van egy szótárra, amely tartalmazza a kutatásunk szempontjából releváns szóalakok szótövesített változatát.

2.4. Információkinyerő módszerek

Amennyiben a szöveg-korpusz olyan információkat szeretnénk „kibányászni", melyek közvetlenül nem visszakereshetőek (pl. azáltal, hogy rákeresünk az adott kifejezésre), akkor információkinyerő módszereket kell használnunk. Ennek számos eljárása ismert, melyek közül itt csak néhányat említünk. Az automatizált **kivonatolás** vagy összefoglalás hasznos lehet akkor, ha az egyes dokumentumok (pl. ítéletek) hosszúak, és fontos, hogy a tartalmukat gyorsan és jó minőségben reprodukáljuk egy rövidített változatban. A névelemfelismeréshez hasonló logikát követ a **fogalomkinyerés,** amely szótári elemek felismerésére irányul, melyekből egy listát állít elő. Ennek egyik alfeladataként határozható meg a **témakinyerés**, amely automatikusan azonosítja a jelentősebb szótári elemeket a dokumentum témájának meghatározása érdekében.[9]

Az eddigiekben olyan technikákat tekintettünk át, amelyek egy dokumentumban szereplő szavak és kifejezések relatív gyakoriságát vették figyelembe függetlenül a kontextustól. Ennél bonyolultabb eljárás az érzelemdetektálás (*sentiment analysis*). Ebben az esetben a dokumentumokat előre definiált jellemzőkkel (jellemzően: pozitív, negatív vagy semleges tartalmú) szeretnénk ellátni, annak érde-

[8] Szarvas, Gy. – Farkas, R. (2007): Információkinyerés. In: Tikk Domonkos (szerk.): Szövegbányászat. Budapest, Typotex Kiadó, 81–101. pp. 91.

[9] https://www.ibm.com/support/knowledgecenter/hu/SS5RWK_3.5.0/com.ibm.discovery.es.nav.doc/iiysgloss.htm#iiysgloss__theme. extraction. Letöltve: 2018. 01. 26.

kében, hogy tartalmukról közvetlenül vissza nem kereshető információt szerezzünk.

A komplexebb szövegbányászati feladatok közül az alábbiakban az **osztályozásra** és **csoportosításra** összpontosítunk, melyek a kvantitatív szövegelemzés leggyakoribb alkalmazási területei közé tartoznak. E két fogalom tükrözi a fejezetben taglalt módszertani megkülönböztetést: az osztályozás (deduktív eljárást követve) előre meghatározott kategóriákba sorolja a korpusz elemeit, míg a csoportosítás az induktív logikának megfelelően a dokumentumok szó-statisztikai jellemzői tárja fel ennek mintázatait prekoncepciók nélkül.

3. Deduktív alkalmazások: Az osztályozás

Az osztályozási feladat (ld. még kategorizálás, klasszifikáció) során a korpusz szövegeit előre megadott osztályokba soroljuk be. Erre többek között azért lehet szükség, hogy a korpusz elemeit bizonyos szempont alapján szétválogassuk és kiszűrjük a számunkra releváns dokumentumokat. A klasszifikáció gyakorlati alkalmazására példaként az MTA TK-ban zajló *Comparative Agendas Project* (CAP) adatbázisának elemzését hozzuk fel. A CAP projekt különböző politikai intézmények (törvényhozás, kormány, média stb.) napirendjéről nyújt információt a releváns dokumentumok (pl. jogszabályok, újságcikkek) közpolitikai téma (pl. egészségügy, honvédelem) szerinti kategorizálásával.[10]

A projekt rendelkezik egy nemzetközi kódkönyvvel, amely lehetővé teszi a résztvevő országok adatainak összehasonlítását az elemzési egységek kategóriákba való besorolását követően. A következő táblázatban szereplő példák a projekt törvény- illetve rendeleti adatbázisból származnak (ld. 2. táblázat).

Rendelet /törvény neve	*CAP kódkönyv szerint kód*	*Kód megnevezése*
83/1990. (IV. 27.) MT rendelet a rendőrséggel és irányításával kapcsolatos egyes rendelkezések módosításáról	12	Igazságügy
90/1990. (XI. 17.) Korm. rendelet a bíróság által felülvizsgálható államigazgatási határozatokról szóló 63/1981. (XII. 5.) MT rendelet módosításáról	20	Kormányzati működés
A Magyar Köztársaság Alkotmányáról szóló 1949. évi XX. törvény módosításáról. Alcím: Kúria elnökének megválasztása	12	Igazságügy
A csődeljárásról és a felszámolási eljárásról szóló 1991. évi XLIX. törvény, a gazdasági társaságokról szóló 2006. évi IV. törvény, a cégnyilvánosságról, a bírósági cégeljárásról és a végelszámolásról szóló 2006. évi V. törvény, továbbá az ezekkel összefüggő egyes törvények módosításáról	15	Belkereskedelem[11]

2. táblázat
Példa a dokumentumkifejezés-mátrixra

A táblázat jól jelzi, hogy milyen módszertani nehézségekkel nézünk szembe ha egy dokumentumot egyértelműen be akarunk sorolni az adott kategóriák valamelyikébe (és csak abba az egybe). Egyetlen ilyen módszertani problémát kiemelve: a szövegek gyakran több témát is lefednek és több aspektusuk is van. Így például a 1990. évi 90. kormányrendelet egyszerre utal a bíróságokra és az államigazgatásra (a rendelet végül tényleges tartalma alapján kapta a kormányzati működéssel kapcsolatos kódot).

A kategóriákba való besorolás technikailag többféleképpen is kivitelezhető: kézzel, gépi támogatással vagy automatizáltan. A kézi kódolás egyik legfőbb előnye a gyakorlatban az, hogy a szöveg jól kiképzett kutatók/szakértők által történő értő olvasása jellemzően érvényes eredményeket ad. Másrészt nagyobb mennyiségnél a kategorizálás inkonzisztens lehet mivel a döntés pillanatában a kutató több szempontot is mérlegel és nem biztos, hogy minden esetben ugyanarra az eredményre jut. A szubjektivitásból adódó szinte elkerülhetetlen pontatlanság kezelhető a kettős vak kódolás alkalmazásával, melynek során két egymástól független szakértő sorolja be a megfigyelési egységeket és az eltérések esetén közösen határozzák meg a végleges besorolást.

Nagy mennyiségű adat esetében ugyanakkor ez az eljárás nem, vagy csak nagy költséggel alkalmazható. Az ilyen esetekben a gépi kódolás a leggyorsabb eljárási mód egy előre meghatározott szótár vagy a mesterséges intelligencia egy területe, a felügyelet gépi tanulás eljárásainak alkalmazásával. Ekkor a szűk keresztmetszetet az ún. tanulóhalmaz jelenti, melynek megfelelően besorolt dokumentumaiból a program „megtanulja" a kategorizálás logikáját melyet ezt követően képes a teljes korpuszra alkalmazni. E módszer hátránya, hogy továbbra is érdemi kézi erőfeszítésre épül (a tanulóhalmaz meghatározása során) és megfelelő programozási ismereteket igényel. Mindemellett az egyes klasszifikációs algoritmusok eltérő eredményeket hozhatnak, s így kiválasztásuk valamint az eredmények értelmezése részben ismét visszahozza az elemzői szubjektivitás problémáját.

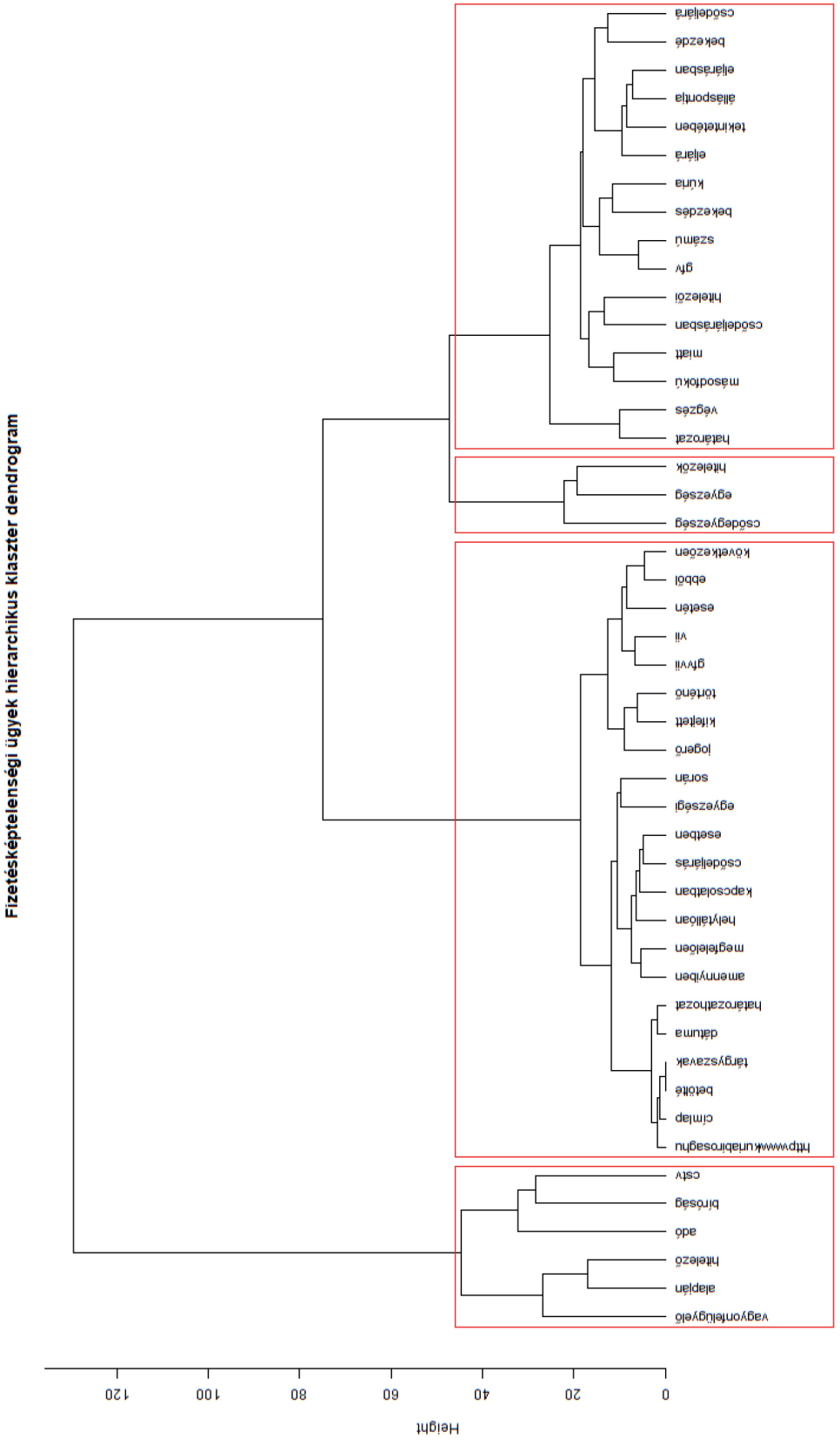
4. Induktív alkalmazások: A hierarchikus klaszterelemzés

Ha nincsen előzetes tudomásunk az adataink struktúrájáról, akkor a klaszterelemzés induktív eljárására támaszkodhatunk. irányt mutathat. E csoportosítási eljárás kivitelezésének két fontosabb eljárása a hierarchikus (felosztó vagy összevonó) illetve nem-hierarchikus megoldás (ld. alább). Mivel az elemzés kezdetekor nem ismerjük a korpusz tulajdonságait, így a kutatási kérdésből és tervből vezethető le, hogy hogyan állítsuk be a klaszterezés paramétereit. Ezek közül a legfontosabb döntés arra vonatkozik, hogy hány tematikus csoportot hozzunk létre a korpusz dokumentumaiból. A hierarchikus klaszterelemzés esetén az ún. dendrogram adhat útmutatást az optimális klaszterszám meghatározására. Ezek az ábrák azt mutatják meg, hogy

[10] Boda, Zs. – Sebők, M. (2015) Előszó: A Hungarian Comparative Agendas Project bemutatása. Politikatudományi Szemle, 2015/4. (33–40.) pp. 39.
[11] Az adatok forrása: cap.tk.mta.hu.

mely kifejezések vannak egymáshoz statisztikai értelemben legközelebb, s az őket tartalmazó dokumentumok így tematikusan egy csoportba tartoznak.

A felosztó klaszterelemzés esetén induláskor a korpusz összes dokumentuma egy nagy csoportba tartozik és az algoritmus iteratív módon osztja fokozatosan több csoportra a teljes halmazt. Összevonó klaszterelemzés esetén a korpusz minden eleme önálló csoportot határoz meg és az iterációk során az algoritmus az egymáshoz hasonlókat sorolja egy csoportba. A 4. ábrán található dendrogram a fizetésképtelenséggel kapcsolatos Kúria-határozatok 36 dokumentumának hierarchikus klaszterbontását tartalmazza, ahol az általunk meghatározott klaszterszám négy.

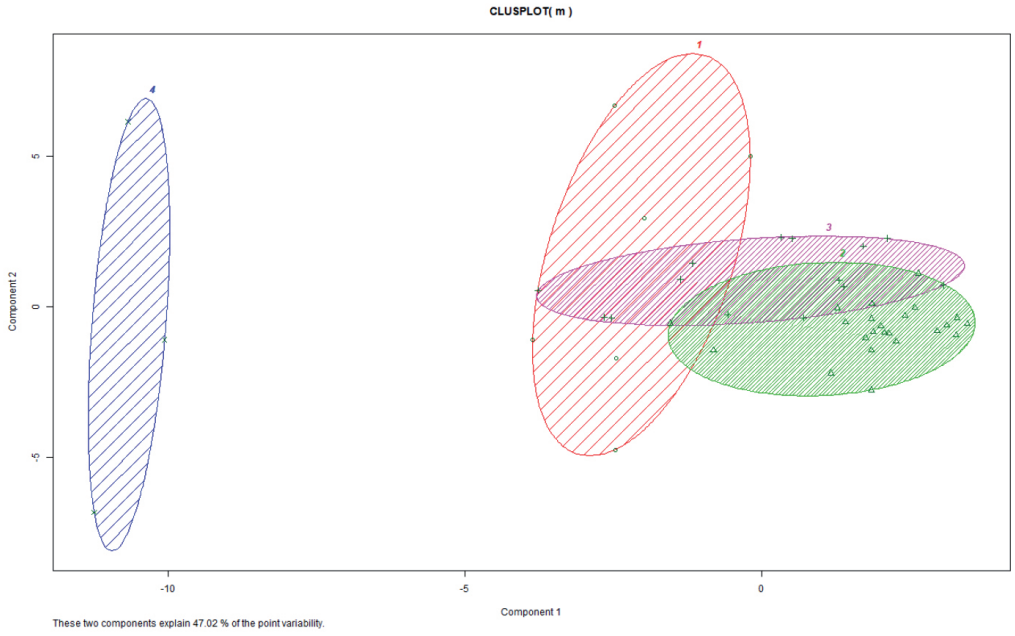


4. ábra
A fizetésképtelenségi határozatok hierarchikus klaszter dendrogramja

A 4. ábra azt mutatja, hogy milyen kulcsszavak jellemzik az algoritmus által létrehozott négy csoport dokumentumait (fontos hangsúlyozni, hogy az ábrán a kifejezések közötti távolságnak konkrét jelentése nincs, csak a kifejezések egymáshoz való viszonya számít). Az első klaszterben szereplő kifejezések – vagyonfelügyelő, hitelező, bíróság, adó – látszólag elkülönülnek a többi klasztertől tartalmilag és szemantikailag egyaránt. Tegyük fel, hogy olyan dokumentumokra vagyunk kíváncsiak, ahol a csődeljárás egyezséggel zárult. Ekkor a balról a harmadik klaszterbe sorolt dokumentumok között kell elsődlegesen keresnünk, míg a negyedik klaszter elemei a statisztikai elemzés alapján inkább az eljárási jellemzőket hangsúlyozzák (csődeljárás, másodfokú stb.). A klaszterelemzés technikai végrehajtása a szövegek adattá (egy vektortérré) való alakítására épül. E térben vizsgáljuk a dokumentumok szóhasználatuk tekintetében vett távolságát (a fentiekben ez az euklidészi távolság volt[12], de a gyakorlati elemzések során más megoldásokat is alkalmaznak).

5. Induktív alkalmazások: a K elemű klaszterelemzés

A szövegbányászati módszereket illusztráló mini esettanulmányunk végén azt vizsgáljuk, hogy a hierarchikus klaszterelemzés során kialakított négy csoport reprodukálható-e egy eltérő csoportosítási eljárás alkalmazásával. Ez a gyakran használt eljárás a k-közép klaszterelemzés, mely a hierarchikus klaszteranalízistől eltérő logikával dolgozik. Az optimális klaszterszám megválasztása itt is a kutató megítélésén múlik. Az algoritmus a klasztereket ezt követően úgy alakítja ki, hogy a dokumentum szóreprezentációs értékeinek egymástól mért (itt: euklideszi) távolságát veszi alapul. Az alkalmazás során először véletlenszerűen kiválasztunk K darab középpontot, majd minden dokumentumot hozzárendelünk a hozzá legközelebbi centrumhoz. Ez a folyamat addig ismétlődik, amíg a klaszterek el nem érnek egy bizonyos állandóságot, és már csak kevés dokumentum kerül át az újabb körökkel egy másik klaszterbe. Az 5. ábrán a fizetésképtelensági-korpuszunk K=4 k-közép klaszterelemzésének eredménye látható, míg a csoportokhoz tartozó leggyakoribb kifejezéseket a 3. táblázat tartalmazza.



5. ábra

A K-közép klaszterezés során a dokumentumok „pontjaival" létrehozott csoportok

Az 5. ábra alapján megállapítható, hogy a 4-es klaszter jól elkülönül a többi csoporttól, miközben a másik három csoport elég nagy átfedést mutat. A hierarchikus klaszterelemzés e 4-es klaszter elkülönülését nem jelzi ilyen egyértelműen (bár az ott szereplő egyik klaszter részhalmaza hasonló szövegeket fed le). Hasonló eredmény ugyanakkor, hogy meghatározható egy olyan csoport, amely általánosabb, minden jel szerint a Kúria eljárásával, a dokumentumok sztenderd szerkezetével kapcsolatos szavakat fed le, s nem a határozatok tartalmát azonosítja csoportlétrehozó szabályként.

Ez a probléma kezelhető az ilyen eljárási kifejezések a korpuszból való eltávolításával, hiszen az így megtisztított korpuszban már csak a tartalmi elemek maradnak, ami új csoportbeosztást eredményez. A két klaszterezési módszer összehasonlítása jól mutatja, hogy a gyakorlati alkalmazás során a szövegbányászati módszerek megválasztása kulcsfontosságú, mely gyakran

[12] E szerint két pont távolsága a pontok különbségének négyzetének gyökével egyezik meg.

	Klaszter1	Klaszter2	Klaszter3	Klaszter4
1	alapján	álláspontja	bekezdé	adó
2	csődegyezség	amennyiben	bekezdés	bíróság
3	egyezség	betölté	csődeljárá	cstv
4	hitelező	címlap	csődeljárásban	
5	hitelezők	csődeljárás	eljárá	
6	vagyonfelügyelő	dátuma	eljárásban	
7		ebből	határozat	
8		egyezségi	hitelezői	
9		esetben	kúria	
10		esetén	másodfokú	
11		gfv	miatt	
12		gfvvii	számú	
13		határozathozat	tekintetében	
14		helytállóan	végzés	
15		httpwwwkuriabirosaghu		
16		jogerő		
17		kapcsolatban		
18		kifejtett		
19		következően		
20		megfelelően		
21		során		
22		tárgyszavak		
23		történő		
24		vii		

3. táblázat
A K-közép klaszterezés csoportjainak leggyakoribb szavai

csak kísérletezéssel, több módszer kipróbálásával megoldható. Az ilyen elemzési döntések meghozásához a korábban bemutatott elemi-leíró statisztikai vizsgálatok fontos előkészítést jelenthetnek.

Utószó

A szövegbányászat módszerei a XXI. században a digitalizáció általános elterjedésével napról-napra nagyobb hangsúlyt kapnak az üzleti, közigazgatási és tudományos adatelemzésben egyaránt. Alkalmazásuk új távlatokat nyithat a joggyakorlat elemzésben számos kurrens dokumentum-kezelési folyamat automatizá-lásával és a közvetlen nem visszakereshető információk kinyerésével. Egy megfelelően kivitelezett szövegbányászati elemzés olyan látens struktúrákat képes feltárni a határozatszövegekben, melyeket manuálisan csak jelentős mértékű kézi munkaerő bevonásával vagy még így sem lehetne reprodukálni.

Kiterjedt alkalmazásukhoz ugyanakkor számos feltételhez kötött. A szövegbányászat kutatási tevékenység, helyes alkalmazásához így a nélkülözhetetlen adatbázis-kezelési és programozói tudás mellett kutatásmódszertani ismerek is szükségesek. Ugyanakkor hiába áll rendelkezésre a szükséges elemzői kapacitás és tudás, ha nincsenek az elemzésre megfelelően előkészített (pl. sztenderd karakterkódolású) szöveges dokumentum-gyűjtemények. A szövegbányászati eljárások által kínált elemzési potenciál kiaknázásához így - akár csak alkalmazásának más területein - megfelelő hardveres, szoftveres és kutatói infrastruktúra, valamint átgondolt, az elérni kívánt célokat világosan tisztázó fejlesztési projektek szükségesek.