Kalcsó Gyula
Digitális Bölcsészeti Központ

Archivált webes tartalom kutatási célú hasznosítása

Adatbányászat, szövegbányászat, adatvizualizáció

Miről lesz szó?

1. A webaratás anyaga mint a digitális bölcsészeti kutatás forrása

2. Szövegkorpuszok építése a webarchívumból

3. Szöveg- és adatbányászat

4. Adatvizualizáció

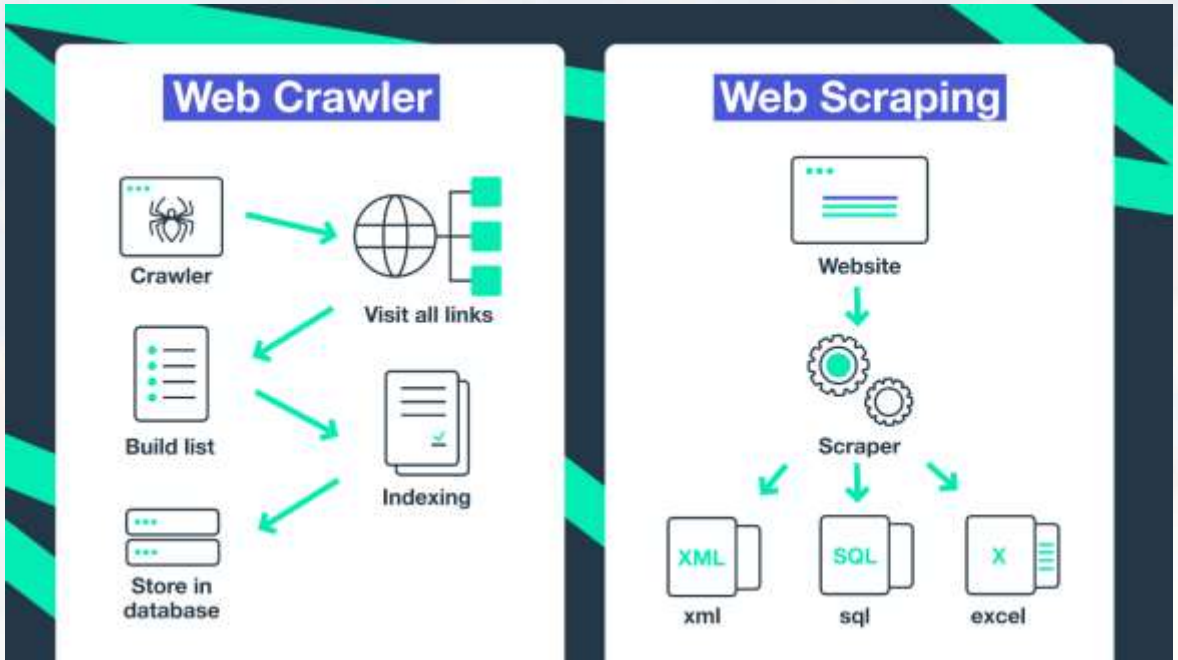
1. A webaratás anyaga mint a digitális bölcsészeti kutatás forrása

A webarchívumok mint a kutatások forrásai

- A weben található hatalmas szövegmennyiség a nyelvmodellek építéséhez (amelyek a nyelvtechnológiai fejlesztésekhez szükségesek)
- A szövegek tartalma (különböző bölcsészet- és társadalomtudományi kutatások: szentimentanalízis, diskurzuselemzés stb.)
- Az AV tartalom elemzése, illetőleg felhasználása a mesterséges intelligencia tanítására (image labeling, object detecting stb.)
- Az adatok (akár nyersen találhatóak a weben, akár a webes tartalomból nyerjük ki őket) összefüggéshálózatainak a kutatása
- (Big data) adatvizualizáció

A webarchívumok mint a kutatások forrásai

- Az adat kétféleképpen állítható elő: web crawling vs. web scraping
- A crawling eredménye nemcsak szöveget, hanem mást is tartalmaz, viszont előnye, hogy több metaadat és kontextuális információ bontható ki belőle
- A crawling munkafolyamata: crawler → aratott tartalom (WARC) → speciális célú script –> szöveg v. más adat
- A scraping munkafolyamata: speciális célú script (scraper) → szöveg v. más adat



Az Ukrajna-gyűjtemény

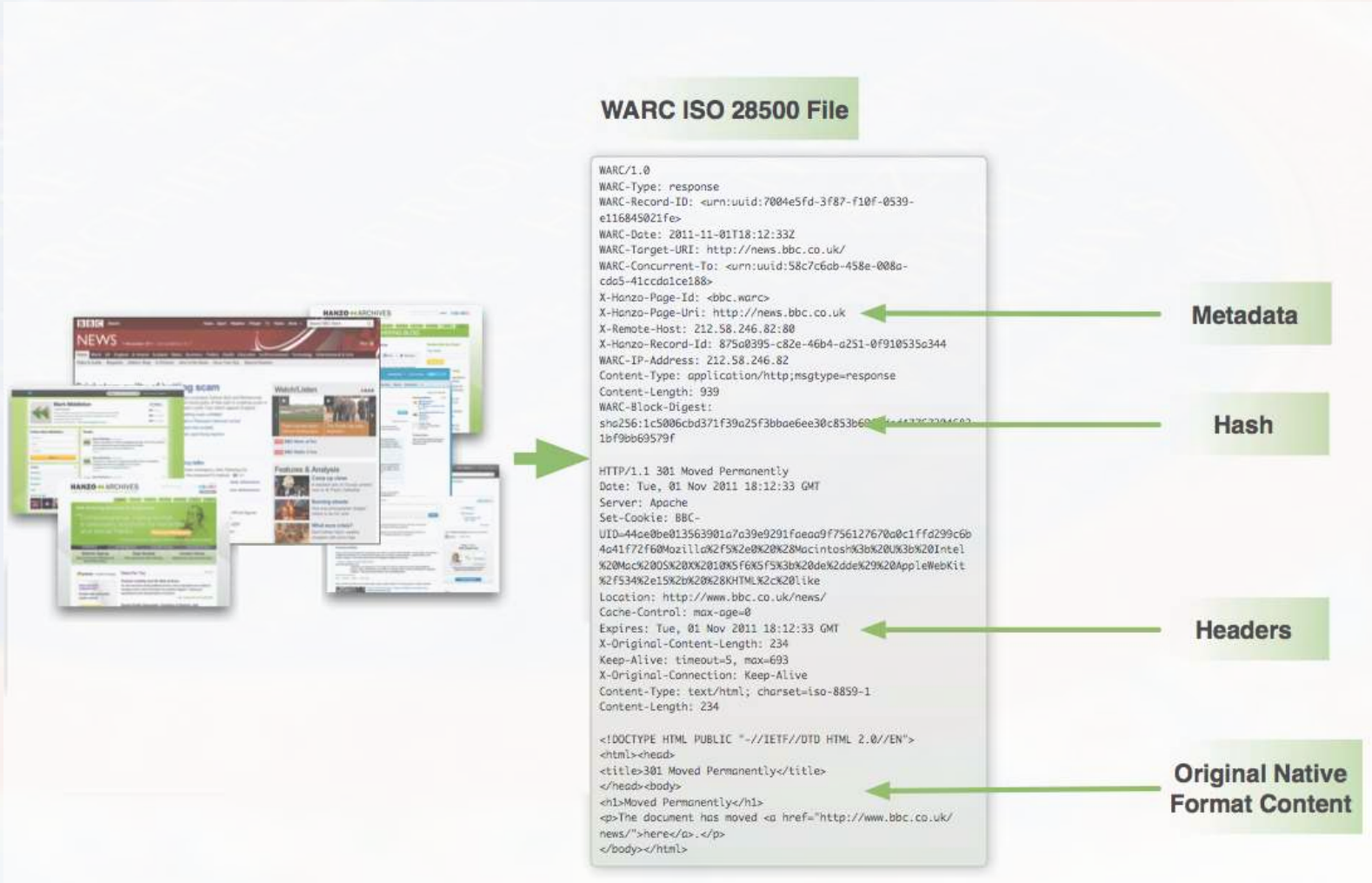
- Az OSZK a webtérszintű és a tematikus aratások mellett eseményalapú gyűjtéseket is készít a jelentősebb kulturális, politikai és sporteseményekről.
- 2022. február 21. óta elkezdtük gyűjteni az orosz–ukrán konfliktussal majd később háborúval kapcsolatos híreket, 75 magyarországi és határon túli hírportálról.
- A hírek gyűjtése alapvetően a portálokon használt címkék vagy kategóriák alapján történik (ez 445 seed URL-t jelent).
- Ezek a mentések hetente egyszer futnak.
- A gyűjtemény jogi okok miatt nem nyilvános, ugyanakkor teljes szövegű keresővel kereshető: https://ukrajnapublic.webharvest.oszk.hu/solrwayback/

Böngészés: Orosz-ukrán konfliktus – 2022

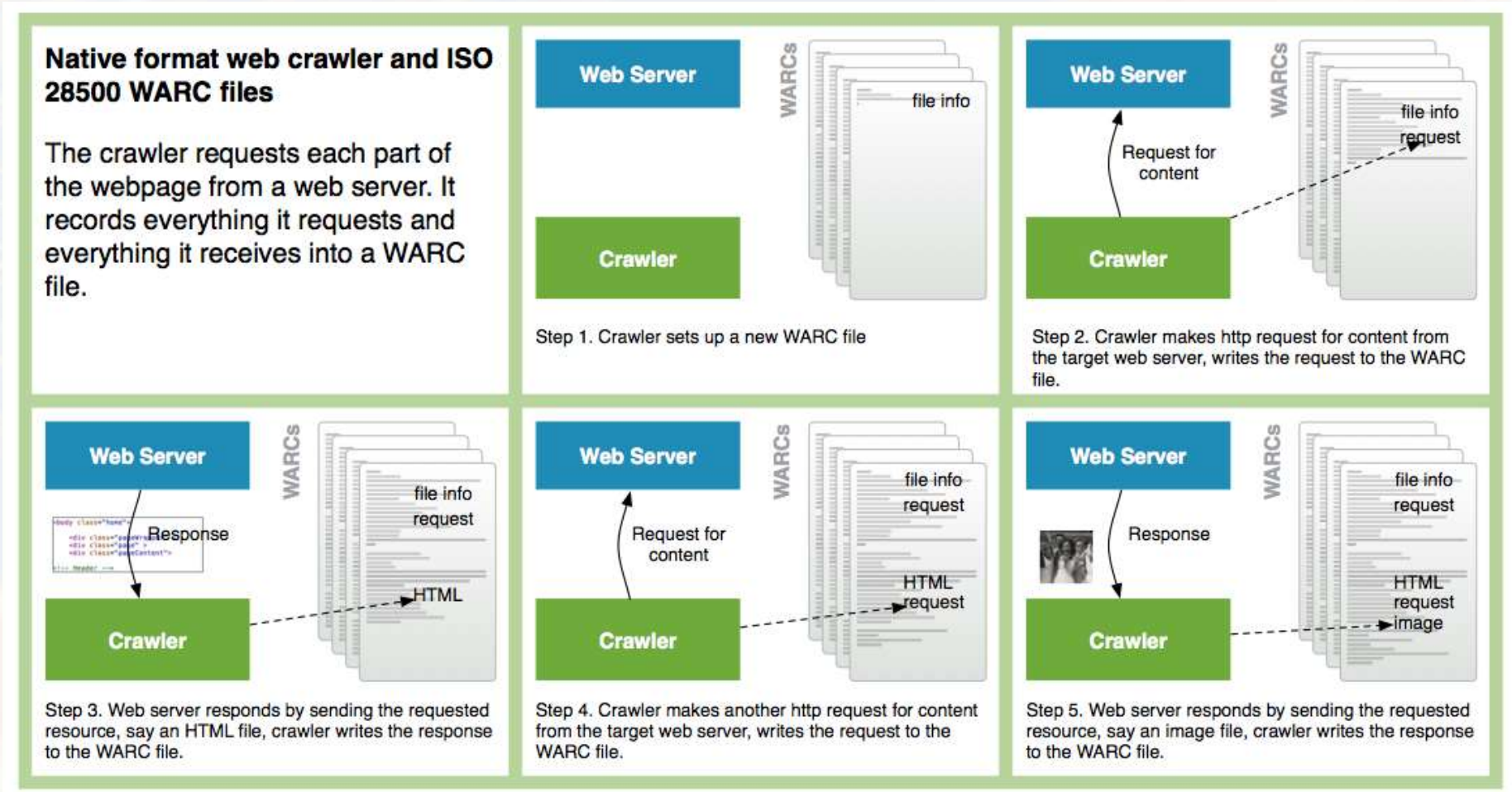
NÉV		URL
Híroldalak és tévécsatornák		
168 óra tematikus címke	https://168.hu/kereses/cimke/orosz-102936	https://168.hu/kereses/cimke/zelens
24 hu tematikus címke és rovat	| https://24.hu/tag/ukrajna/	https://24.hu/tag/ukran-orosz-konflikt
444 tematikus címke és rovat	https://444.hu/ukran-haboru	https://444.hu/tag/ukrajna | https://44
Alfahír tematikus címke	https://alfahir.hu/cimkek/orosz_ukran_konfliktus	https://alfahir.hu/cimkek/donyec_mec
ATV tematikus címke	https://www.atv.hu/cimke/ukrajna/	https://www.atv.hu/cimke/orosz-ukrar
AzÜzlet tematikus címke	https://azuzlet.hu/cimke/ukrajna/	https://azuzlet.hu/cimke/orosz-ukran
Azonnali tematikus címke	https://azonnali.hu/dosszie/ukrajna	https://azonnali.hu/cimke/ukran-oros
Blikk tematikus címke	https://www.blikk.hu/ukrajna-27129	https://www.blikk.hu/orosz-ukran-hab
Borsod24 tematikus címke	https://borsod24.hu/tag/ukrajna/	https://borsod24.hu/tag/putyin/ | https
Bumm.sk tematikus címke	https://www.bumm.sk/cimke/ukrajna	https://www.bumm.sk/cimke/kijev | ht
Daily News Hungary tematikus címke	https://dailynewshungary.com/tag/ukraine/	https://dailynewshungary.com/tag/wa
Demokrata tematikus címke	https://demokrata.hu/cimke/ukran-orosz-konfliktus/	https://demokrata.hu/cimke/ukrajna/
ERDON tematikus címke	https://www.erdon.ro/cimke/ukran-valsag	https://www.erdon.ro/cimke/orosz-uk
Eurázsia tematikus címke	https://www.eurazsia.hu/tag/ukrajna/	https://www.eurazsia.hu/category/oro
Euronews tematikus címke	https://hu.euronews.com/tag/ukrajna	https://hu.euronews.com/tag/orosz-u
Femina tematikus címke	https://femina.hu/cimke/orosz-ukr%C3%A1n+konfliktus	https://femina.hu/cimke/ukrajna
Független Hírügynökség tematikus címke	https://fuhu.hu/cimke/ukrajna/	https://fuhu.hu/cimke/oroszorszag/ |
G7 tematikus címke	https://g7.hu/tag/orosz-ukran-konfliktus/	https://g7.hu/tag/ukrajna/ | https://g7
Helló Magyar tematikus címke	https://hellomagyar.hu/tag/ukrajna/	https://hellomagyar.hu/tag/oroszorsz
Hír TV keresés	https://hirtv.hu/kereses?search_txtf=Ukrajna	https://hirtv.hu/kereses?search_txtf=
Hír.ma tematikus címke	https://hir.ma/tag/Orosz-Ukr%C3%A1n+h%C3%A1bor%C3%BA	https://hir.ma/tag/Ukrajna | https://hir
hirado.hu tematikus címke	https://hirado.hu/hely/ukrajna/	https://hirado.hu/tema/ukran-valsag/
hírek.sk tematikus címke	https://www.hirek.sk/cimke/orosz-ukran-konfliktus	https://www.hirek.sk/cimke/orosz-ukr
Hírkereső keresés	https://www.hirkereso.hu/search?q=ukrajna&timelimit=720	https://www.hirkereso.hu/search?q=u
Hírstart tematikus címke	https://www.hirstart.hu/fk/ukrajna	https://www.hirstart.hu/fk/moszkva | h
Hungary Today tematikus címke	https://hungarytoday.hu/tag/ukrainian-war/	https://hungarytoday.hu/tag/ukrainian
HVG online tematikus címke	https://hvg.hu/cs/orosz-ukran%20konfliktus	https://hvg.hu/360/cimke/orosz-ukr%
Index tematikus címke és rovat	https://index.hu/24ora/?cimke=ukrajna	https://index.hu/24ora/?cimke=orosz
Infostart tematikus címke	https://infostart.hu/cimke/ukrajna	https://infostart.hu/cimke/oroszorsz%
Itt honról haza tematikus címke	https://itthonrolhaza.hu/tag/ukrajna/	https://itthonrolhaza.hu/tag/donyeck/

2. Szövegkorpuszok építése a webarchívumból

A webarchívumok szabványos formátuma: WARC

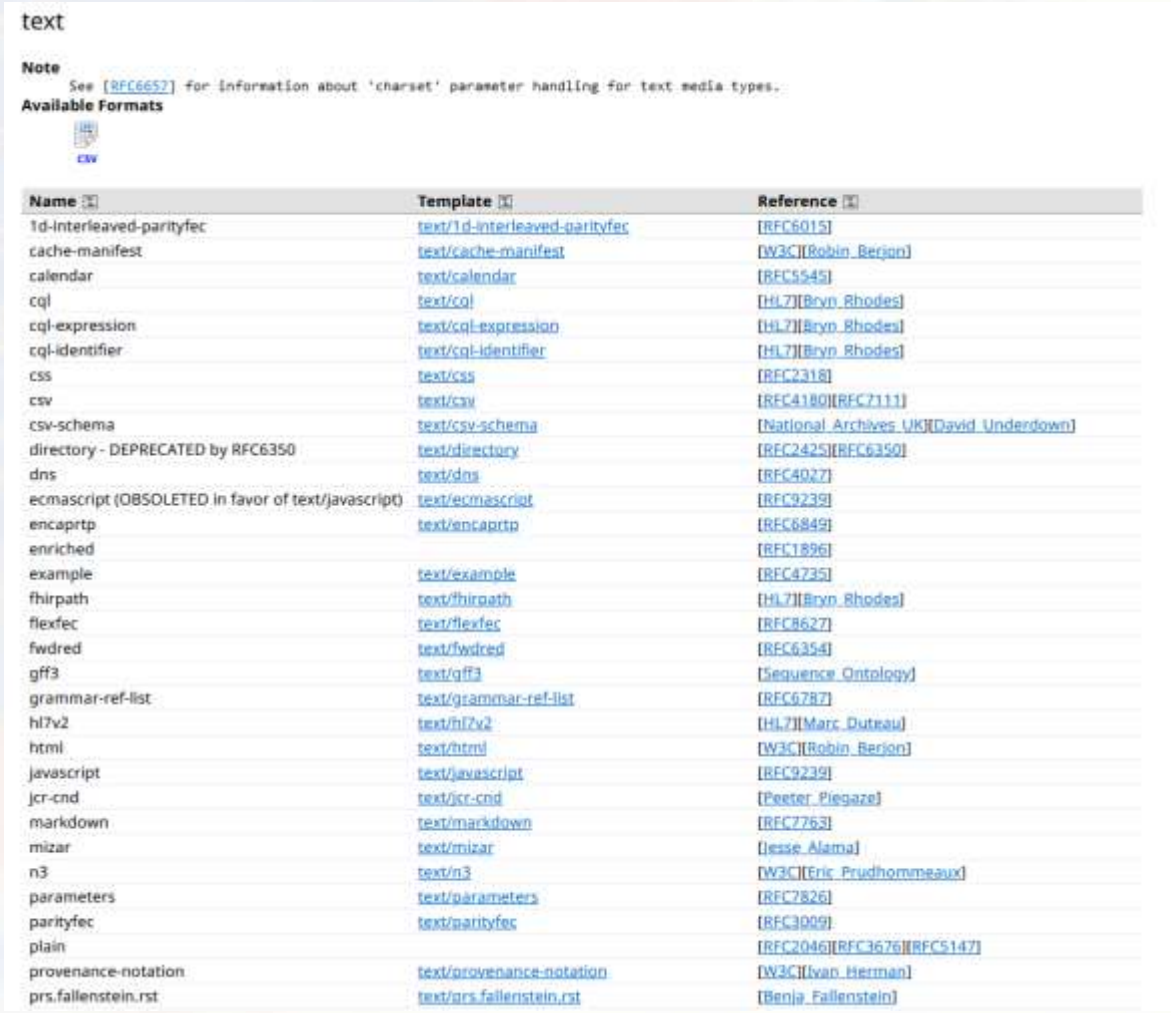


A webarchívumok szabványos formátuma: WARC



Hol a szöveg a WARC-ban? (WARC2TXT)

- Hivatalosan 92 MIME-típus van, amely text formátumú.
- A hírek szövege a HTML-ben található.
- Az első lépés tehát: WARC2HTML



text

Note

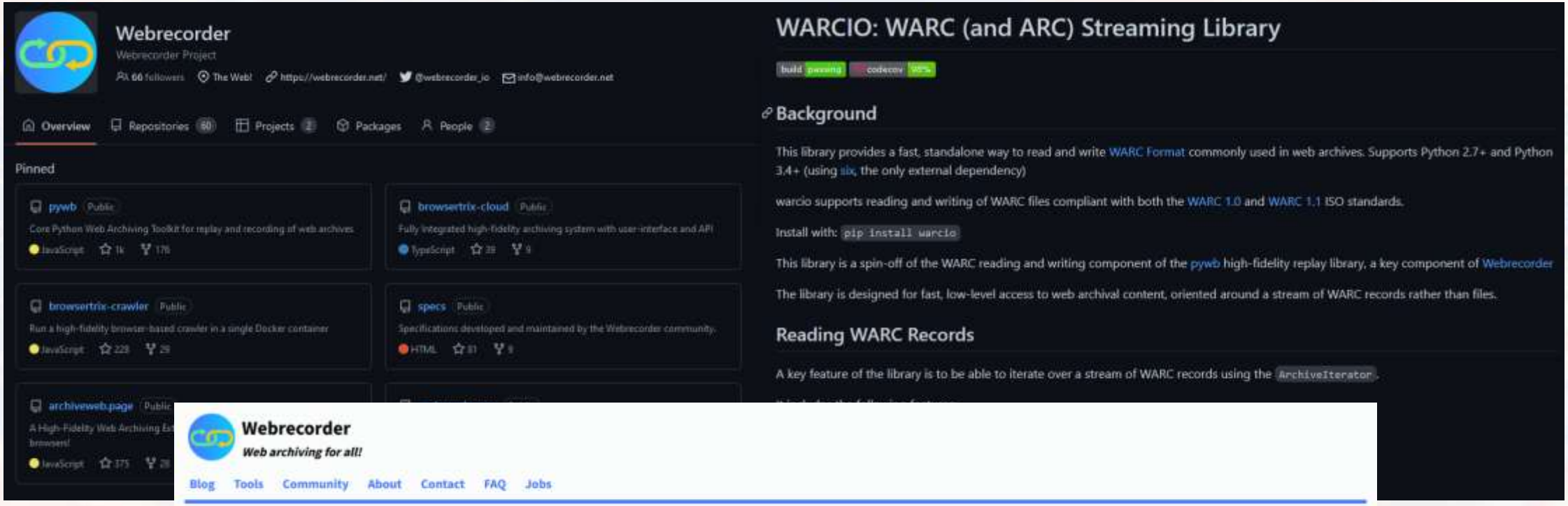
See [RFC6657] for information about 'charset' parameter handling for text media types.

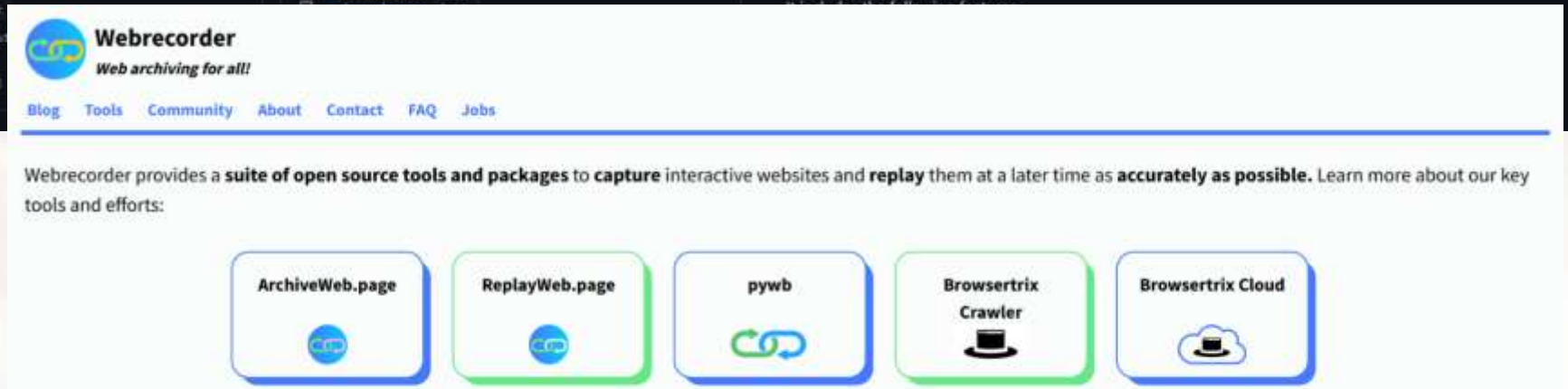
Available Formats

csv

Name	Template	Reference
1d-interleaved-parityfec	text/1d-interleaved-parityfec	[RFC6015]
cache-manifest	text/cache-manifest	[W3C][Robin_Berjon]
calendar	text/calendar	[RFC5545]
cql	text/cql	[HL7][Bryn_Rhodes]
cql-expression	text/cql-expression	[HL7][Bryn_Rhodes]
cql-identifier	text/cql-identifier	[HL7][Bryn_Rhodes]
css	text/css	[RFC2318]
csv	text/csv	[RFC4180][RFC7111]
csv-schema	text/csv-schema	[National_Archives_UK][David_Underdown]
directory - DEPRECATED by RFC6350	text/directory	[RFC2425][RFC6350]
dns	text/dns	[RFC4027]
ecmascript (OBSOLETED in favor of text/javascript)	text/ecmascript	[RFC9239]
encaprtp	text/encaprtp	[RFC6849]
enriched		[RFC1896]
example	text/example	[RFC4735]
fhirpath	text/fhirpath	[HL7][Bryn_Rhodes]
flexfec	text/flexfec	[RFC8627]
fwdred	text/fwdred	[RFC6354]
gff3	text/gff3	[Sequence_Ontology]
grammar-ref-list	text/grammar-ref-list	[RFC6787]
hl7v2	text/hl7v2	[HL7][Marc_Duteau]
html	text/html	[W3C][Robin_Berjon]
javascript	text/javascript	[RFC9239]
jcr-cnd	text/jcr-cnd	[Peeter_Piegaze]
markdown	text/markdown	[RFC7763]
mizar	text/mizar	[Jesse_Alama]
n3	text/n3	[W3C][Eric_Prudhommeaux]
parameters	text/parameters	[RFC7826]
parityfec	text/parityfec	[RFC3009]
plain		[RFC2046][RFC3676][RFC5147]
provenance-notation	text/provenance-notation	[W3C][Ivan_Herman]
prs.fallenstein.rst	text/prs.fallenstein.rst	[Benja_Fallenstein]

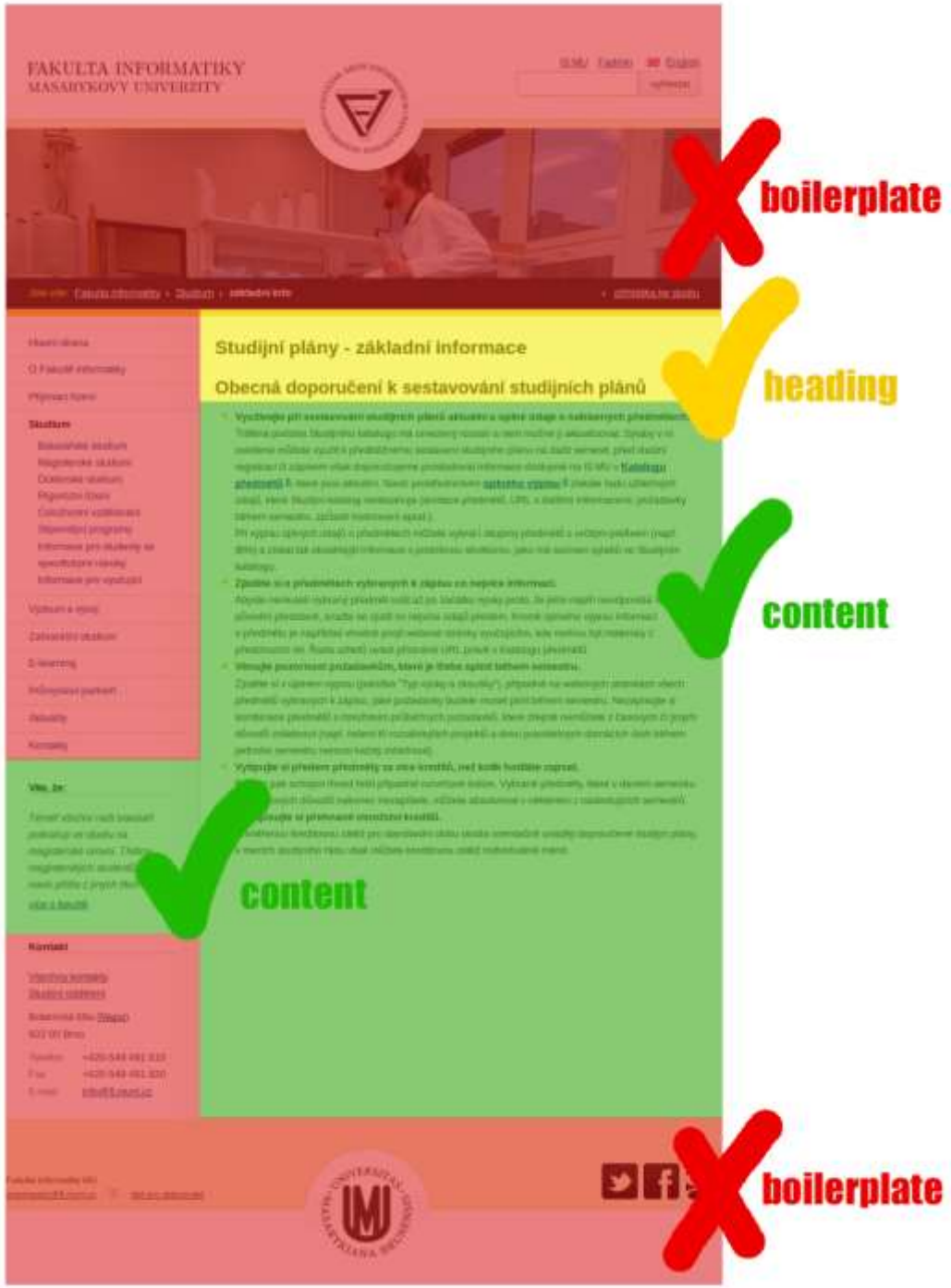
WARC2HTML


Webrecorder
Webrecorder Project
Overview
Repositories
Projects
Packages
People
Pinned
pywb
Core Python Web Archiving Toolkit for replay and recording of web archives
browsertrix-cloud
Fully integrated high-fidelity archiving system with user-interface and API
browsertrix-crawler
Run a high-fidelity browser-based crawler in a single Docker container
specs
Specifications developed and maintained by the Webrecorder community.
archiveweb.page
WARCIO: WARC (and ARC) Streaming Library
Background
This library provides a fast, standalone way to read and write WARC Format commonly used in web archives. Supports Python 2.7+ and Python 3.4+ (using six, the only external dependency)
warcio supports reading and writing of WARC files compliant with both the WARC 1.0 and WARC 1.1 ISO standards.
Install with: pip install warcio
This library is a spin-off of the WARC reading and writing component of the pywb high-fidelity replay library, a key component of Webrecorder
The library is designed for fast, low-level access to web archival content, oriented around a stream of WARC records rather than files.
Reading WARC Records
A key feature of the library is to be able to iterate over a stream of WARC records using the ArchiveIterator


Webrecorder
Web archiving for all!
Blog
Tools
Community
About
Contact
FAQ
Jobs
Webrecorder provides a suite of open source tools and packages to capture interactive websites and replay them at a later time as accurately as possible. Learn more about our key tools and efforts:
ArchiveWeb.page
ReplayWeb.page
pywb
Browsertrix Crawler
Browsertrix Cloud

WARC2HTML

Dátum	HTML	Tokenszám
2022. február 21.	82000	26000000
2022. február 28.	77913	26480366
2022. március 7.	78737	26846341
2022. március 14.	80500	26828470
2022. március 21.	80807	26000000
2022. március 22.	94046	28000000
2022. március 28.	102096	31000000
2022. április 5.	97427	29000000
2022. április 11.	95666	31000000
2022. április 18.	94260	31000000
2022. április 25.	100361	33000000
2022. május 2.	98847	33808641
2022. május 9.	100667	34760808
2022. május 16.	105468	35081005
2022. május 23.	102542	34389562
2022. május 30.	101333	32987217
2022. június 6.	105505	33273179
Összesen	**1598175**	**519455589**



HTML2TXT

jusText Python-library:

- két lépésben (a kontextus és a nyelvi tartalom alapján) leválasztja az ún. boilerplate-et a HTML-ből
- az eredmény: plain text

3. Szöveg- és adatbányászat

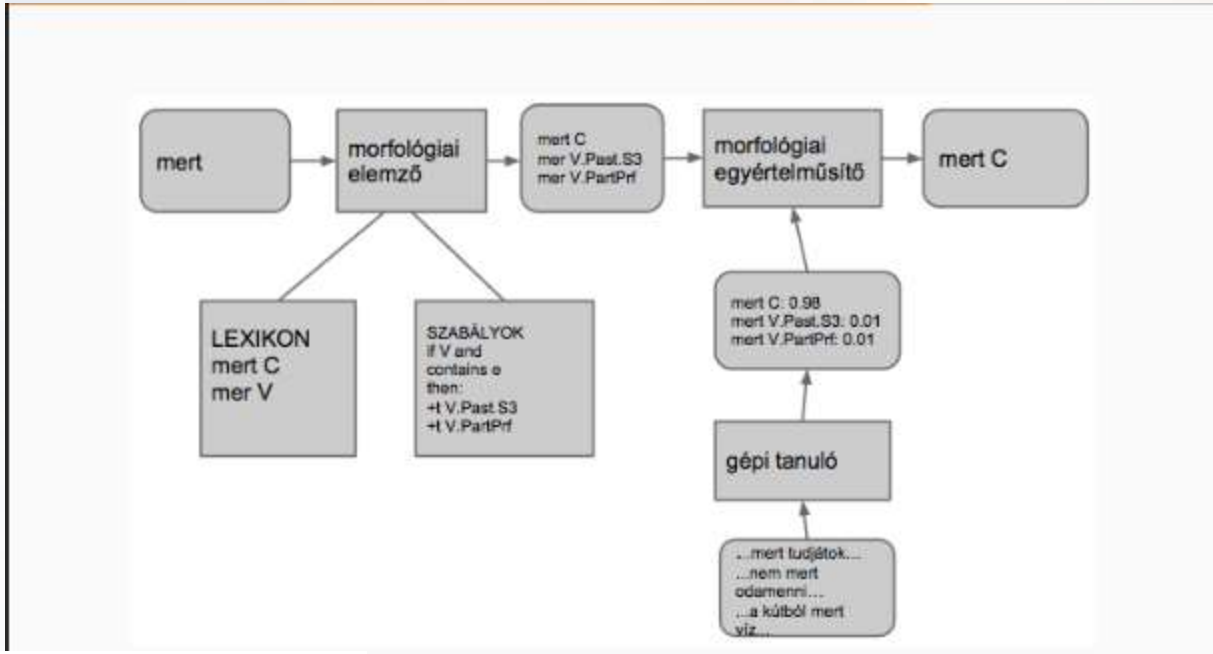
Nyelvi elemzés

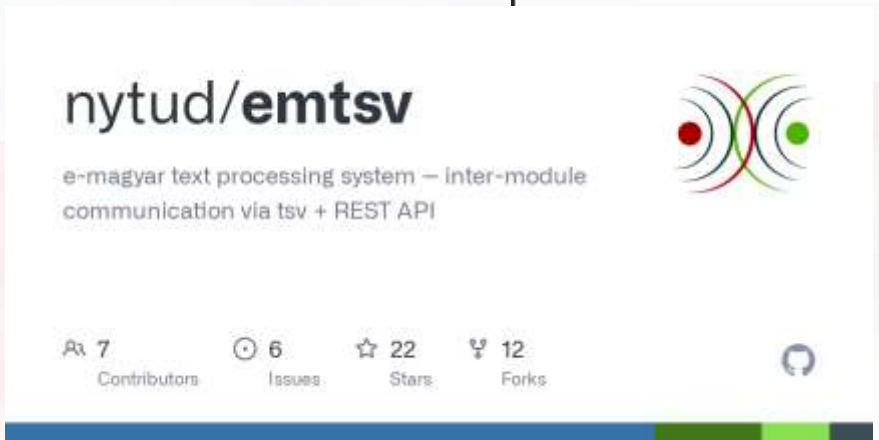
A nyelvi elemzés célja:

- mondatokra bontás
- szavakra bontás
- morfológiai elemzés
- morfológiai egyértelműsítés

Miért szükséges?

- külön kezelhetők a szófajok (vár)
- kivehetőek az adatbányászatban irreleváns elemek (kötőszavak, névutók stb.)





Nyelvileg elemzett szöveg

```
1  form    wsafter anas    lemma   xpostag upostag feats   id      deprel  head
2  Komolyan        " "     [{"lemma": "komoly", "tag": "[/Adj][_Manner/Adv]", "morphana": "komoly[/Adj]=komoly+an[_Manner/Adv]=an", "readable": "komoly[/Adj] + an[_Mann
3  gondolom        ""      [{"lemma": "gondol", "tag": "[/V][Prs.Def.1Sg]", "morphana": "gondol[/V]=gondol+om[Prs.Def.1Sg]=om", "readable": "gondol[/V] + om[Prs.Def.1Sg
4  .       " "     [{"lemma": ".", "tag": "[Punct]", "morphana": "", "readable": "", "twolevel": ""}]      .       [Punct] PUNCT   _       3       PUNCT   0
5
6  Független       " "     [{"lemma": "független", "tag": "[/Adj][Nom]", "morphana": "függ[/V]=függ+etlen[_NegPtcp/Adj]=etlen+[Nom]=", "readable": "függ[/V] + etlen[_Ne
7  főpolgármester-jelöltként       " "     [{"lemma": "főpolgármester-jelölt", "tag": "[/Adj][EssFor:ként]", "morphana": "fő[/Adj]=fő+polgármester[/N]=polgármester+-[Hy
8  a       " "     [{"lemma": "a", "tag": "[/Det|Art.Def]", "morphana": "a[/Det|Art.Def]=a", "readable": "a[/Det|Art.Def]", "twolevel": "a:a :[/Det|Art.Def]"}, {"lemma"
9  Sétáló  " "     [{"lemma": "sétáló", "tag": "[/Adj][Nom]", "morphana": "sétál[/V]=sétál+ó[_ImpfPtcp/Adj]=ó+[Nom]=", "readable": "sétál[/V] + ó[_ImpfPtcp/Adj] + [Nom]
10 Budapest        " "     [{"lemma": "Budapest", "tag": "[/N][Nom]", "morphana": "Budapest[/N]=Budapest+[Nom]=", "readable": "Budapest[/N] + [Nom]", "twolevel": "B:B u
11 vízióját        " "     [{"lemma": "vízió", "tag": "[/N][Poss.3Sg][Acc]", "morphana": "vízió[/N]=vízió+ja[Poss.3Sg]=já+t[Acc]=t", "readable": "vízió[/N] + ja[Poss.3S
12 képviselem      ""      [{"lemma": "képvisel", "tag": "[/V][Prs.Def.1Sg]", "morphana": "képvisel[/V]=képvisel+em[Prs.Def.1Sg]=em", "readable": "képvisel[/V] + em[Prs
13 ,       " "     [{"lemma": ",", "tag": "[Punct]", "morphana": "", "readable": "", "twolevel": ""}]      ,       [Punct] PUNCT   _       8       PUNCT   7
14 ez      " "     [{"lemma": "ez", "tag": "[/Det|Pro][Nom]", "morphana": "ez[/Det|Pro]=ez+[Nom]=", "readable": "ez[/Det|Pro] + [Nom]", "twolevel": "e:e z:z :[/Det|Pro]
15 az      " "     [{"lemma": "az", "tag": "[/Det|Art.Def]", "morphana": "az[/Det|Art.Def]=az", "readable": "az[/Det|Art.Def]", "twolevel": "a:a z:z :[/Det|Art.Def]"},
16 egyetlen        " "     [{"lemma": "egyetlen", "tag": "[/Num][Nom]", "morphana": "egyetlen[/Num]=egyetlen+[Nom]=", "readable": "egyetlen[/Num] + [Nom]", "twolevel":
17 programpontom   ""      [{"lemma": "programpont", "tag": "[/N][Poss.1Sg][Nom]", "morphana": "program[/N]=program+pont[/N]=pont+om[Poss.1Sg]=om+[Nom]=", "readable": "
18 ,       " "     [{"lemma": ",", "tag": "[Punct]", "morphana": "", "readable": "", "twolevel": ""}]      ,       [Punct] PUNCT   _       13      PUNCT   7
19 ezt     " "     [{"lemma": "ez", "tag": "[/Det|Pro][Acc]", "morphana": "ez[/Det|Pro]=ez+t[Acc]=t", "readable": "ez[/Det|Pro] + t[Acc]", "twolevel": "e:e z:z :[/Det|P
20 fogom   " "     [{"lemma": "fog", "tag": "[/V][Prs.Def.1Sg]", "morphana": "fog[/V]=fog+om[Prs.Def.1Sg]=om", "readable": "fog[/V] + om[Prs.Def.1Sg]", "twolevel": "f:f
21 a       " "     [{"lemma": "a", "tag": "[/Det|Art.Def]", "morphana": "a[/Det|Art.Def]=a", "readable": "a[/Det|Art.Def]", "twolevel": "a:a :[/Det|Art.Def]"}, {"lemma"
22 következő       " "     [{"lemma": "következő", "tag": "[/Adj][Nom]", "morphana": "következik[/V]=következ+ő[_ImpfPtcp/Adj]=ő+[Nom]=", "readable": "következik[/V]=kö
23 egy     " "     [{"lemma": "egy", "tag": "[/Det|Art.NDef]", "morphana": "egy[/Det|Art.NDef]=egy", "readable": "egy[/Det|Art.NDef]", "twolevel": "e:e g:g y:y :[/Det|A
24 évben   " "     [{"lemma": "év", "tag": "[/N][Ine]", "morphana": "év[/N]=év+ben[Ine]=ben", "readable": "év[/N] + ben[Ine]", "twolevel": "é:é v:v :[/N] b:b e:e n:n :[
25 propagálni      ""      [{"lemma": "propagál", "tag": "[/V][Inf]", "morphana": "propagál[/V]=propagál+ni[Inf]=ni", "readable": "propagál[/V] + ni[Inf]", "twolevel":
26 ,       " "     [{"lemma": ",", "tag": "[Punct]", "morphana": "", "readable": "", "twolevel": ""}]      ,       [Punct] PUNCT   _       21      PUNCT   15
27 ezért   " "     [{"lemma": "ez", "tag": "[/Det|Pro][Cau]", "morphana": "ez[/Det|Pro]=ez+ért[Cau]=ért", "readable": "ez[/Det|Pro] + ért[Cau]", "twolevel": "e:e z:z :[
28 fogok   " "     [{"lemma": "fog", "tag": "[/V][Prs.NDef.1Sg]", "morphana": "fog[/V]=fog+ok[Prs.NDef.1Sg]=ok", "readable": "fog[/V] + ok[Prs.NDef.1Sg]", "twolevel": "
29 mindenhol       " "     [{"lemma": "mindenhol", "tag": "[/Adv|Pro]", "morphana": "mindenhol[/Adv|Pro]=mindenhol", "readable": "mindenhol[/Adv|Pro]", "twolevel": "m:m
30 kiállni ""      [{"lemma": "kiáll", "tag": "[/V][Inf]", "morphana": "ki[/Prev]=ki+áll[/V]=áll+ni[Inf]=ni", "readable": "ki[/Prev] + áll[/V] + ni[Inf]", "twolevel": "
```

Adatbányászat az elemzett szövegből

- Adattisztítás: az írásjelek és egyéb, szükségtelen adatok (pl. benne maradt biolerplate) eltávolítása
- A megfelelő adatelemek kinyerése: a TSV feldolgozása (a fejlécek alapján Linux-parancsokkal feldaraboljuk, a szükséges oszlopokat tartjuk meg)
- További szűrések: a szófelhőhöz pl. az ún. viszonyszók eltávolítása (névelők, névutók, névmások, kötőszók)

```
cat <jobnév_dátum>_text2emtsv.txt | cut -f
6,7 | tr "\t" "_" | sed 's/\[\/Supl\]//' |
sed 's/^.*\[Punct\].*$//' | sed
's/^.*\[\/Prev\].*$//'| sed
's/^.*\[\/X\].*$//'| sed
's/^.*\[\/Num.*$//' | sed
's/^.*\[\/Cnj.*$//' | sed
's/^.*\[\/Det.*$//' | sed
's/^.*\[\/Post.*$//' | sed
's/^.*|Pro.*$//' | sed '/^$/d' | cut -f
1,2 -d "[" | sort | uniq -c | sort -nr |
sed 's/^ \+//' | sed 's/ /\t/' | sed
's/_/\t/' >
/mnt/cifs/fs01/dbk/warc2text-automate-test
/06-wordfreq-list-lemma/<dátumos_mappa>/au
to-filtered-list/<jobnév_dátum>_wordfreq-l
ist-lemma_auto-filtered-list.tsv
```

Kreatív tartalmak

Adatvizualizáció

Az orosz–ukrán háború a magyar online sajtóban

A magyar hírportálok orosz–ukrán háborúról szóló cikkeinek szóhasználata képekben

Megnéztük, hogyan változik a háborús hírek szókészlete február 21-e óta.

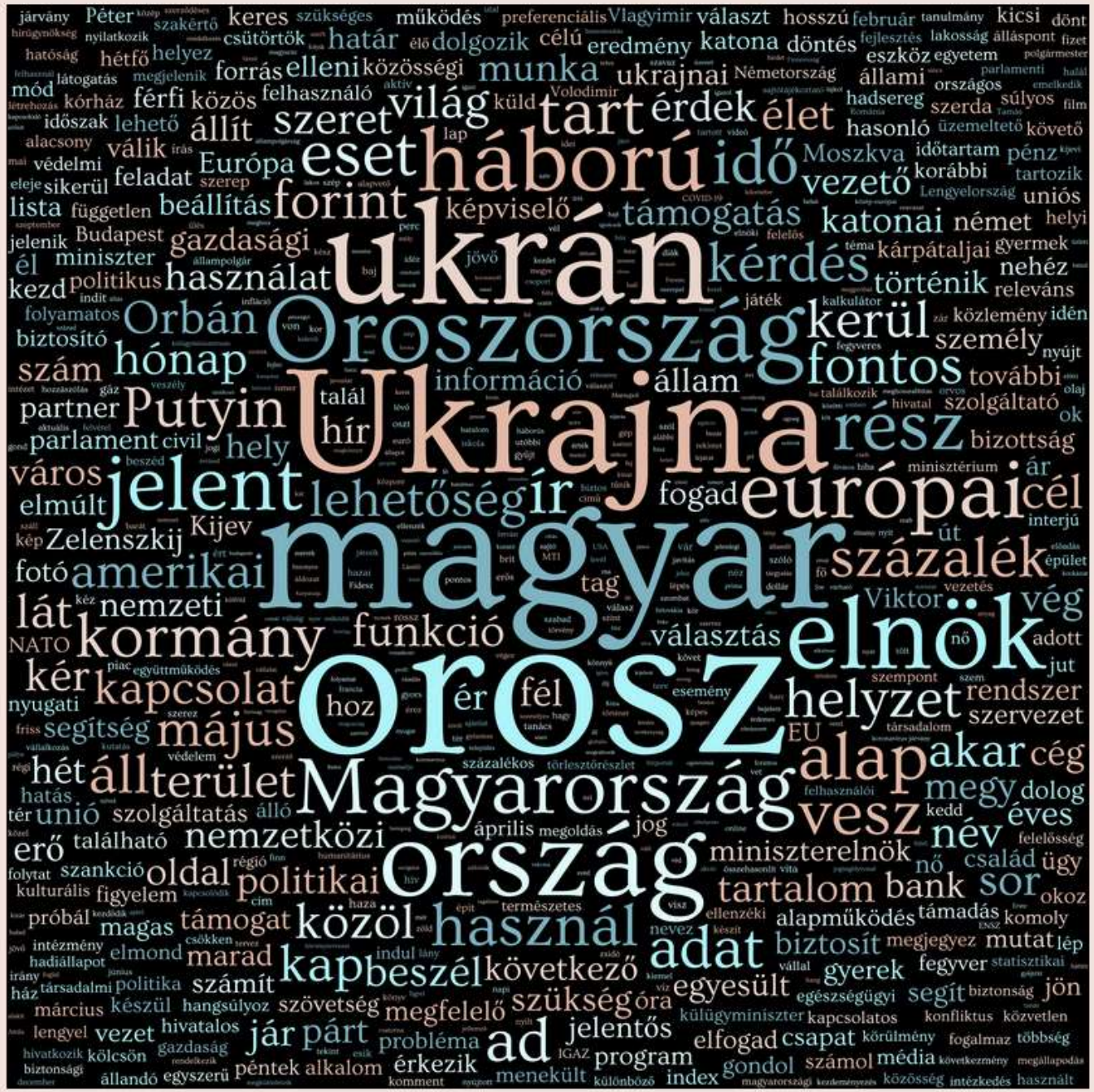
Interaktív

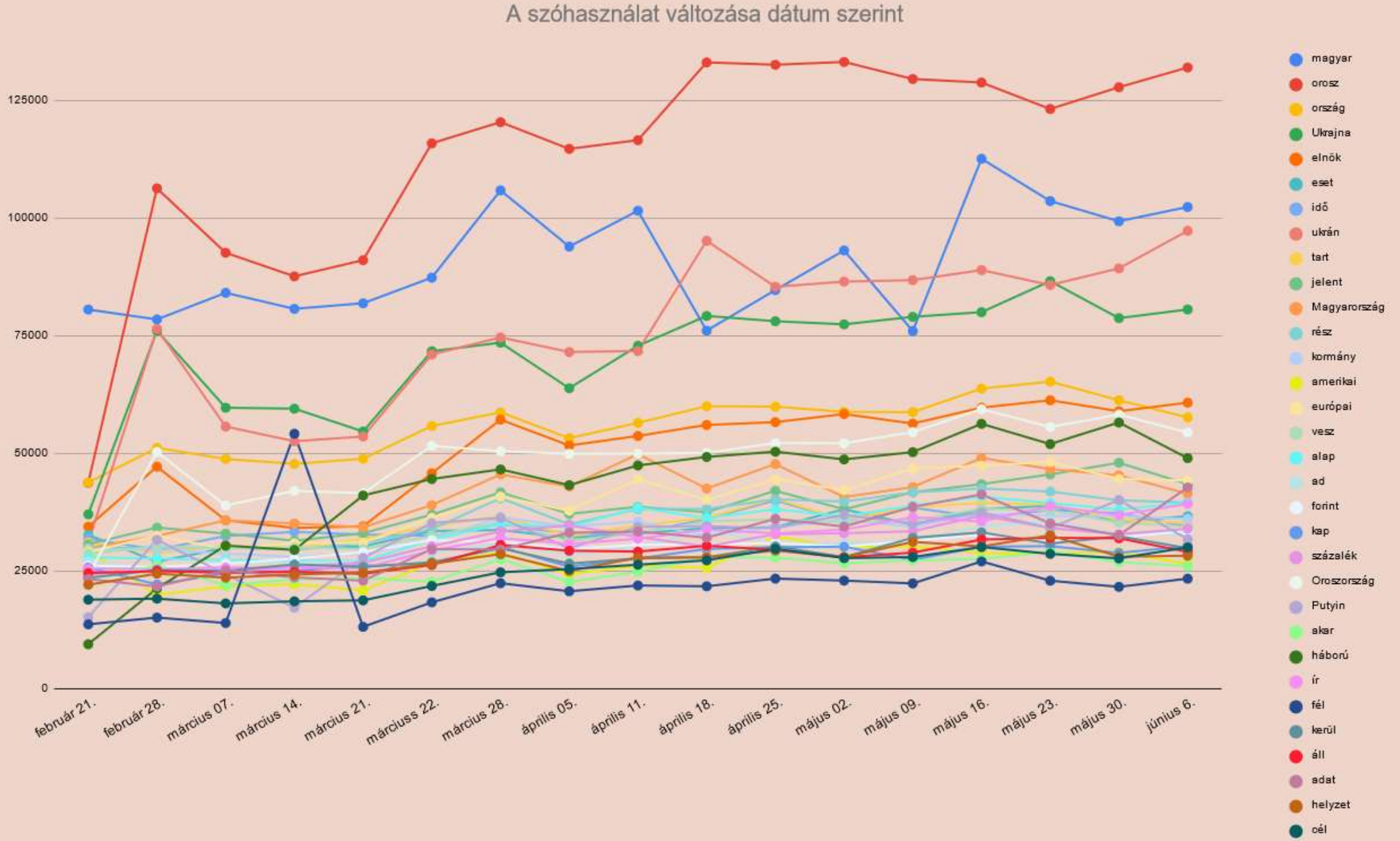
Az orosz–ukrán háború a magyar online sajtóban

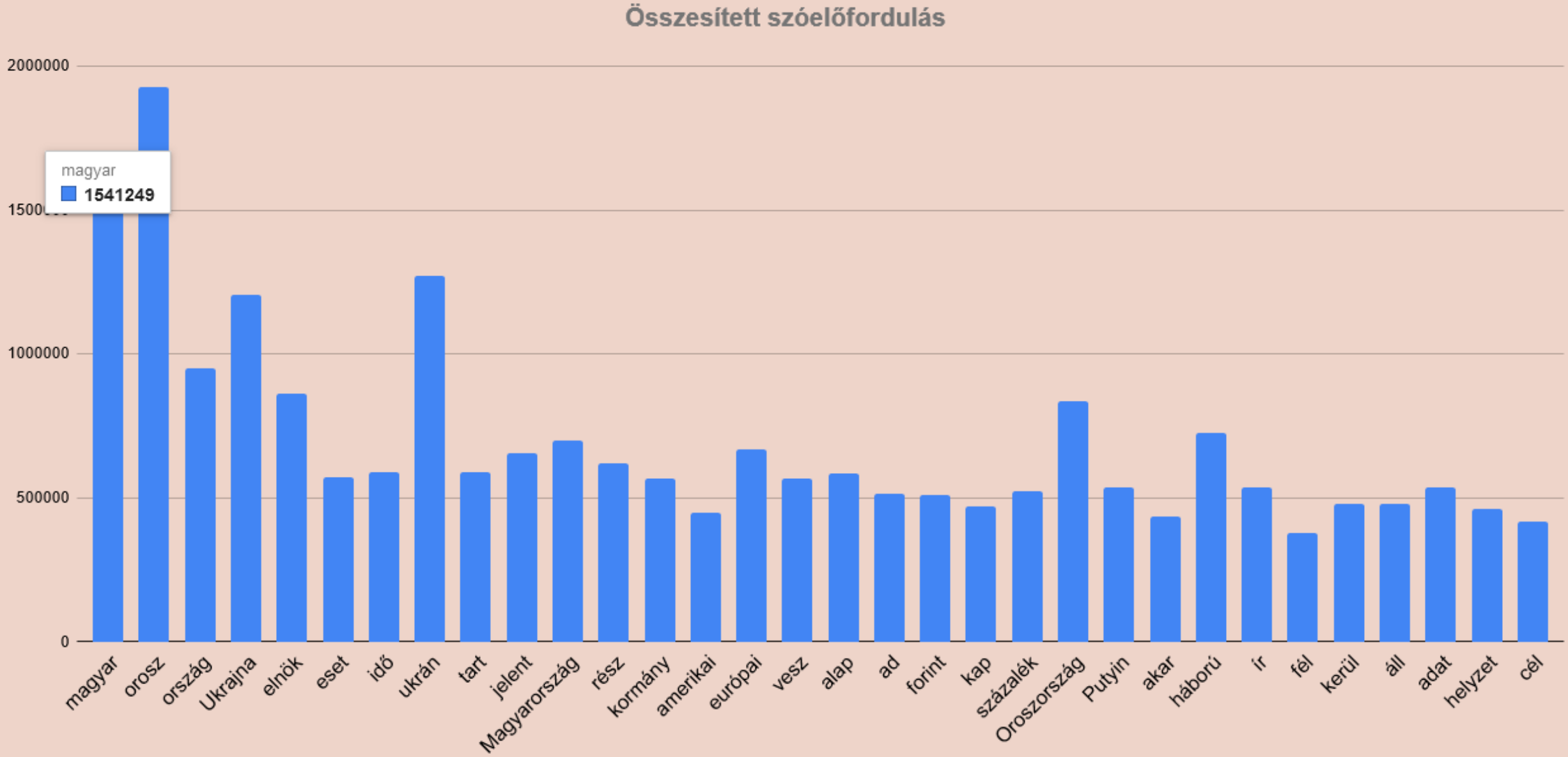
A magyar hírportálok orosz–ukrán háborúról szóló cikkeinek szóhasználata böngészhető, lekérdezhető, testreszabható, interaktív vizualizációk formájában

Fedezd fel az adatok mögött rejtőző információkat!

2022. május 23.




A szóhasználat változása dátum szerint
125000
100000
75000
50000
25000
0
február 21.
február 28.
március 07.
március 14.
március 21.
márciuss 22.
március 28.
április 05.
április 11.
április 18.
április 25.
május 02.
május 09.
május 16.
május 23.
május 30.
június 6.
magyar
orosz
ország
Ukrajna
elnök
eset
idő
ukrán
tart
jelent
Magyarország
rész
kormány
amerikai
európai
vesz
alap
ad
forint
kap
százalék
Oroszország
Putyin
akar
háború
ír
fél
kerül
áll
adat
helyzet
cél


Összesített szóelőfordulás
2000000
1000000
500000
0
magyar
1541249
magyar
orosz
ország
Ukrajna
elnök
eset
idő
ukrán
tart
jelent
Magyarország
rész
kormány
amerikai
európai
vesz
alap
ad
forint
kap
százalék
Oroszország
Putyin
akar
háború
ír
fél
kerül
áll
adat
helyzet
cél

A kétezer leggyakoribb szó

Gyakoriság	Szó	Szófaj
131928	orosz	[/Adj]
97231	ukrán	[/Adj]
80518	Ukrajna	[/N]
80372	magyar	[/Adj]
60714	elnök	[/N]
57589	ország	[/N]
54383	Oroszország	[/N]
48932	háború	[/N]
44162	európai	[/Adj]
43397	jelent	[/V]
42702	adat	[/N]
41371	Magyarország	[/N]
39476	alap	[/N]
39474	rész	[/N]
39192	ír	[/V]
36593	eset	[/N]
36016	kormány	[/N]
35154	tart	[/V]
34424	idő	[/N]
34367	vesz	[/V]
33911	százalék	[/N]
33460	használ	[/V]
33426	forint	[/N]
32754	ad	[/V]
31716	Putyin	[/N]
30751	hónap	[/N]
30658	közöl	[/V]
30265	lehetőség	[/N]
30230	kap	[/V]
29970	cél	[/N]
29801	kér	[/V]
29795	kerül	[/V]
29780	tartalom	[/N]
10754361		

Szófelhő

Szókereső

Keresés

- Az összes kije...
- ad
- adag
- adat
- adó
- adomány
- adott
- aggodalom
- aggódik
- agresszió
- ajánl
- ajánlat
- akad
- akadályoz
- akar
- akció
- aktív
- aktivál
- aktuális
- alábbi
- alacsony
- alagút

Idővonal

M W
Hónap

jún. 2022 - jún. 2022

febr. márc. ápr. máj. jún.

Dátumválasztó

június (Hónap) + 6 (Nap)

A kétezer leggyakoribb szó

Gyakoriság	Szó	Hónap	Nap
131928	orosz	június	6
102292	magyar	június	6
97231	ukrán	június	6
80518	Ukrajna	június	6
60714	elnök	június	6
57589	ország	június	6
54383	Oroszország	június	6
48932	háború	június	6
44162	európai	június	6
43397	jelent	június	6
42702	adat	június	6
41371	Magyarország	június	6
39476	alap	június	6
39474	rész	június	6
39192	ír	június	6
36593	eset	június	6
36016	kormány	június	6
35154	tart	június	6
34424	idő	június	6
34367	vesz	június	6
33911	százalék	június	6
33460	használ	június	6
33426	forint	június	6
32754	ad	június	6
31716	Putyin	június	6
31520	vezető	június	6
30751	hónap	június	6
30658	közöl	június	6
30265	lehetőség	június	6
30230	kap	június	6
29970	cél	június	6
29801	kér	június	6
29795	kerül	június	6
29780	tartalom	június	6
29019	áll	június	6
28627	terület	június	6
28325	élet	június	6
28207	helyzet	június	6
28157	fontos	június	6
28059	kérdés	június	6
28018	funkció	június	6
27714	világ	június	6

A szófajok gyakorisága

Gyakoriság	Szófaj
5874389	[/N]
2607228	[/Adj]
2259698	[/V]
13046	[/Adj|Attr]
10754361	

Szófajválasztó

- [ ] [/Adj]
- [ ] [/Adj|Attr]
- [ ] [/N]
- [ ] [/V]

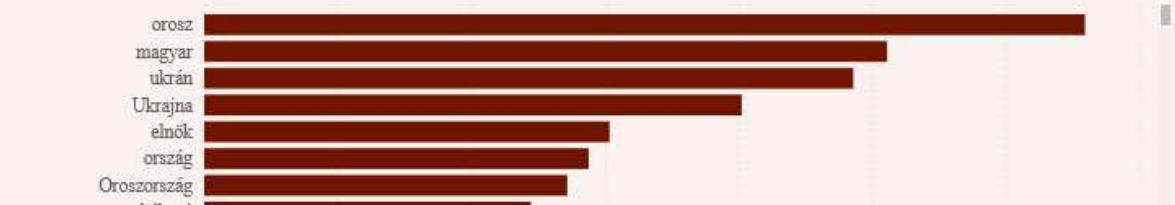
Szókereső

Keresés

- [ ] ad
- [ ] adag
- [ ] adat
- [ ] adó

Szófelhő

Szavak gyakoriság szerint



Dátumválasztó

június (Hónap) + 6 (Nap)

A kétezer leggyakoribb szó

Gyakoriság	Szó	Hónap	Nap
80518	Ukrajna	június	6
60714	elnök	június	6
57589	ország	június	6
54383	Oroszország	június	6
48932	háború	június	6
42702	adat	június	6
41371	Magyarország	június	6
39476	alap	június	6
39474	rész	június	6
36593	eset	június	6
36016	kormány	június	6
34424	idő	június	6
33911	százalék	június	6
33426	forint	június	6
31716	Putyin	június	6
30751	hónap	június	6
30265	lehetőség	június	6
29970	cél	június	6
29780	tartalom	június	6
28627	terület	június	6
28325	élet	június	6
28207	helyzet	június	6
28059	kérdés	június	6
28018	funkció	június	6
27714	világ	június	6
26578	sor	június	6
26121	kapcsolat	június	6
25931	erő	június	6
25724	érdek	június	6
25546	név	június	6
24019	vezető	június	6
23695	használat	június	6
23330	fél	június	6
23268	város	június	6
22960	szám	június	6
22887	hír	június	6
22288	hét	június	6
22152	munka	június	6
22070	vég	június	6
22058	információ	június	6
21920	magyar	június	6
21886	személy	június	6

A szófajok gyakorisága

Gyakoriság	Szófaj
5874389	[/N]
5874389	

Szófajválasztó

- [ ] [/Adj]
- [ ] [/Adj|Attr]
- [x] [/N]
- [ ] [/V]

Szókereső

Keresés

- [ ] adag
- [ ] adat
- [ ] adó
- [ ] adomány

Szófelhő

Szavak gyakoriság szerint

Továbblépési lehetőségek

- A szövegkorpuszok építésében:
  - A metaadatok segítségével finomítani a tartalom osztályozását, szűrését
  - Mesterséges intelligencia bevonásával automatizálni (pl. topic modeling)
  - A cél: több szempont alapján szűrt tematikus korpuszok építésének a lehetősége

- Más webarchivált tartalom adatbányászatában:
  - Egyéb MIME-típusú tartalmak kinyerése, feldolgozása (kép, AV, dokumentumok stb.)
  - Mesterséges intelligencia bevonásával automatizálni (pl. image labeling, object detection, speech to text)

Köszönjük a figyelmet!

Drótos László, Kalcsó Gyula, Makkai Csilla, Mihály Eszter, Simon Eszter, Szűcs Kata, Varga Emese, Végvári Ágnes, Visky Ákos, Vitéz Gábor