



Eszterházy Károly
Főiskola



Kivonatoló program kontra emberi kivonatolás

Lengyelné dr. Molnár Tünde
Eszterházy Károly Főiskola



Kivonatoló program



- Magyar nyelvű offline kivonatoló program
 - kvantitatív tartalomelemzés
 - egységeit a szöveg szavai képezik,
 - az output: a szöveg mondatai
 - Jelenlegi állapot: elkészült a program első verziója
 - Tesztelés:
 - emberi kivonatokkal való összevetéssel

Content analysis has been defined as a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding (Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1985). Holsti (1969) offers a broad definition of content analysis as, "any technique for making inferences by objectively and systematically identifying specified characteristics of messages" (p. 14). Under Holsti's definition, the technique of content analysis is not restricted to the domain of verbal communication, but is also used in other areas such as coding student drawings (Wheelock, Haney, & Bebel, 2000), or coding of actions observed in videotaped child play (Schaefer, Kawaguchi, Knoff, & Serrano, 1992). In order to allow for replication, however, the technique can only be applied to data that are available in writing.

Content analysis enables researchers to sift through large volumes of data with relative ease in a systematic fashion (GAO, 1996). It can be a useful technique for allowing us to discover and describe the focus of individual, group, institutional, or social attention (Weber, 1985). It also allows inferences to be made which can then be corroborated using other methods of data collection. Krippendorff (1980) notes that "in much content analysis research, inferences are made by the search for techniques to infer from symbolic data what would be easier to determine, or longer possible, or too elaborate by the use of other techniques."

Practical Applications of Content Analysis

Content analysis can be a powerful tool for estimating authorship. For example, an investigator can begin with a list of suspected authors, examine their prior writings, and correlate the frequency of nouns or function words to help evaluate the relative probability of each person's authorship of the data at hand. Schaefer and Wallace (1964) used this basic technique to determine word patterns of 157 papers published by the authors of the Federalist papers; Krippendorff (1980) used a more holistic approach in order to determine the identity of the author of a letter in the 18th-century Primary Colors.

Content analysis is also useful for examining trends and patterns in documents. For example, Berelson (1952) conducted a content analysis of school curricula to determine "to make some inferences about the nature of the curriculum for education in the United States" (p. 10). Whether the goal is to determine the nature of a program, whether it is a government program, or whether it is a program of a private organization, content analysis can be a useful tool.

Content analysis provides an empirical basis for monitoring shifts in public opinion. For example, Berelson and Berelson (1952) conducted a content analysis of news stories to determine "the nature of the public's opinion of political events" (p. 10). Content analysis can also be used to monitor shifts in public opinion in a number of other areas, such as in the area of religion, education, and social issues.

Content analysis is also used to make valid inferences from the text. It is important that the researcher understand the procedure for making inferences from the text. Holsti (1969) notes, "To make valid inferences from the text, it is important that the researcher understand the procedure for making inferences from the text. People should code the same text in the same way" (p. 14). However, further notes, "reliability problems usually arise out of the ambiguity of the coding rules. Coding rules, or other coding rules" (p. 15). Holsti (1969) also notes that "the researcher should be able to explain to the people who have coded the text the coding rules and the procedure that they have established and the reasons for the coding. The key here is that the reliability coefficient they report is not a reliability coefficient (Holsti, 1969). In order to use data from a content analysis to make inferences, the researcher must be able to explain the coding rules and the procedure that they have established and the reasons for the coding. The key here is that the reliability coefficient they report is not a reliability coefficient (Holsti, 1969). In order to use data from a content analysis to make inferences, the researcher must be able to explain the coding rules and the procedure that they have established and the reasons for the coding. The key here is that the reliability coefficient they report is not a reliability coefficient (Holsti, 1969)."

Perhaps the most common method of content analysis research is that of content analysis. This simply means doing a word frequency count. The assumption made is that the words that are used in the text are the words that reflect the overall concepts. While this may be true in some cases, there are several caveats. First, content analysis is a descriptive method, not an inferential method. It is used to make inferences about numbers of occurrences.

The primary goal of content analysis is to determine what may be used for statistical analysis. In Holsti's (1969) definition, a document is "a collection of words that are used in the text. The researcher must be able to explain to the people who have coded the text the coding rules and the procedure that they have established and the reasons for the coding. The key here is that the reliability coefficient they report is not a reliability coefficient (Holsti, 1969). In order to use data from a content analysis to make inferences, the researcher must be able to explain the coding rules and the procedure that they have established and the reasons for the coding. The key here is that the reliability coefficient they report is not a reliability coefficient (Holsti, 1969)."

Content analysis is also used to determine the nature of a program, whether it is a government program, or whether it is a program of a private organization, content analysis can be a useful tool. Content analysis provides an empirical basis for monitoring shifts in public opinion. For example, Berelson and Berelson (1952) conducted a content analysis of news stories to determine "the nature of the public's opinion of political events" (p. 10). Content analysis can also be used to monitor shifts in public opinion in a number of other areas, such as in the area of religion, education, and social issues. Content analysis is also used to make valid inferences from the text. It is important that the researcher understand the procedure for making inferences from the text. Holsti (1969) notes, "To make valid inferences from the text, it is important that the researcher understand the procedure for making inferences from the text. People should code the same text in the same way" (p. 14). However, further notes, "reliability problems usually arise out of the ambiguity of the coding rules. Coding rules, or other coding rules" (p. 15). Holsti (1969) also notes that "the researcher should be able to explain to the people who have coded the text the coding rules and the procedure that they have established and the reasons for the coding. The key here is that the reliability coefficient they report is not a reliability coefficient (Holsti, 1969). In order to use data from a content analysis to make inferences, the researcher must be able to explain the coding rules and the procedure that they have established and the reasons for the coding. The key here is that the reliability coefficient they report is not a reliability coefficient (Holsti, 1969)."

Text Mining

Kivonatolás

Szótár használat



Kivonatoló File Beállítások Kilépés

Luhn-módszer systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of

• Szótáralap

Szótár tulajdonságainak beállítása

Szótár kiválasztása

Alapértelmezett Egyéni

Kérem adja meg, az Ön által kiválasztott gyakorisági szótárban mely sorszámú szavak kerüljenek figyelembe vételre!

minimuma:

maximuma:

OK

Output



Kivonatoló

File Beállítások Kilépés

Mentés Nyomtatás

Content analysis has been defined as a systematic, reliable technique for compressing many words of text into a fewer content-bearing (Berelson, 1952; GAO, 1956; Krippendorff, 1980; and Wiersma, 1990). Holsti (1969) offers a broad definition of content analysis

10 A következő meghatározás is mutatja, hogy nincs éles határ a kivonat és a referátum között.
11 Egy szöveg szavaiból képzett előfordulási lista négy részre osztható.
12 A jövőben véleményem szerint egyre több gyakorisági szótárral fogunk találkozni, melyek közül néhánynak az alapját talán a Magyar Nemzeti Szövegtár
13 szolgáltatja.
14 Berelson a szisztematikus és mennyiségi leírására alkalmas kutatási technikákat tekinti tartalomelemzésnek.
15 Berelson a tartalomelemzést a következőképpen definiálja: a kommunikáció valóságos tartalmának objektív, szisztematikus és mennyiségi leírására
16 alkalmas kutatási technika.
17 Akkor használjuk, ha az egységek a vizsgálat szempontjából heterogének.
18 A nem véletlen mintavételi eljárások alapján levont következtetésekből nem általánosíthatunk.
19 Zipf-törvénye alapján a szignifikáns szavak listájának meghatározása során figyelembe kell venni az adott terület kifejezéseiből készített gyakorisági
20 szótárakat.
21 Viszont ezt a kategóriát nehéz figyelembe venni a későbbi elemzés során.
22 Mint ahogy emítettem és az eddig ismertetett szótárak is mutatják a szépirodalomtól eltérő tudományághoz kapcsolódó átfogó gyakorisági szótár
23 egyenlőre nem létezik magyar nyelven.
24 Nem az explicit tartalom vizsgálata a cél, hanem a szövegekörnyezet kontextusából szeretnénk következtetéseket levonni.
25 A függelékben található gyakorisági szótár a 10-nél többször előforduló szavakat tartalmazza.
26 A gyakorisági vizsgálatok nem állnak meg a szöveg szavaiból képzett előfordulási csoportok felállításának szintjén.
27 Míg a korábbi vizsgálatok jobbra kvantitatív elemzések voltak, a II. világháború idején a propaganda-elemzés vált meghatározóvá.
28 Az elmúlt években ugrásszerűen megnőtt a gyakorisági szótárak készítése, kiadása, és egyre több szótárt építő vállalkozással találkozhatunk, azonban
29 szakterülethez kapcsolódó átfogó gyakorisági szótár továbbra sem létezik magyar nyelven.
30 A szöveg szavainak szótöveiből képzett szógyakorisági lista felállítását a szöveget jellemző szignifikáns szavak meghatározása követi.
31 A szóhasználat gyakorisági értékeinek meghatározásához a 150 millió szót tartalmazó Magyar Nemzeti Szövegtárat vették alapul.
32 Ha a szöveget manuálisan előkészítjük és a következő módon történik a gépelésük, akkor a problémák nagy része kiküszöbölhető: Rövidített szavak
33 eredeti formájában történő begépelése
34 A dokumentációs nyelvészeti csoport korszakát (1967-1971) az 1966-ban, az MTA Számítóközpontjában, a gépi nyelvészeti munkacsoportból átalakult
35 Dokumentációs Nyelvészeti Csoport munkája jellemzi.
36 Az eredeti dokumentum ismerete nélkül is érthető.
37 A mondat meghatározások is több szempontot kell figyelembe venni.

Kivonatoló



Eszterházy Károly
Főiskola



Felmérés

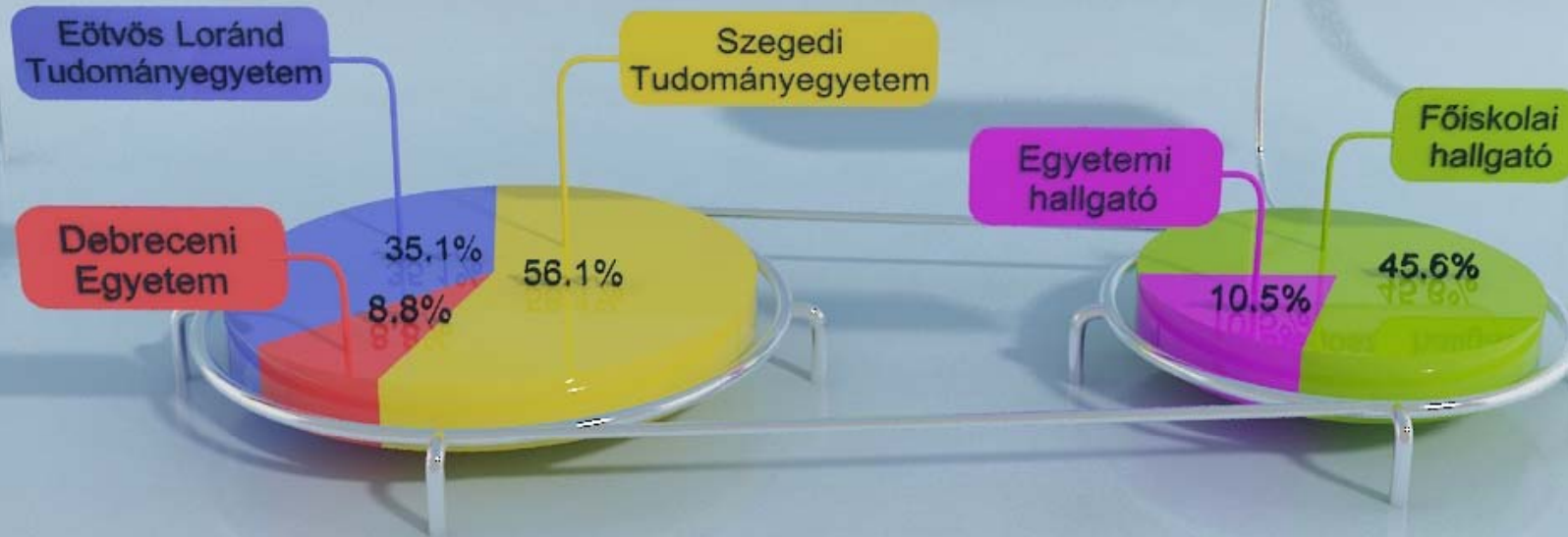


A felmérés alapja

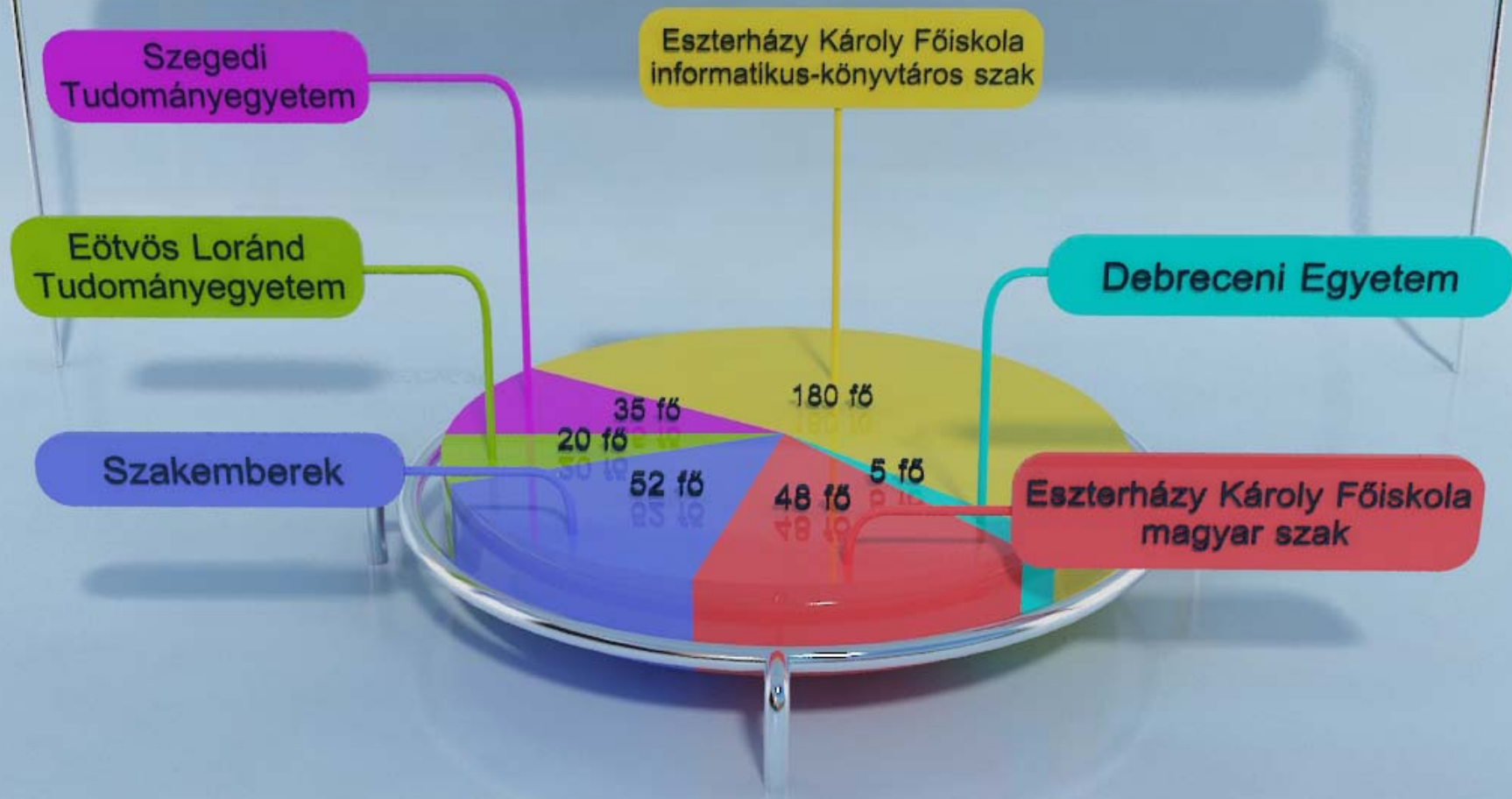


- KOLTAY Tibor:
Szöveg, információ, relevancia: néhány adalék a témakörhöz
- PROKNÉ Palik Mária:
A tartalmi feltárás problémái online könyvtári katalógusokban
- Kitöltés módja:
 - online
 - papír alapú
- Feladat:
 - 20%-s tömörítés → 17-17 releváns mondat kiválasztása és rangsorolása
 - Kulcsszó megjelölés (korlát nélkül)

A részt vevő egyetemek megoszlása a felmérésben szereplő hallgatói létszám alapján



A felmérésben részt vevők megoszlása





Súlyozás



- **Borda-módszer alapján**
 - A rangsorbeli első helyezetteket **n** súllyal vettem figyelembe,
 - a másodikat **$n-1$** -s értékkel,
 -
 - az utolsó helyezett **1** -s súllyal fog szerepelni az értékelésben.



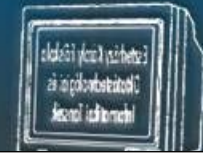
Eszterházy Károly
Főiskola



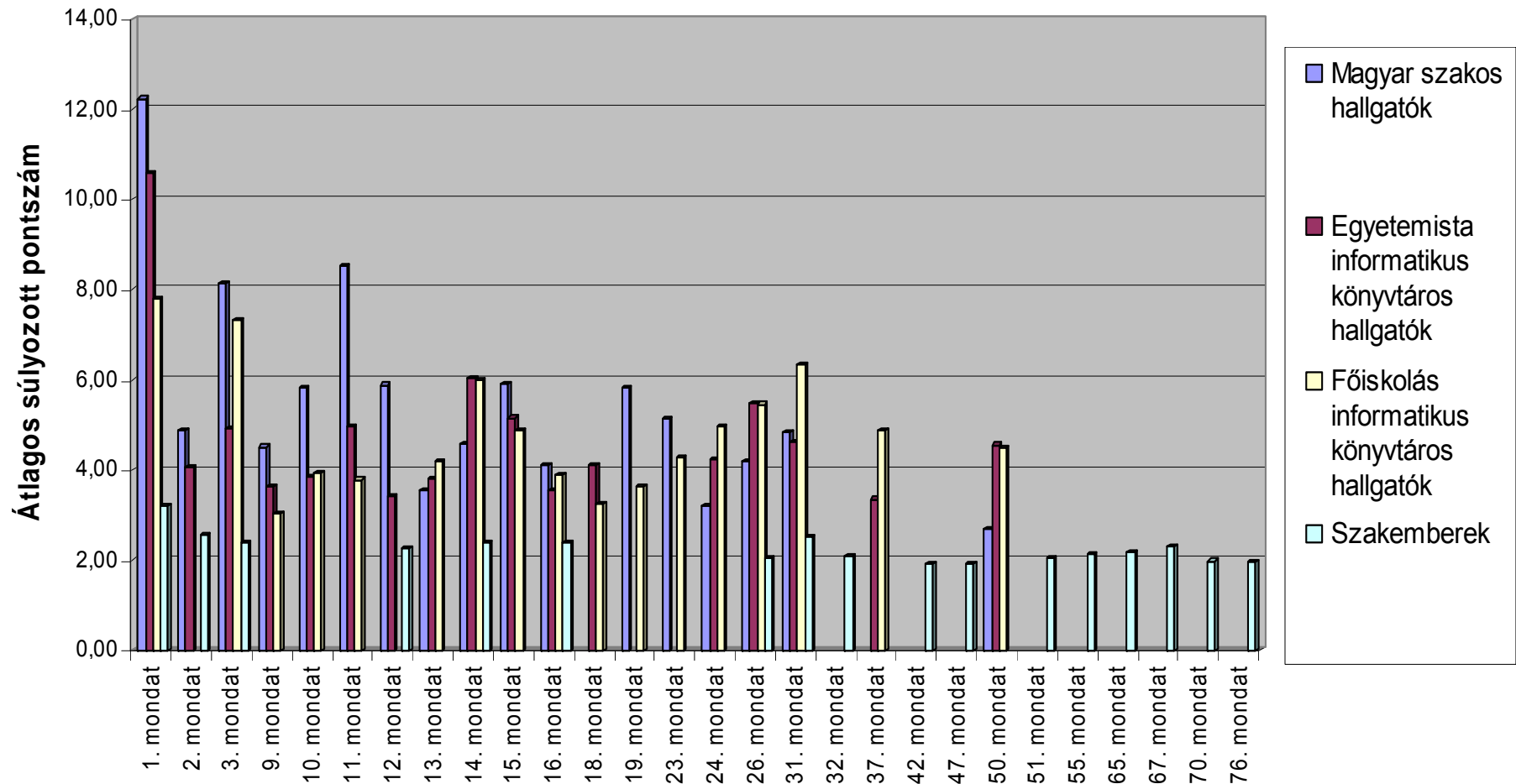
KOLTAY Tibor: *Szöveg, információ, relevancia: néhány adalék a témakörhöz*



Sorrend	A főiskolások legrelevánsabb mondatai	Az egyetemisták legrelevánsabb mondatai
1.	1. mondat	1. mondat
2.	3. mondat	14. mondat
3.	31. mondat	26. mondat
4.	14. mondat	15. mondat
5.	26. mondat	11. mondat
6.	24. mondat	3. mondat
7.	37. mondat	31. mondat
8.	15. mondat	50. mondat
9.	50. mondat	24. mondat
10.	23. mondat	18. mondat
11.	13. mondat	2. mondat
12.	10. mondat	10. mondat
13.	16. mondat	13. mondat
14.	11. mondat	9. mondat
15.	19. mondat	16. mondat
16.	18. mondat	12. mondat
17.	9. mondat	37. mondat



A kivonat mondatai a négy mintacsoportnál





Súlyozott pontszámokból képzett korrelációs mátrix



	Informatikus– könyvtáros egyetemista hallgatók	Informatikus –könyvtáros főiskolás hallgatók	Szakemberek	Magyar szakos hallgatók
Informatikus könyvtáros egyetemista hallgatók	1			
Informatikus könyvtáros főiskolás hallgatók	0,917	1		
Szakemberek	0,486	0,445	1	
Magyar szakos hallgatók	0,872	0,856	0,441	1



Kivonatok mondatainak eloszlása



A szöveg	A kivonat eloszlása Koltay Tibor cikke esetén				Szerző
	Egyetemista informatikus könyvtáros hallgatók	Főiskolás informatikus könyvtáros hallgatók	Szak-emberek	Magyar szakos hallgatók	
I. negyede	70,59%	64,71%	35,29%	70,59%	46,67%
II. negyede	23,53%	29,41%	23,53%	23,53%	40,00%
III. negyede	5,88%	5,88%	23,53%	5,88%	6,67%
IV. negyede	-	-	17,65%	-	6,67%



Egyezés a szerző kivonatával



- A minta 340 tagja közül 7 fő jelölte meg ugyanazon mondatokat mint a szerző
- A mintacsoportok súlyozott kivonatait alapul véve az egyezés:
 - Egyetemista informatikus könyvtáros hallgatók esetén **8 mondat**
 - Főiskolás informatikus könyvtáros hallgatók esetén **8 mondat**
 - Szakemberek esetén **5 mondat**
 - Magyar szakos hallgatók esetén **7 mondat**



Számítógépes output



- A két módszer esetén 10 mondat egyezik meg.
- A mintacsoportokkal az egyezés:

<i>Mintacsoportok</i>	Luhn módszere alapján		Szószablya szótár alapján	
	Egyező monda- tok száma	Egyezés aránya	Egyező monda- tok száma	Egyezés aránya
Informatikus könyvtáros egyetemista hallgatók	3	17,65%	3	17,65%
Informatikus–könyvtáros főiskolás hallgatók	4	23,53%	3	17,65%
Szakemberek	4	23,53%	6	35,29%
Magyar szakos hallgatók	3	17,65%	3	17,65%
Szerző kivonata	5	29,41%	3	17,65%



Eszterházy Károly
Főiskola



PROKNÉ Palik Mária: *A tartalmi feltárás problémái online könyvtári katalógusokban*



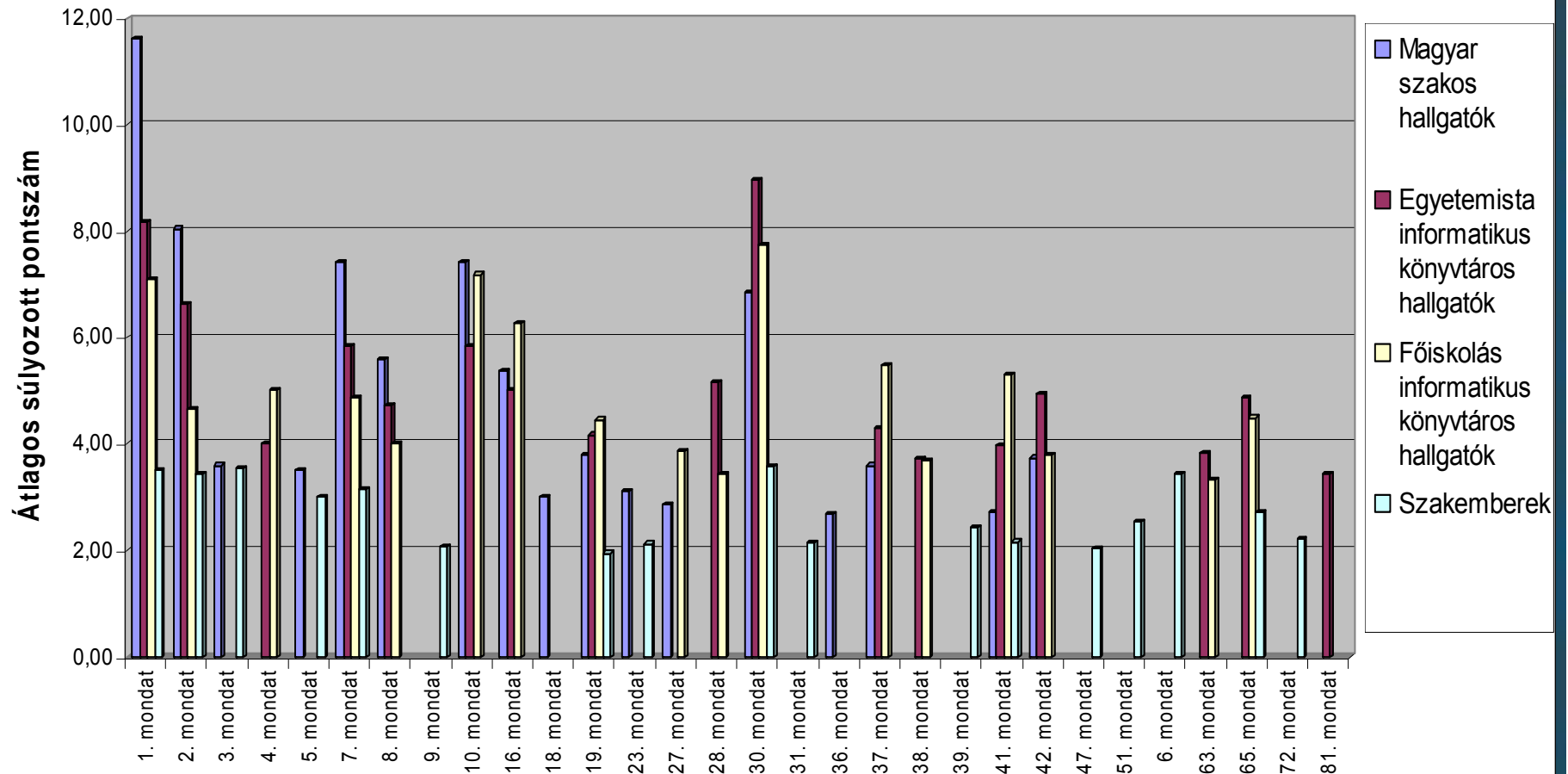
Sajátosságok



- Az egyetemista és főiskolás informatikus-könyvtáros hallgatók kivonata 94%-ban megegyezik
- Az első bekezdés 4-4 illetve 3-3 mondata megtalálható a mintacsoportok kivonatában
- A szakembereknél eltérőek a vélemények:
 - 52 szakember 37 különböző mondatot tett első helyre
 - egyetlen mondat kapott 25% feletti jelölést (a hallgatóknál 80, illetve 70%-70% fölötti)



A kivonat mondatai a négy mintacsoportnál





Korrelációs mátrix a megjelölt mondatok alapján

	Informatikus könyvtáros egyetemista hallgatók	Informatikus könyvtáros főiskolás hallgatók	Szakemberek	Magyar szakos hallgatók
Informatikus könyvtáros egyetemista hallgatók	1			
Informatikus könyvtáros főiskolás hallgatók	0,936	1		
Szakemberek	0,526	0,471	1	
Magyar szakos hallgatók	0,878	0,849	0,576	1



Kivonatok mondatainak eloszlása



A szöveg	A kivonat eloszlása Prokné Palik Mária cikke esetén				
	Egyetemista informatikus könyvtáros hallgatók	Főiskolás informatikus könyvtáros hallgatók	Szak-emberek	Magyar szakos hallgatók	Szerző
I. negyede	47,06%	47,06%	47,06%	58,82%	5,88%
II. negyede	23,53%	29,41%	23,53%	29,41%	29,41%
III. negyede	11,76%	11,76%	17,65%	11,76%	23,53%
IV. negyede	17,65%	11,76%	11,76%	0,00%	41,18%

A szerző többször nem 1-1 mondatot jelöl meg, hanem 1-1 bekezdésnyi részt.



Egyezés a szerző kivonatával



- Összesen **7** olyan fő van a 340 mintatag közül, akiknek a kivonata 9 mondatban megegyezik a szerző kivonatának mondataival (sorrend elhagyásával)
- A mintacsoportok súlyozott kivonatait alapul véve az egyezés:
 - Egyetemista informatikus könyvtáros hallgatók kivonatával **6 mondat**
 - Főiskolás informatikus könyvtáros hallgatók kivonatával **6 mondat**
 - Szakemberek kivonatával **2 mondat**
 - Magyar szakos hallgatók kivonatával **4 mondat**



Számítógépes output



- A két módszer esetén 11 mondat egyezik meg.
- A mintacsoportokkal az egyezés:

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



Véggövetkeztetés



- Nincs két ember, aki ugyanazt a referátumot állítaná elő.
 - Bár a 340 felmérésben részt vevő személy közül 7 illetve 6 személy van, akinek mondatai megegyeznek, de eltérő sorrendben
 - kijelenthető, hogy nagyon kevés személy ítéli meg egyformán a cikkek lényeges mondatait.

- Nincs nagy betűs KIVONAT

**Kivonatoló program kontra emberi kivonatolás:
DÖNTETLEN**



Eszterházy Károly
Főiskola



Köszönöm a figyelmet!

mtunde@ektf.hu