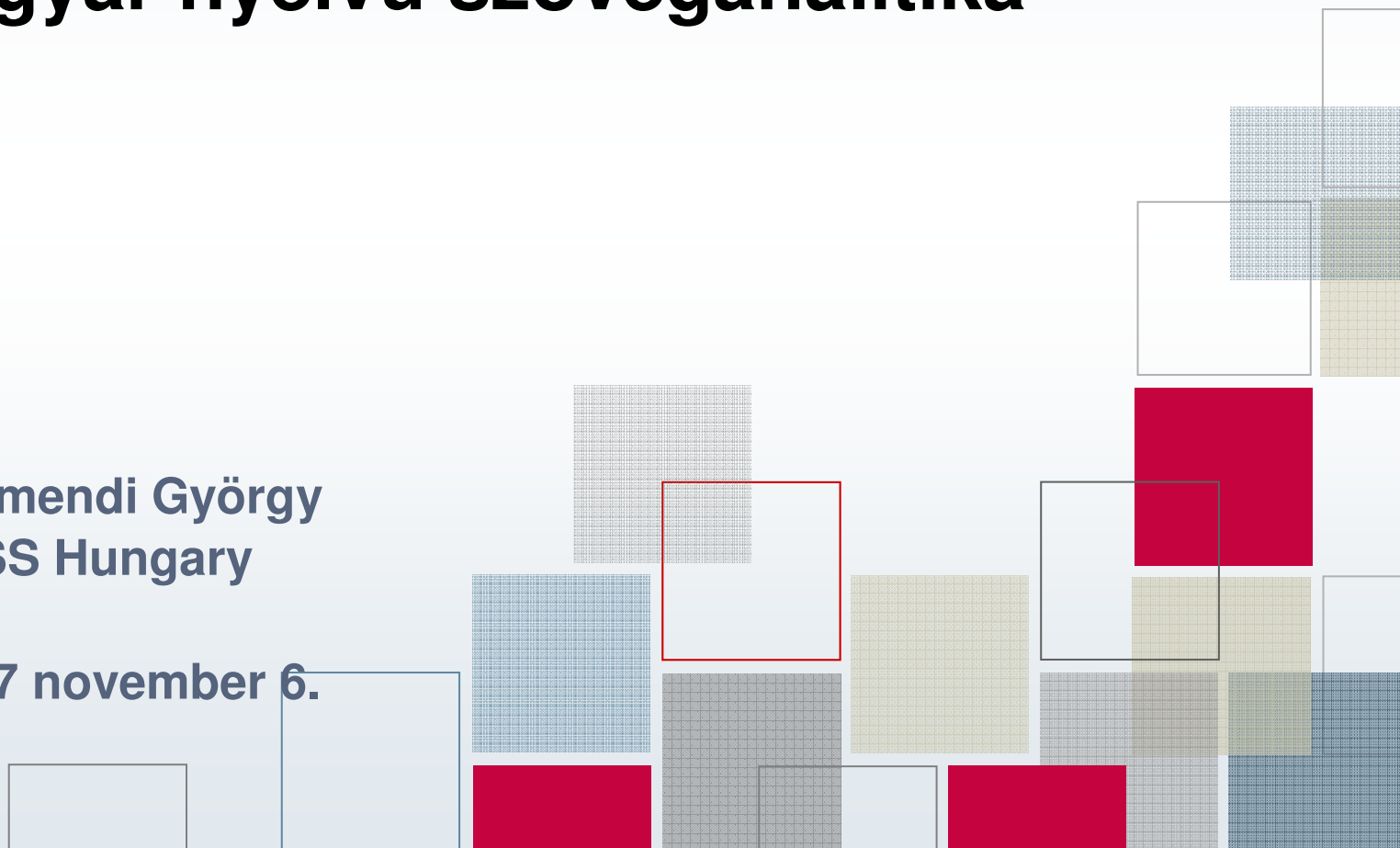


> Magyar nyelvű szöveganalitika

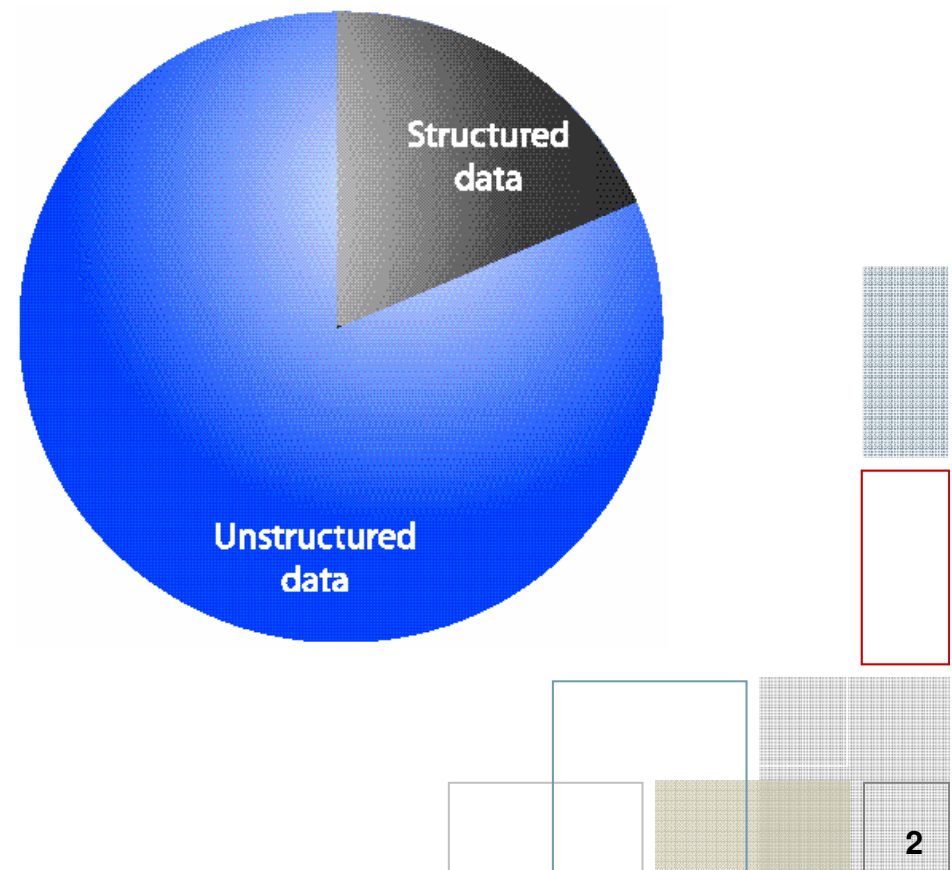
Körmendi György
SPSS Hungary

2007 november 6.



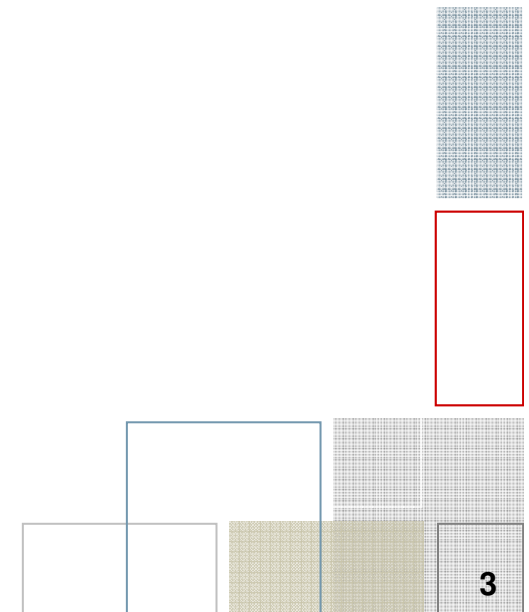
> Mit várunk a szöveganalitikától?

- A vállalatoknál tárolt információk 80%-a struktúrátlan adatokban /IDC/
- Információ kinyerése!
- Hol tart?



> Mit NEM tárgyalunk?

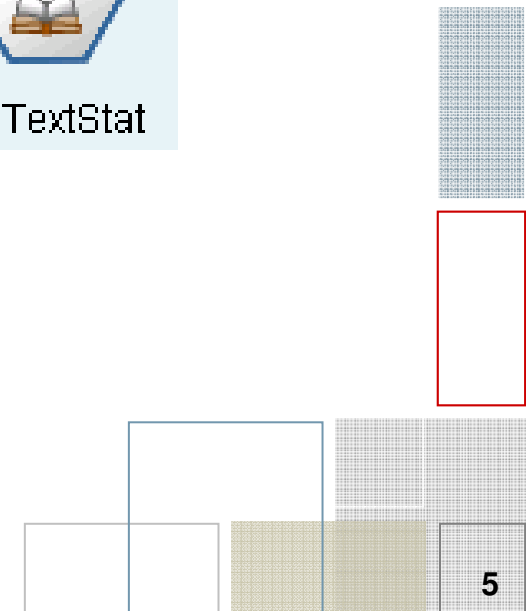
- Keresés
- Speciális feladatok (pl. spamszűrés)



>Text analitikai eszköztár TMFC 5.0



> Web feed node



> Web feed node

Buvos Szakacs
 Refresh
 URL: **RSS** http://buvosszakacs.blog.hu/rss2
 Source Preview

Title: Népek morzsái
 Short Description:
 Author: MBTBD
 Contributors:
 Published Date: Wed Oct 31 17:09:00 CET 2007
 Modified Date:
 Description:

A bűvös szakács - Népek morzsái
 2007.10.31. 17:09
 A paníroz szó szerint annyit tesz, „meg- kenyerezni” (panis, pain, pan, pão = kenyér). Ez a gasztronómiai találmány először az aranyo
 -->
 A bejegyzés trackback címe:
<http://blog.hu/goldmine/hedgehog193.php/214443Töröld ki az URL belsejéből a .php előtt található háromjegyű számot!>
 Kommentek, trackbackek, visszapingek:
 Remy
 Nagy erővel keresem a régen beígért olajokról szóló cikket, vagy még nincs is ilyen?
 Először tanulj meg panírozni. ;)

Kedves Remy, a cikk még készül, addig is figyelmébe ajánliuk a Piszkos aranyak c. cikket.

Record start tag:

Field	HTML Tag

> Web feed node

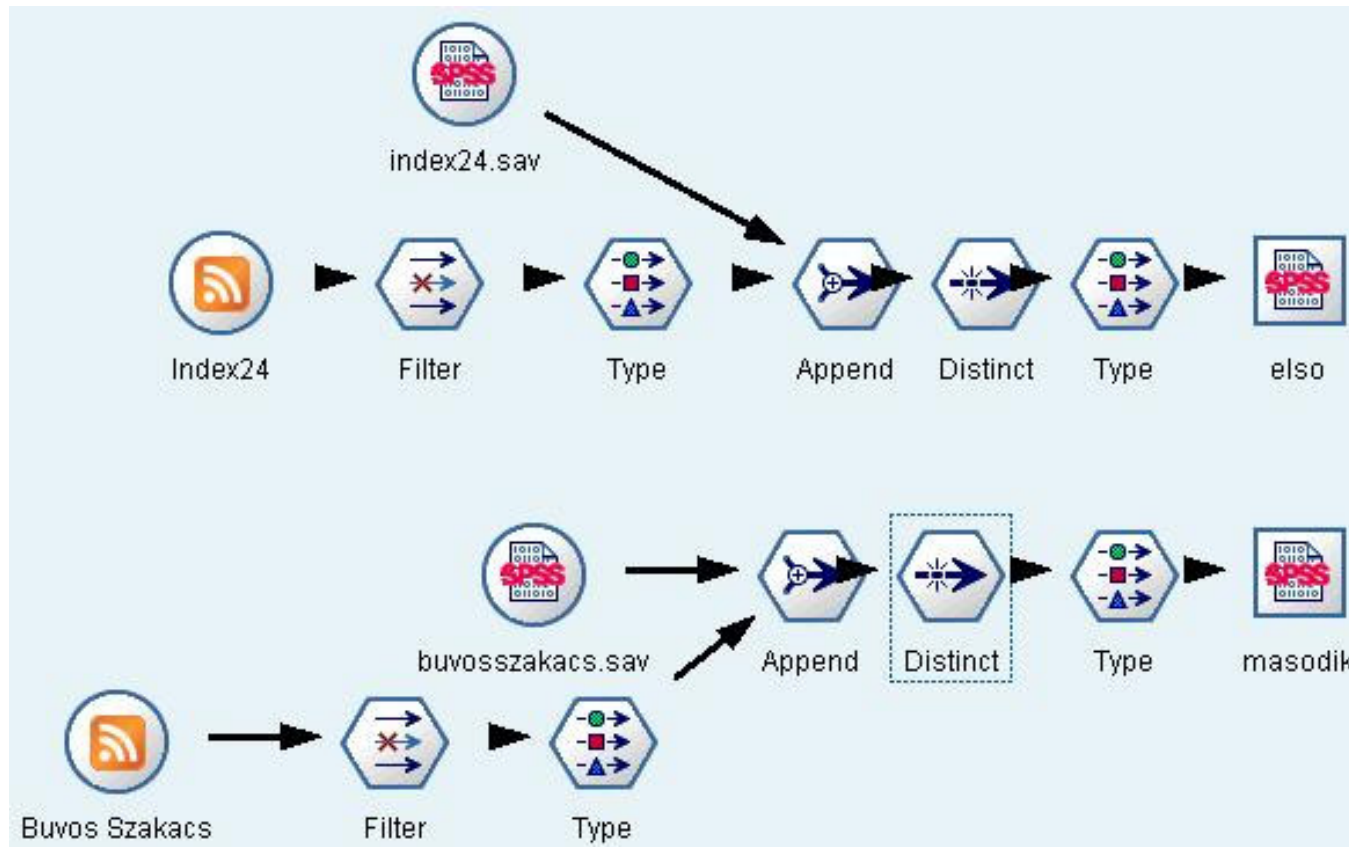
Table (7 fields, 50 records)

	Title	Short_Description	Description
1	Nobu	A londoni Old Park Lane-en fekv? Metro...	A b?vös szakács - Nobu2007.08.11. 11:03 A londoni Old Park Lane-en fekv? Metropolitan
2	Szaracén desszert	Ha Firenzében tizennyolc lakásból áll e...	A b?vös szakács - Szaracén desszert2007.08.09. 13:47 Ha Firenzében tizennyolc lakásb
3	Toszkán hentések	2006. január 25-én Firenzében vörös s...	A b?vös szakács - Toszkán hentések2007.08.08. 20:53 2006. január 25-én Firenzében v
4	Itt van Góliáth	Muskotály Küvé hozott Góliáth-paradics...	A b?vös szakács - Itt van Góliáth2007.08.07. 15:09 Muskotály Küvé hozott Góliáth-paradic
5	A lécs és a tudat	T?z a nap. Nagy bajuszú, népies visele...	A b?vös szakács - A lécs és a tudat2007.08.06. 11:37 T?z a nap. Nagy bajuszú, népies vis
6	Gary Danko az Ikarusban	A San Franciscó-i Gary Danko menüi s...	A b?vös szakács - Gary Danko az Ikarusban2007.08.04. 09:14 A menüi szerepelnek egé
7	San Francisco és Gary Danko	A 19. század közepén kitört aranyláz ide...	A b?vös szakács - San Francisco és Gary Danko2007.08.04. 08:39 A 19. század közepén
8	Plachutta	Bécs belvárosában van a Wollzeile nev...	A b?vös szakács - Plachutta2007.08.03. 16:35 Bécs belvárosában van a Wollzeile nev? u
9	Táfelspicc és társai	„A tányérhús szaft nélküli hús” - véleke...	A b?vös szakács - Táfelspicc és társai2007.08.03. 15:56 „A tányérhús szaft nélküli hús” -
10	Kedves Látogató!	Néhány napra elutaztunk. Úti cél: a salz...	A b?vös szakács - Kedves Látogató!2007.07.31. 19:18 Néhány napra elutaztunk. Úti cél: a
11	Konyha határok nélkül	Szakácskongresszus lesz New Yorkba...	A b?vös szakács - Konyha határok nélkül2007.07.29. 09:34 Szakácskongresszus lesz Ne
12	Burgonyahelyzet	Anélkül, hogy részletesebben kitérnék...	A b?vös szakács - Burgonyahelyzet2007.07.27. 10:49 Anélkül, hogy részletesebben kitérn
13	Madridfúsió 2007	Ötödször rendezték meg 2007-ben a „...	A b?vös szakács - Madridfúsió 20072007.07.26. 13:22 Ötödször rendezték meg 2007-b
14	Gasthaus Ubl	Ha a bécsi Naschmarktnál a Rechte W...	A b?vös szakács - Gasthaus Ubl2007.07.26. 06:06 Ha a bécsi Naschmarktnál a Rechte W
15	Mesclun	A „mesclun” szó a provanszál mesclu...	A b?vös szakács - Mesclun2007.07.25. 09:40 A „mesclun” szó a provanszál mesclumób
16	Pinotxo	A katalán Ferran Adria, a világ egyik leg...	A b?vös szakács - Pinotxo2007.07.24. 15:41 A katalán Ferran Adria, a világ egyik legism
17	Barcelona és a tapasbárok	A tapas „kis adag ingyencség, f?étkezés...	A b?vös szakács - Barcelona és a tapasbárok2007.07.21. 15:26 A tapas „kis adag ingyenc
18	A salzburgi Ikarus	Mi történik, ha társul egy Ausztr...	A katalán Ferran Adria, a világ egyik legismertebb szakácsa a tengerparti Rosesben vezet...
19	Roland Trettl tavaszi menüi	A salzburgi Ikarus étterembe	éttermét, melyet minden évben hat hónapra bezár. Ilyenkor visszavonul, hogy néhány az egymás
20	N?nek a paradicsomok	Miután egyik bloglátogatónk,	munkatársával kidolgozza a következ? év több tucat apró fogásból álló menüjét. M?helye - látogatónk
21	A Taillevent	A hatvanas években Mexicó	melyet sokan alkímista boszorkányko

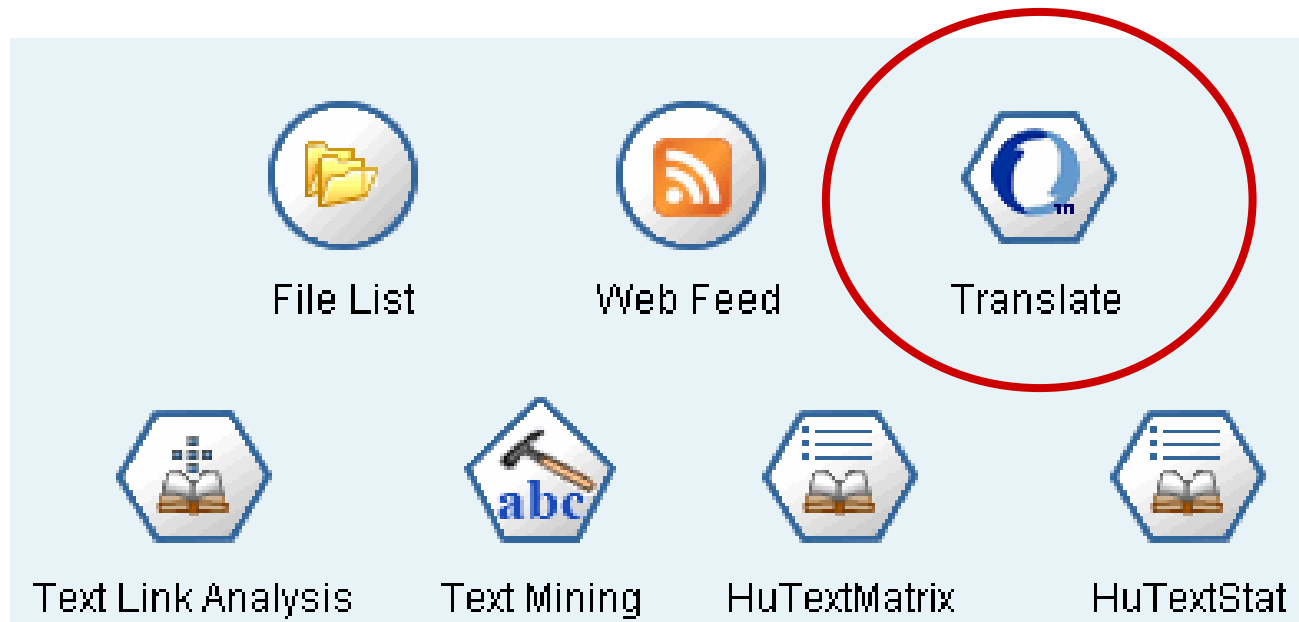
Table Annotations

OK

> Web feed node



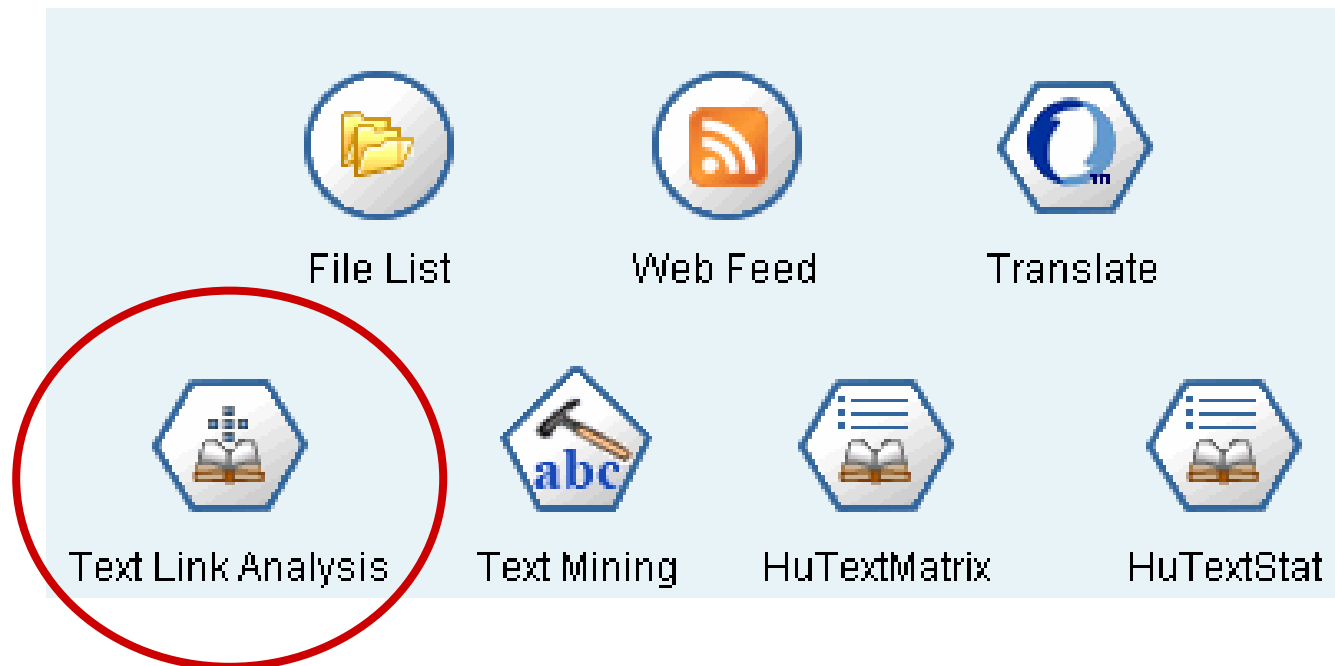
> Statisztikai gépi fordító – Language Weaver



> A textanalitika központi eleme



> Különálló Text Link Analysis



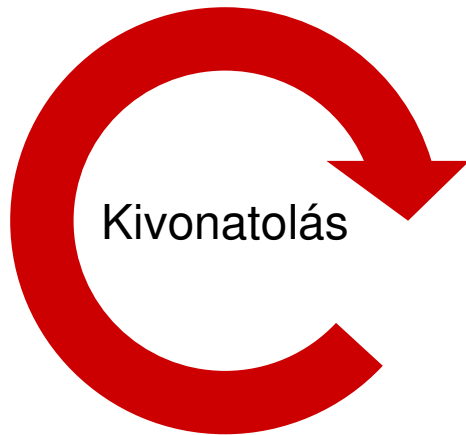
> Magyar nyelvű kivonatolás



> Szemantikus szöveganalitika



Szótárak, Könyvtár



Kivonatolás

Az **ipafai papnak fapipája van.**
A **Clementine** a **legjobb**
adatbányász szoftver.

Nyelvi minták azonosítása

> A textanalitika központi eleme



>Szótár

Interactive Text Mining of hozzaszolas

File Edit View Generate Tools Help

All Libraries

Resource Editor

Auto

- Hun_Hun
- OpinionHun

Term	Match	Inflect	Type	Library
4x4	Entire Term	<input type="checkbox"/>	Jellemzok	Hun_Hun
a	Entire Term	<input type="checkbox"/>	Kot	Hun_Hun
abbahtagyott	Entire Term	<input checked="" type="checkbox"/>	ige	Hun_Hun
abbahtagytuk	Entire Term	<input checked="" type="checkbox"/>	ige	Hun_Hun
ablakos	Entire Term	<input type="checkbox"/>	Jellemzok	Hun_Hun
ablaktörl?	Entire Term	<input type="checkbox"/>	Alkatesz	Hun_Hun
ablaktörl?d	Entire Term	<input type="checkbox"/>	Alkatesz	Hun_Hun
ablaktörl?kar	Entire Term	<input type="checkbox"/>	Alkatesz	Hun_Hun
ablaktörl?t	Entire Term	<input type="checkbox"/>	Alkatesz	Hun_Hun
abroncsokkal	Entire Term	<input type="checkbox"/>	Alkatesz	Hun_Hun
abrüzi	Entire Term	<input type="checkbox"/>	Szemely	Hun_Hun
abs	Entire Term	<input type="checkbox"/>	Felszereltség	Hun_Hun
accord	Entire Term	<input type="checkbox"/>	Auto	Hun_Hun
ad	Entire Term	<input checked="" type="checkbox"/>	ige	Hun_Hun
adac	Entire Term	<input type="checkbox"/>	Organization	Hun_Hun
addisonnak	Entire Term	<input type="checkbox"/>	Szemely	Hun_Hun
addisonon	Entire Term	<input type="checkbox"/>	Szemely	Hun_Hun
addisont	Entire Term	<input type="checkbox"/>	Szemely	Hun_Hun
adi	Entire Term	<input checked="" type="checkbox"/>	ige	Hun_Hun

Target	Synonyms	Library
ablaktörlő	ablaktörl?, ablaktörl?d, ablaktörl?kar, ablaktörl?t	Hun_Hun
ad	ad, adj, adja, adják, adnak, adni, adok, adom, adott, adtak, adtam	Hun_Hun
akar	akarja, akarják, akarna, akarnak, akarnam, akarod, akarok, akarom, akarsz	Hun_Hun

Synonyms Optional

2 Libraries 23 Types 889 Terms 181 Excludes 540 Synonyms 0 Optional

Switch Resources

Template	Owner	Version	Date	Annotation	TLA
Basic Resources (Dutch)	kormendi	1	máj.-10-2007 16:40		
Competitive Intelligence (English)	kormendi	1	ápr.-12-2007 15:33		+
CRM (English)	kormendi	1	ápr.-12-2007 18:01		
Basic Resources (English)	kormendi	1	jún.-04-2007 23:11		
Gene Ontology (English)	kormendi	1	ápr.-12-2007 15:40		
Genomics (English)	kormendi	1	ápr.-12-2007 15:42		+
IT (English)	kormendi	1	ápr.-12-2007 15:47		
MeSH (English)	kormendi	1	ápr.-12-2007 15:49		
Opinions (English)	kormendi	1	jún.-16-2007 10:53		+
Security Intelligence (English)	kormendi	1	ápr.-12-2007 16:46		+
Basic Resources (French)	kormendi	1	márc.-10-2007 18:58		
Basic Resources (German)	kormendi	1	máj.-10-2007 15:21		
Basic Resources (Italian)	kormendi	1	febr.-12-2007 10:53		
CRM (Portuguese)	kormendi	1	febr.-12-2007 11:12		
Basic Resources (Portuguese)	kormendi	1	febr.-12-2007 10:54		
Basic Resources (Spanish)	kormendi	1	márc.-06-2007 15:19		
Security Intelligence (Spanish)	kormendi	1	jún.-04-2007 22:59		+
Film	kormendi	2	okt.-31-2007 17:10		+
Auto	kormendi	1	okt.-31-2007 12:35		+

Select Cancel Help

>Text Link Analysis

Edit Advanced Resources

File Edit View Help

Session Library Patterns

Use POS patterns from:
Hun_Hun

Use Text Link Analysis patterns from:
Hun_Hun

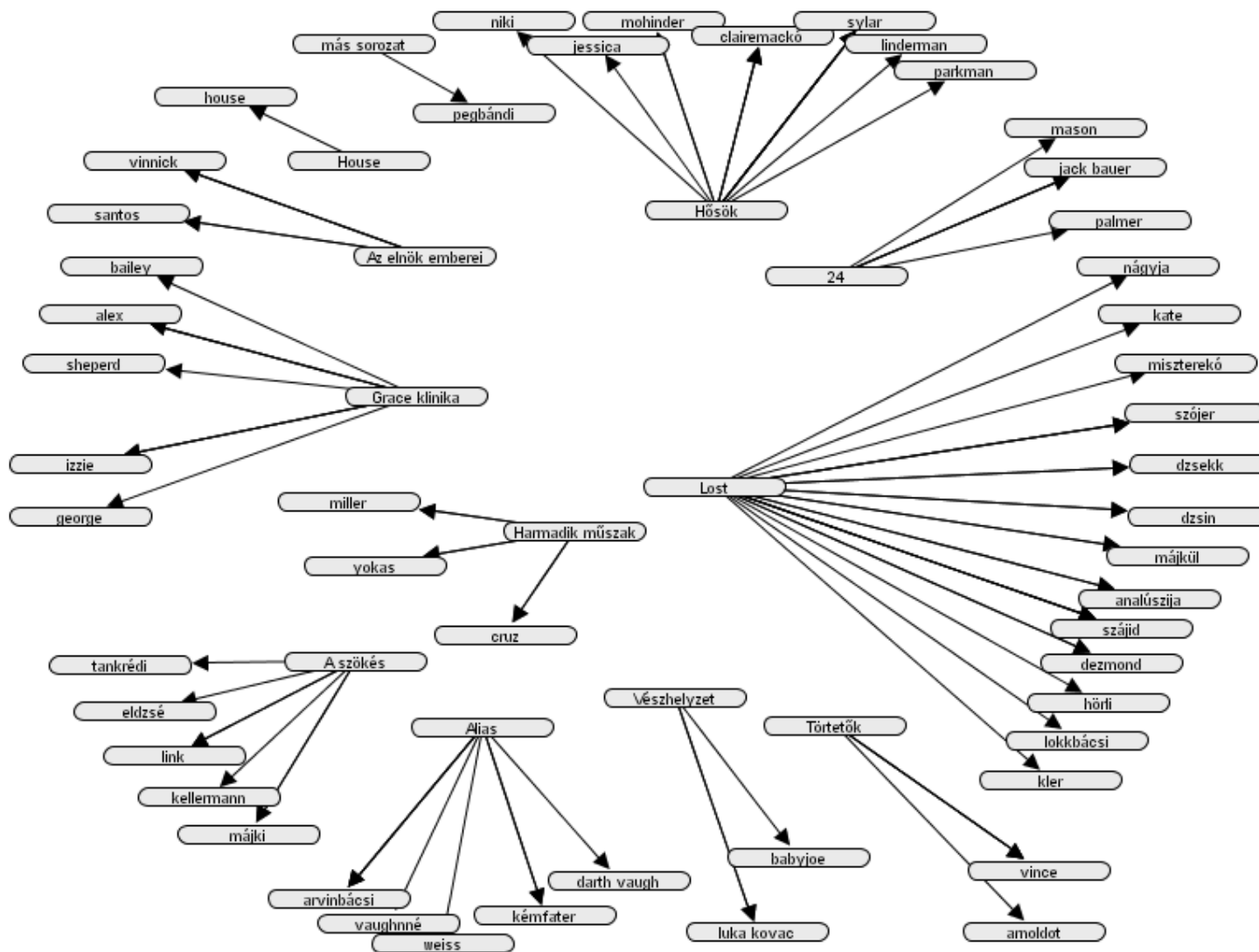
- Hun_Hun
 - Dynamic POS Patterns
 - Forced Definitions
 - Text Link Analysis

```
#  
[pattern(6)]  
name=6  
value= $vNegative @(0,5) $vIge ($vAik|$vAuto)  
output=$1t#1t$3t#3  
#  
[pattern(7)]  
name=7  
value= $vAuto @(0,5) $vLegek  
output=$1t#1t$3t#3  
#  
[pattern(8)]  
name=8  
value= $vAuto @(0,5) $vJel  
output=$1t#1t$3t#3  
#  
[pattern(9)]  
name=9  
value= $vAuto @(0,5) $vAik @(0,10) $vNegative?  
output=$1t#1t$3t#3t$5t#5  
#  
[pattern(10)]  
name=10  
value= ($vPositive|$vNegative) @(0,5) $vHas @(0,5) ($vPositive|$vNegative)  
output=$1t#1t$3t#3t$5t#5t$7t#7  
#  
[pattern(11)]  
name=11  
value= $vAuto @(0,10) $vSzarm  
output=$1t#1t$3t#3
```

> Magyar példák

The screenshot displays the SPSS Clementine 11.1 software interface. The main workspace shows a workflow starting with a 'Table' node connected to a 'sorozatcim' node. Below this, a 'sorozat_blog.sav' file is loaded into a 'Description' node. The workflow continues through several nodes: 'Filler', '(generated)', 'Description', 'Type', 'Filter', 'Distinct', and another '(generated)' node. A 'Link Analysis' node is also present. The right-hand side of the interface features a 'Clementine' panel with 'Streams' and 'Outputs' tabs, listing 'Stream1', 'huntex_demo3', and 'sorozatok_elemzes1'. Below this is a 'CRISP-DM' panel with a 'Classes' tree showing stages like 'Business Understanding', 'Data Understanding', 'Data Preparation', 'Modeling', 'Evaluation', and 'Deployment'. The bottom of the window includes a toolbar with various modeling tools and a status bar showing 'Server: Local Server' and '183MB / 342MB'.

>Magyar példák



>Text link analysis

Interactive Text Mining of Description

File Edit View Generate Categories Tools Help

Text Link Analysis

47 patterns

Global	Slot1 Type	Slot2 Type	Slot3 Type
9	<Szemely>	<kul>	<Unknown>
7	<Unknown>	<Kot>	<Szemely>
7	<Unknown>	<Ige>	
7	<Szemely>	<Szereplo>	
6	<Positive>	<Szemely>	
6	<Szemely>	<Ige>	<Szemely>
6	<Szereplo>	<Negative>	
6	<Szereplo>	<Unknown>	
6	<Negative>	<Szereplo>	
5	<Ige>	<Szereplo>	
5	<Szemely>	<Negative>	
5	<Szereplo>	<Positive>	

5 Selected: 5 patterns

Global	Docs	Slot1 Concep	Slot2 Concep	Slot3 Concep
1	1	luka	tuti	
1	1	pegbándi	jó	
1	1	analúszija	kedves	
1	1	májkül	szép	
1	1	yokas	jó	

Concept Web

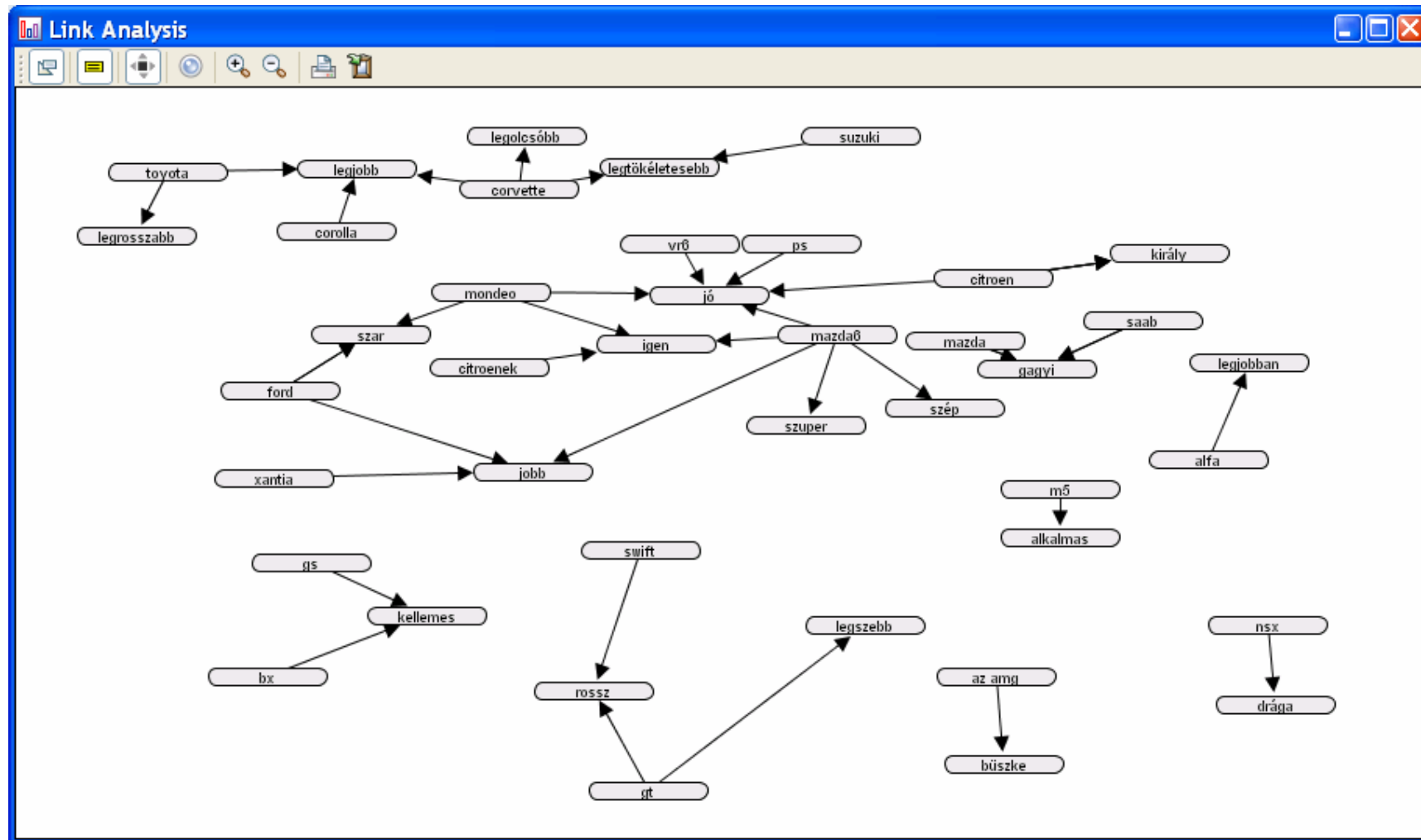
Positive Szereplo

Global Count

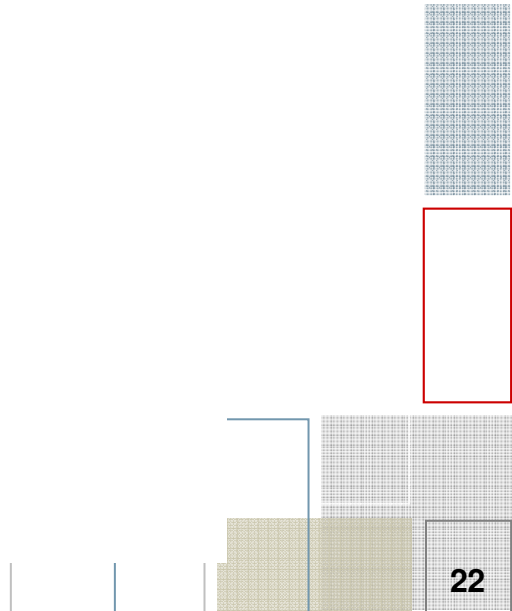
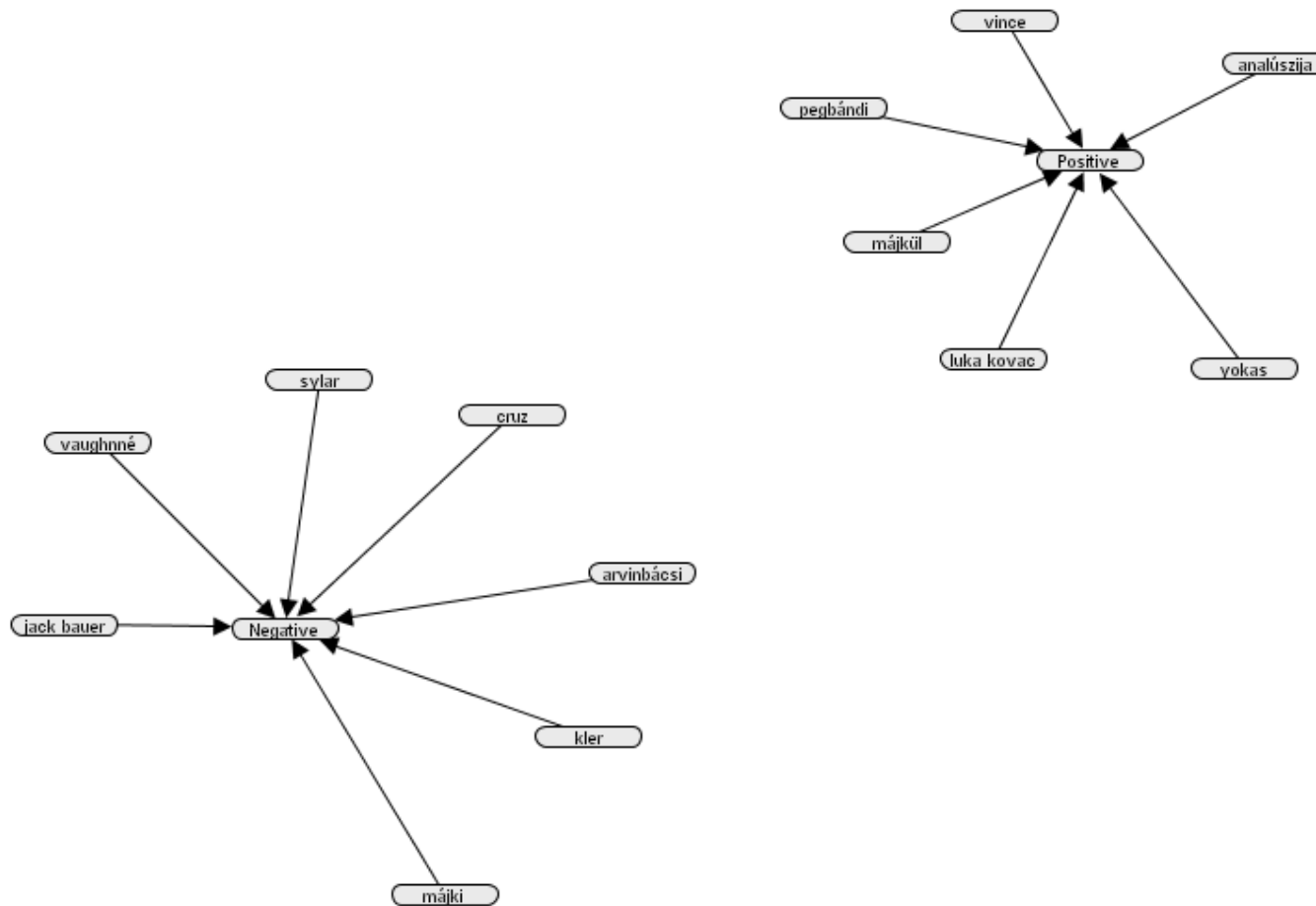
2,0
1,5
1,0
0,5
0,0

To populate data pane:
Make a selection in a table and click Display

> Magyar példák



> Magyar példák



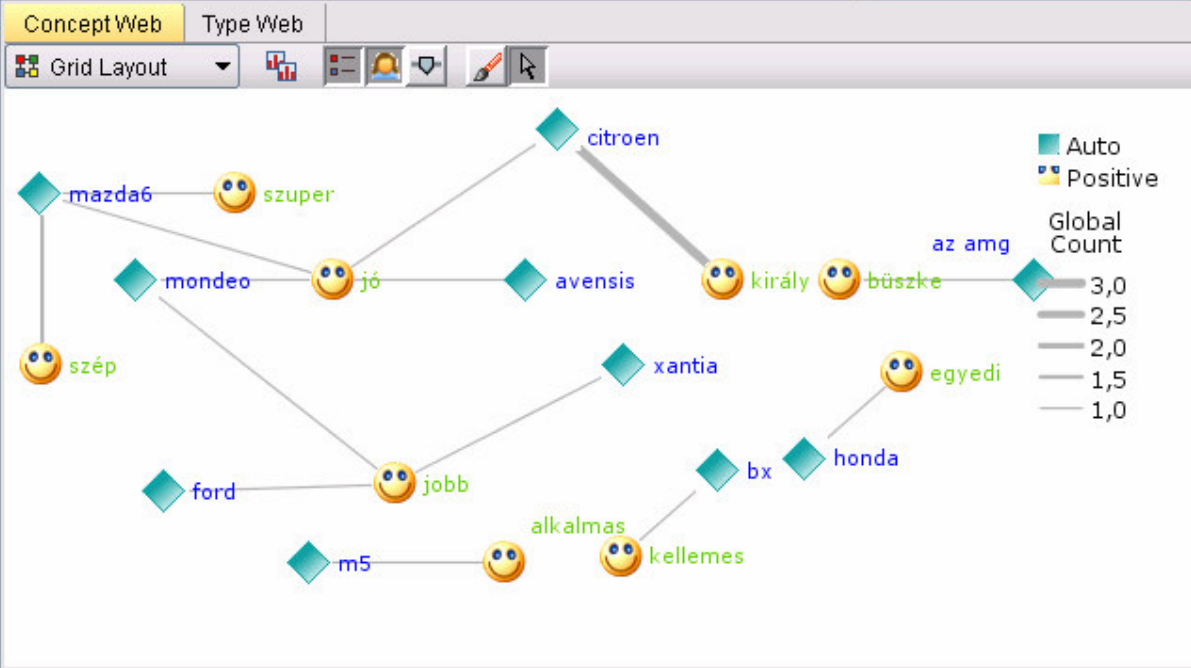
Interactive Text Mining of Path

File Edit View Generate Categories Tools Help

Text Link Analysis

Extract 8 patterns

Global	Slot1 Type	Slot2 Type
17	<Auto>	<Positive>
8	<Auto>	<Negative>
8	<Auto>	<Legek>
2	<Product>	<Positive>
1	<Auto>	<Person>
1	<Unknown>	<Negative>
1	<Negative>	<Ige>
1	<Organization>	<Szemely>



Extract Selected: 14 patterns

Global	Docs	Slot1 Concept	Slot2 Concept
3	3	citroen	király
2	2	mazda6	szép
1	1	xantia	jobb
1	1	mondeo	jó
1	1	az amg	büszke
1	1	citroen	jó
1	1	ford	jobb
1	1	avensis	jó
1	1	honda	egyedí
1	1	mazda6	szuper
1	1	mazda6	jó
1	1	bx	kellemes
1	1	mondeo	jobb
1	1	m5	alkalmas

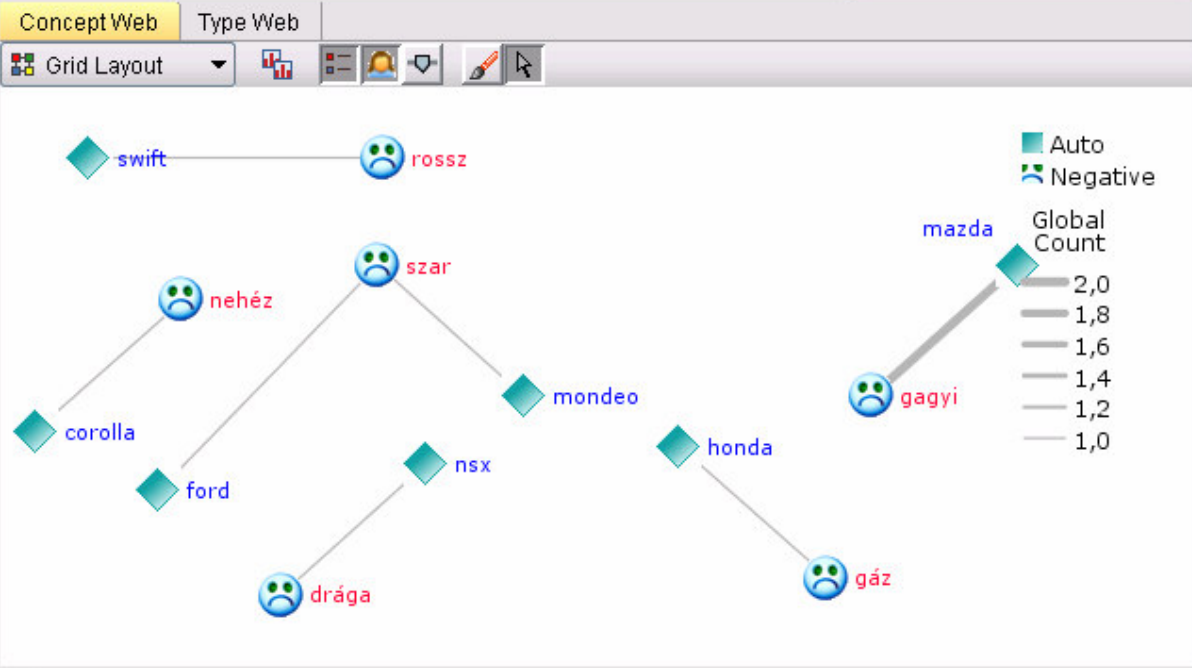
To populate data pane:
Make a selection in a table and click Display

Extract [Icons] 8 patterns Display

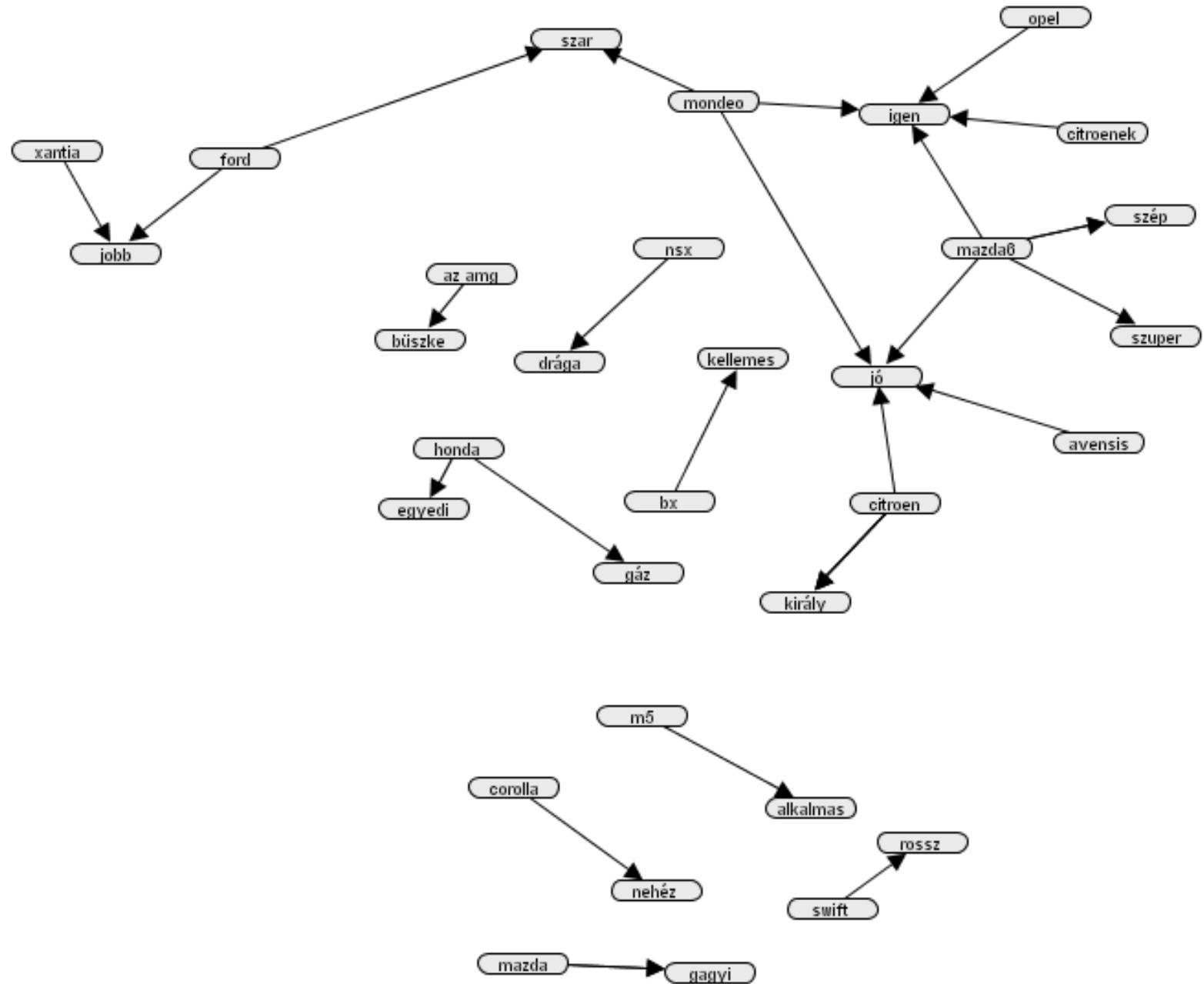
Global	Slot1 Type	Slot2 Type
17	<Auto>	<Positive>
8	<Auto>	<Negative>
8	<Auto>	<Legek>
2	<Product>	<Positive>
1	<Auto>	<Person>
1	<Unknown>	<Negative>
1	<Negative>	<Ige>
1	<Organization>	<Szemely>

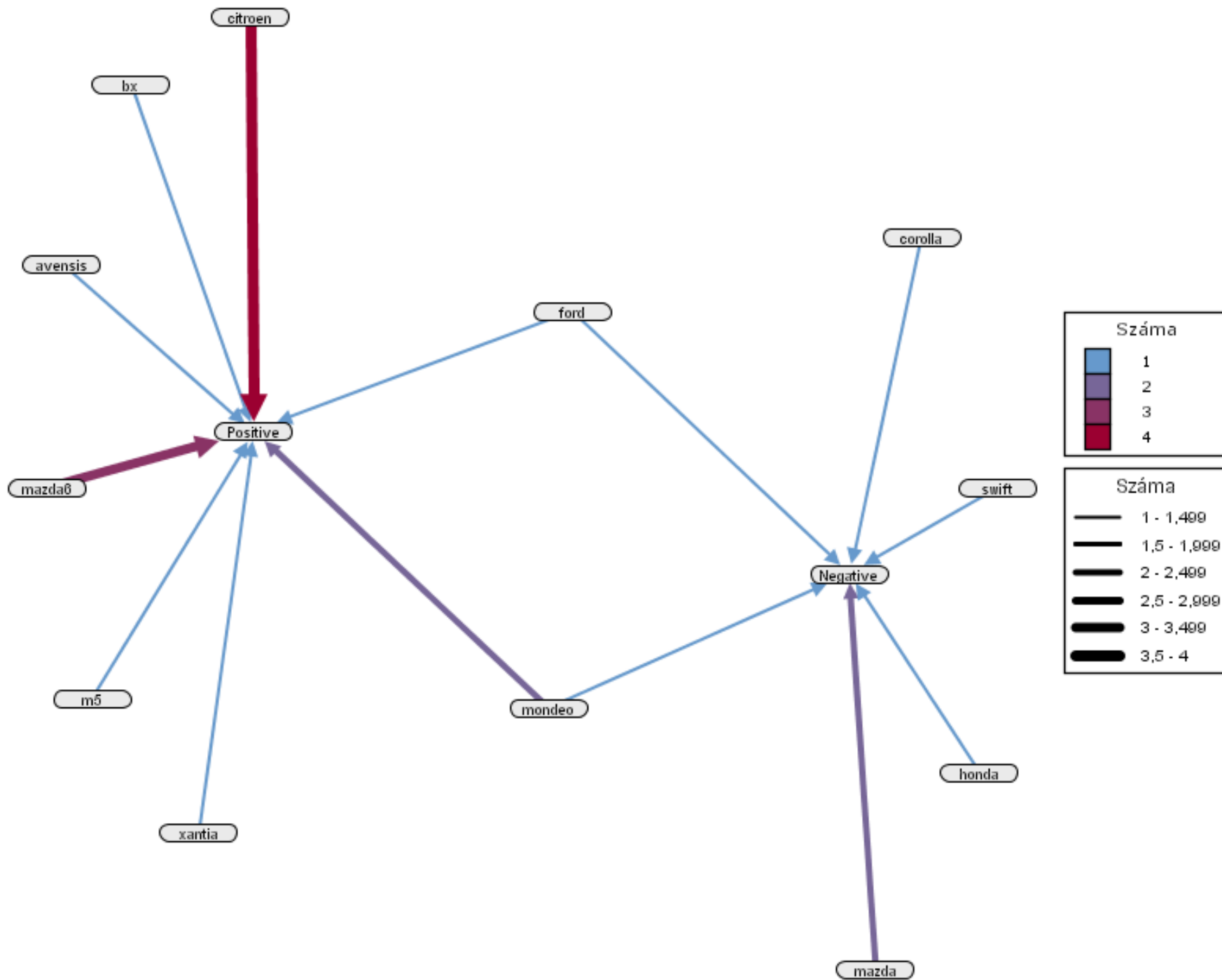
Extract [Icons] Selected: 7 patterns Display

Global	Docs	Slot1 Concept	Slot2 Concept
2	2	mazda	gagyi
1	1	honda	gáz
1	1	ford	szar
1	1	mondeo	szar
1	1	nsx	drága
1	1	corolla	nehéz
1	1	swift	rossz



To populate data pane:
Make a selection in a table and click Display





>TMC 3.0 Advanced Features (5) Link Analysis & Pattern Matcher

- The pattern matcher allows you to find relationships between concepts identified during the text extraction process.

Examples:

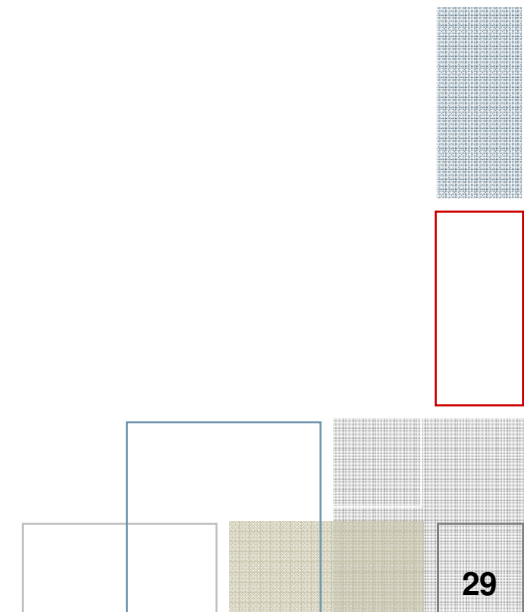
- Bioinformatics: Gene1 <inhibits> Gene2
 - CRM: Customer1 <unhappy> handset
 - Homeland security: person <member of> organization
- TMC 3.0 gives you access to the pattern matcher via the Text Link Analysis node
 - CEMI node – load through CEMI dialogue
 - Projects involving the pattern matcher should involve the Text Mining Task Force (contact Olivier Jouve)



Text Link Analysis

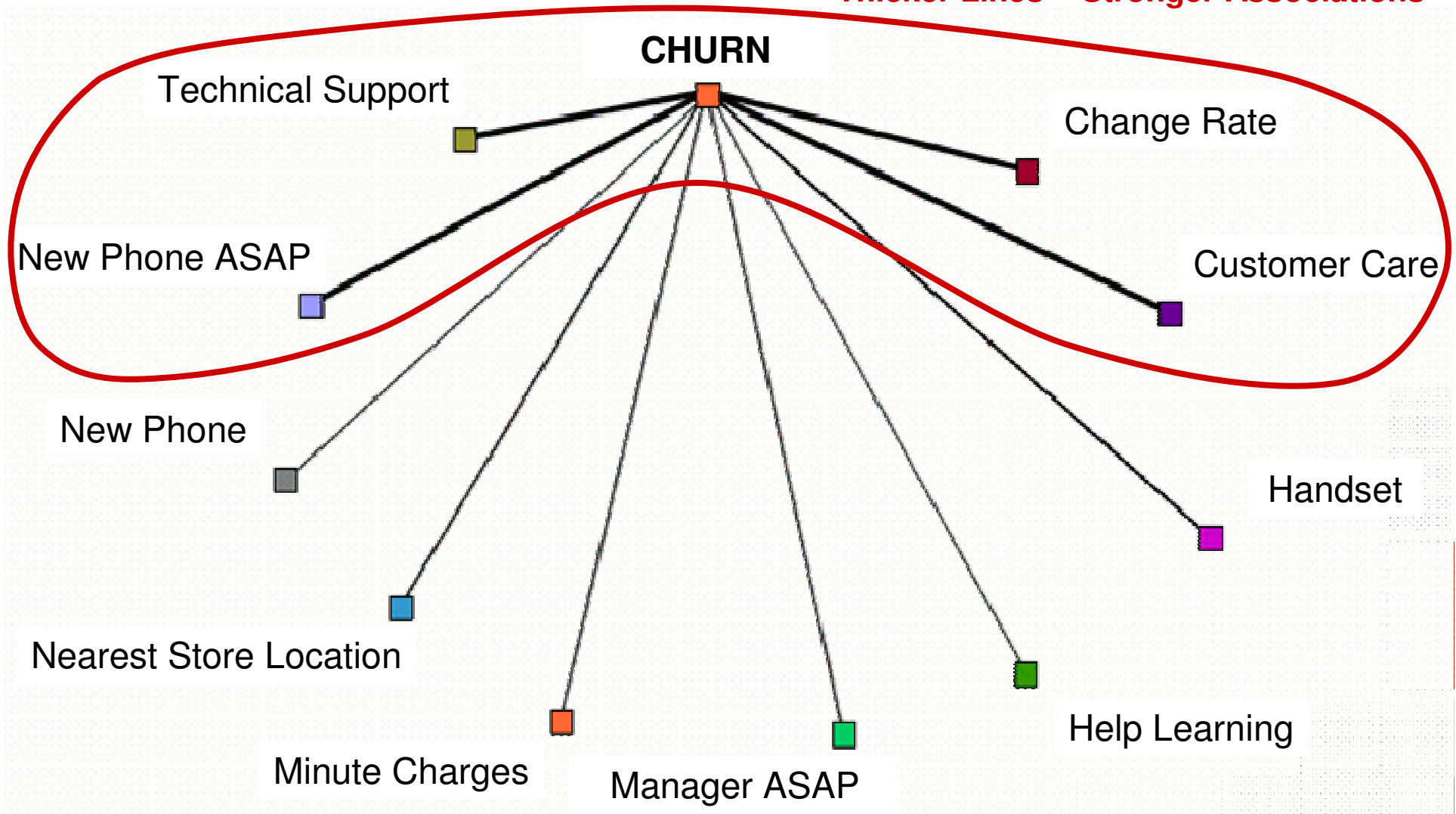
> Mire jó a szöveganalitika?

- Security
- Attitűd (vélemény) azonosítás (pl. ügyfélszolgálat, piackutatás)
- Pharma
 - vényadatok
 - digitalizált levéltárak

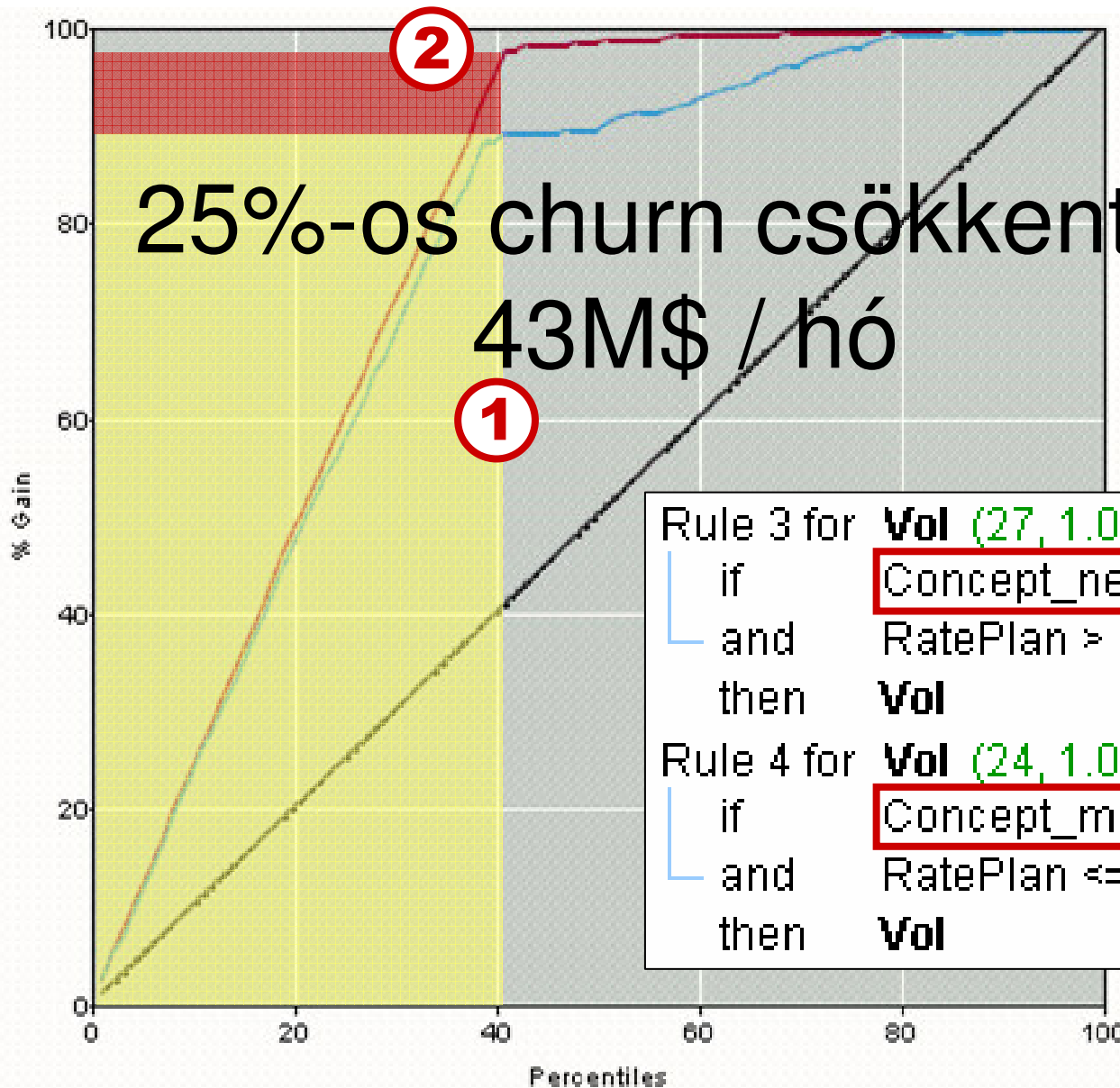


> Churnmodell javítása

Thicker Lines = Stronger Associations



> CRM - telco

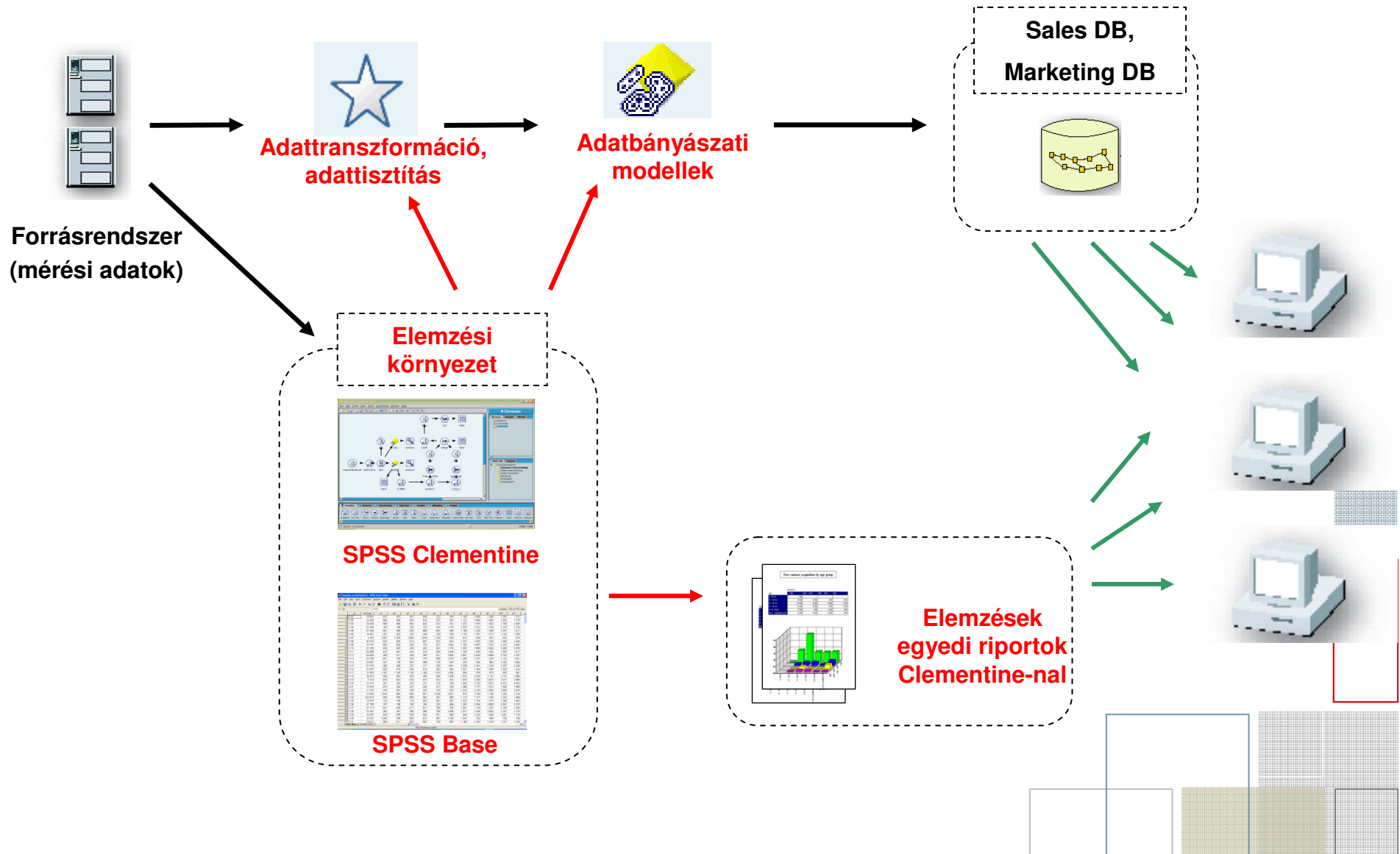


1. Hagyományos churn modell
2. Text mining eredményekkel

```
Rule 3 for Vol (27, 1.0)
  if Concept_new_phone_asap = 1
  and RatePlan > 3
  then Vol
Rule 4 for Vol (24, 1.0)
  if Concept_minute_charges = 1
  and RatePlan <= 1
  then Vol
```

> Alkalmazás

SPSS



> Kérdések?

Körmendi György
gykormendi@spss.hu
30-400-1854