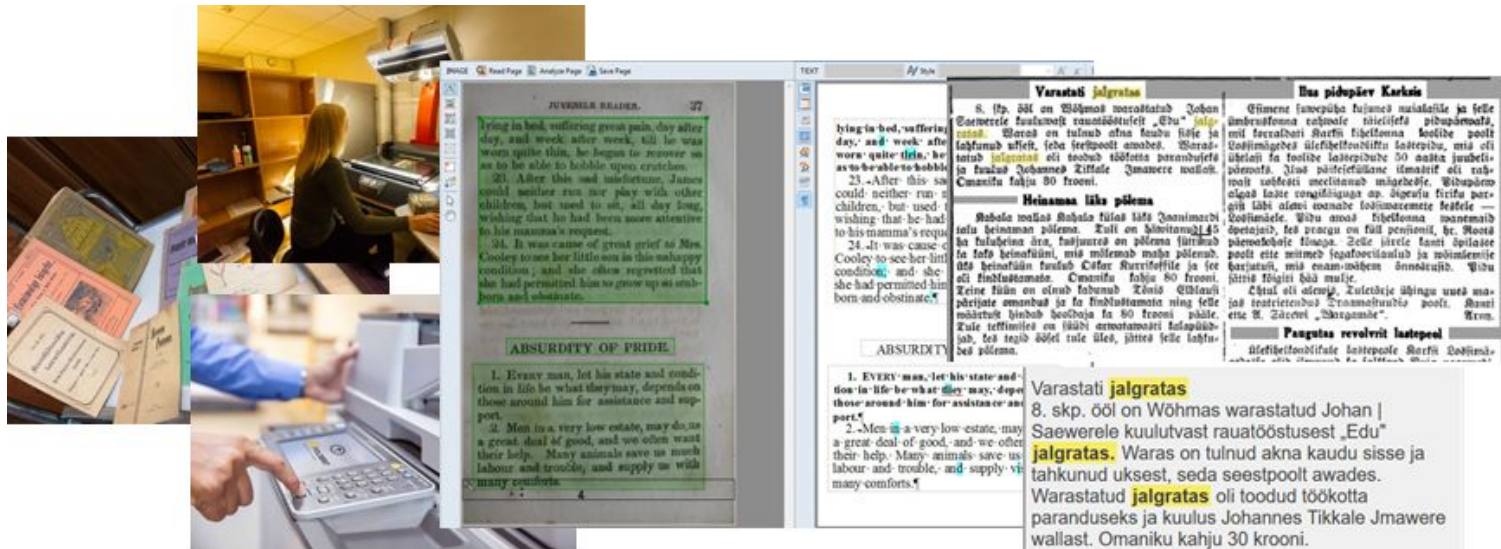

Tekstiainese digikasutus Rahvusraamatukogus: avatud andmed ja avatud kood

Peeter Tinitš, 26. okt 2020
Oskar Kallase päev, Eesti Kirjandusmuuseum

Suured digikollektsioonid



Uurija vajadused

Tekstid



Metaandmed







Tööriistad



Digari otsingud






Otsi ainult kättesaadavaid

Detailne otsing

OTSI

Täistekst

"oskar kallas"



Otsingute ajalugu

relevantsus

Otsisõnale "oskar kallas" leiti 1,376 vastet

<

1

2

3

4

5

...

>

PIIRA OTSINGUT

Väljaande tüüp

Ajakirjad (54)

Ajalehed (1,269)

Jätkväljaanded (53)

Väljaanne

Postimees (1886-1944) (137)

Kaja (92)

Päevaleht (65)

Lisa valik minu loendisse | Lisa kõik minu loendisse

1. Oskar Kallas

☐ Päevaleht 10 august 1925

Püsilink: <https://dea.digar.ee/article/paevalehtew/1925/08/10/12>

Oskar Kallas

Lisa loendisse

Lisa teema

2. Oskar Kallas.

☐ Maa Hääl : maarahva ajaleht 14 juuli 1934

Lisa loendisse

Lisa teema

Digari otsingud

AVALEHTOTSINGVÄLJAANDEDILMUMISAEGTEEMADABI

LOGI SISSEENGESTPYC

> Kaja > 13 aprill 1935 tr. 1

Väljaanne

Artikkel

Dr. Oskar Kallas

<https://dea.digar.ee/article/kaja/1935/04/13/1/69>

Tekst

NB! Tekst võib sisaldada vigu. Loe lähemalt...
Paranda seda teksti. Logi sisse raamatukogu kasutajatunnuse, ID-kaardi või Mobiil-ID-ga

Dr. Oskar Kallas

MI / kutsusutud Helsingi Kalctvala seltsi välismaiseks (Ulkomainen) liikmeks.

Teemad (0)

Väljaanne

Otsingu tulemused

Illede näitus.

amäht illede näituselt oli laupäeval möga...
Samaal püüdnud pühade vahetaval...
...koosil külastasid näitusel...
...nõuad näitusel...
...id olid näitajaid...
...il oli tegemist, et juhtida näitusruumes...
...ist nii, et näitus näeks, mis on mõlga...

Krediidipanga peakoosolek.

Krediidipanga peakoosolekul peeti reedel Tal...
...naas, koosolekul kuulus äge küsimus, kes...
...naas pool tundi.

Clearing Rootsiaga hinn

Mabariigi valituse otsusega...
...ja pandi ajutiselt maksma 26. mär...
...Stofholmis nootide vahetamise leel...
...tud Eesti-Roosji clearingtoffulepe.

Rahvahakultuuri nõuku

esimeheks prof. P. Crei...
...Rahvahakultuuri nõukogu pidas laupäe...
...linnas funktsioneeris mõistes kaitis oma i...
...et koosolekul. Peamotaks oli näitaj...
...torra arutamine, eestikeelse valimine.

Keerulisemad küsimused

DIGAR
EESTI ARTIKLID



 Otsi ainult kättesaadavaid

Detailne otsing

OTSI

Täistekst ▼

mis nimed esinesid samas artiklis oskar kallasega



Otsingute ajalugu

Otsisõnale mis nimed esinesid samas artiklis oskar kallasega leiti 0 vastet

Arengud raamatukogudes

Delpher



jupyter
nbviewer

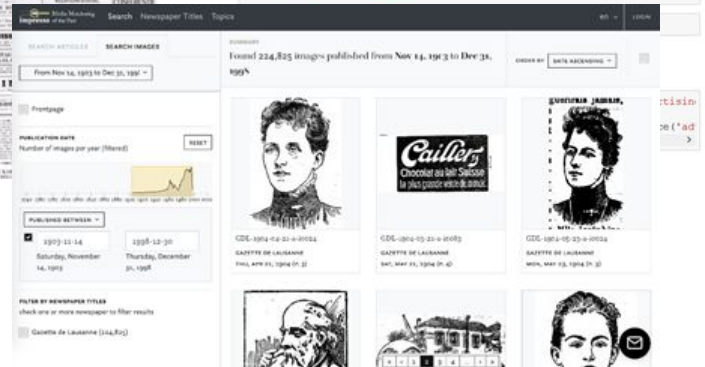
JUPYTER FAQ </> [Icons]

Show when the articles were published

```
In [ ]: alt.Chart(df).mark_line().encode(  
    x='year(date):T',  
    y='count()',  
    tooltip=[alt.Tooltip('year(date):T', title='Year'), alt.Tooltip('cou  
    ).properties(width=600)  
< >
```

Find the longest article

```
In [ ]: # Which is the longest article(s)?  
df[df['words'] == df['words'].max()]
```



Tekstid kui andmed



An illustration of the life of a researcher using
TDM: Beatrice's delivery of the
Nineteenth Century Books Coll
disk for the Palimpsest project.

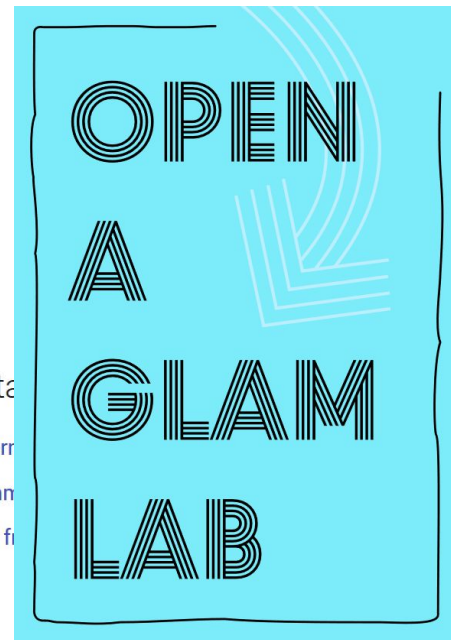
bibendum ac
commodo, vi
dui, eu laore
est, ac venen
commodo. Nulla
id nisi euismod, her
Etiam dolor nulla, pla
felis. Aliquam ultric
egestas sed. Sed impe
felis quis mar
leo, lacinia i
commodo, od
massa, nec ve
vel ipsum fer

Text as Data

Pellentesque
um vestibulum
semper tempor
iles at tortor in
Sed
per, vulputate dolor.
n at, volutpat tincidunt
e, quis sollicitudin ex
at dictum. Integer vitae
Praesent tellus
em. Curabitur
r felis pharetra
vestibulum dui
i. Curabitur sit

GLAM datasets from government data port

- Human readable list of GLAM datasets harvested from govern
- CSV formatted list of GLAM datasets harvested from govern
- CSV formatted list of GLAM datasets (CSVs only) harvested from
(March 2020)



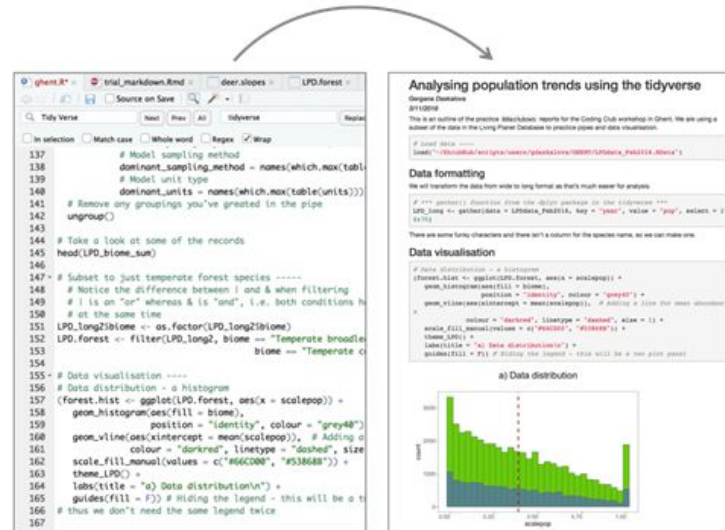
Avatus

FAIR e. “ausad andmed”

- Leitavad
- Ligipääsetavad
- Teineteisega haakuvad
- Taaskasutatavad



Avaandmed teaduses



Data and code availability

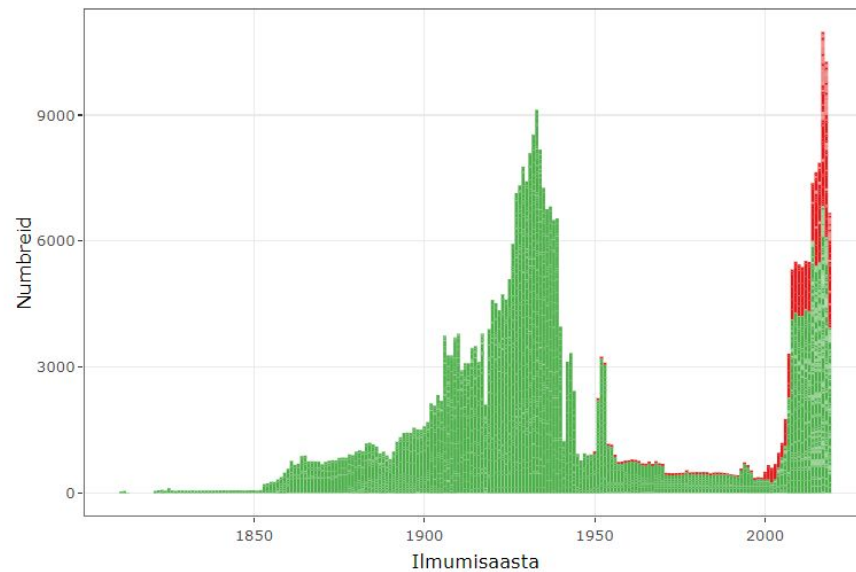
Data and code to reproduce the analysis and figures are available at <https://iof.io/6ysda/>



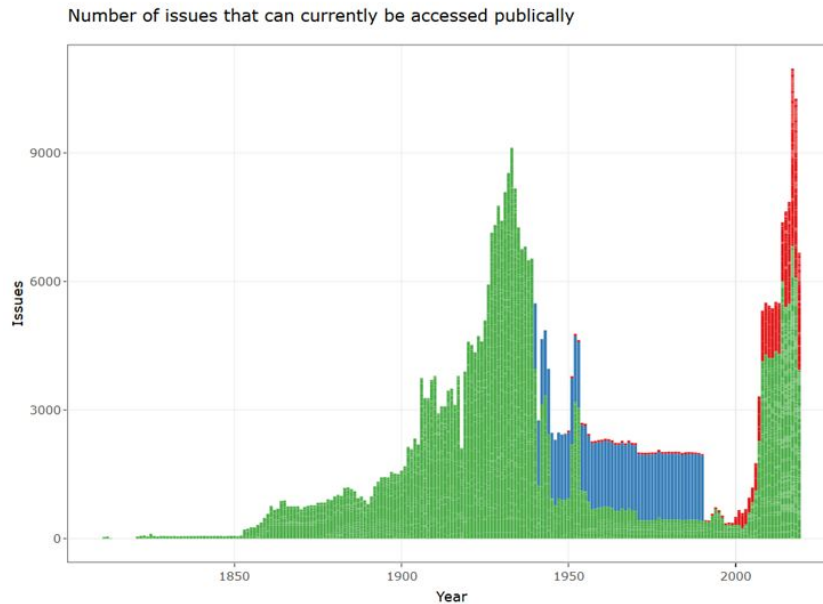
(Heunis 2020)

Avatud materjal

Ligipäas ajalehenumbritele



Avatud materjal



Avatud kood

Andmekogu

Andmekoguna kasutame Eesti Rahvusraamatukogu digilarhiivi Eesti artikleid, millele on olemas tekstikaueveligipääs. Kollektstoont materjalidest saab ülevaate siit http://data.digar.ee/text/dea_info.html. Ligipääs on hetkel ainult koodi läbi

```
```{r}
Loe sisse metaandmete fail hpc serverilt.
all_issues <- fread("unzip -p /gpfs/hpc/projects/digar_txt/text/all_issues_access.zip", sep="\t")[access_now==T]

Valime AJALEHD, 1920 ja 1940 vahel, kus on väljaande koodiks postimeesew
subset <- all_issues[str_detect(DocumentType,"NEWSPAPER")&year>1920&year<1940&keyid=="postimeesew"]

Meile vajalike failide nimekirj
files <- subset[zippath_sections!="",unique(zippath_sections)]
collectionname <- "/gpfs/hpc/projects/digar_txt/text"
filelist <- paste0(collectionname,"/text_sections/", files)

```
```

Tekstide metafailid on samamoodi indekseeritud. Järgmine koodijupp kogub kokku meie otsinguga seotud metainfo.

```
```{r}
metafiles <- subset[zippath_sections!="",unique(zippath_sections_meta)]
metafilelist <- paste0(collectionname,"/meta_sections/", metafiles)

subset_meta <- rbindlist(lapply(paste0("unzip -p ",metafilelist),fread,fill=T),idcol=T)

write_tsv(subset_meta,"subset_meta_postimeesew1.tsv")
```
```

Avatud andmed



ETAIS arvutusklastris

Sign in

Username:

Password:

Sign In

Andmekogu

Andmekoguna kasutane Eesti Rahvusraamatukogu digiarhiivi Eesti artikleid, millele on olemas tekstikaeveligipääs. Kollektiooni materjalidest saab ülevaate siit http://data.digar.ee/text/dea_info.html. Ligipääs on hetkel ainult koodi labi

```
'''{r}
# Loe sisse metaandmete fail hpc serverilt.
all_issues <- fread("unzip -p /gpfs/hpc/projects/digar_txt/text/all_issues_access.zip",sep="\t")[access_now==T]

# Valime AJALEHD, 1928 ja 1940 vahel, kus on väljaande koodiks postneesev
subset <- all_issues[str_detect(DocumentType,"NEWSPAPER")&year~1928&year~1940&keyid=="postneesev"]

# Meile vajalike failide nimekiri
files <- subset(zippath_sections=="",unique(zippath_sections))
collectionname <- "/gpfs/hpc/projects/digar_txt/text"
filelist <- paste0(collectionname,"/text_sections/", files)

...

Tekstide metafailid on samamoodi indekseeritud. Järgmine koodijupp kogub kokku mete otsinguga seotud metainfo.

'''{r}
metafiles <- subset(zippath_sections=="",unique(zippath_sections_meta))
metafilelist <- paste0(collectionname,"/meta_sections/", metafiles)

subset_meta <- rbindlist(lapply(paste0("unzip -p ",metafilelist),fread,fill=T),idcol=T)

write_tsv(subset_meta,"subset_meta_postneesev1.tsv")

...

```

Your server is starting up.

You will be redirected automatically when it's ready for you.

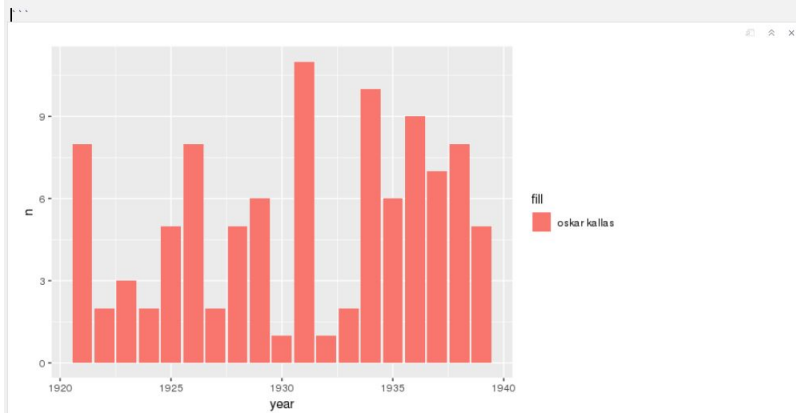


Spawning server...

Rakendus

```
searchfile <- "oskar_kallas.txt"
texts <- read_tsv(searchfile,col_names = c("id","txt"))
texts <- texts %>%
  mutate(year=as.numeric(str_extract(id,"[0-9]{4}")) # kask str_extract võtab ainult esimese vaste, ehk siinkohal nell
  esimest numbrit.

texts %>%
  count(year) %>%
  ggplot(aes(x=year,y=n,fill="oskar kallas"))+
  geom_col()
```



Rakendused

relevantsus Otsisõnale **marlene dietrich** leiti 1,564 vastet

1 2 3 4 5 ... >

PIIRA OTSINGUT

Väljaande tüüp

Ajakirjad (27)

Ajalehed (1,530)

Jatkväljaanded (7)

Lisa valik minu loendisse | Lisa kõik minu loendisse

1. MARLENE DIETRICH

☐ Film ja Elu : Huvitav shumaalleht 26 aprill 1935

Püsiliik: <https://dea.digar.ee/article/filmjaelu/1935/04/26/21>

Lisa loendisse

Lisa teema

jupyter access_pilot (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Find texts with particular keywords within the set

```
In [ ]:
# We can look for transportation devices within the dataset, e.g. motorcycles (motor bikes)
# And we can look
# One thing to note,
searchterm <- "marlene dietrich"
filename <- "marcol.txt"
preloadedcollectionname <- "preloadedcollection1"

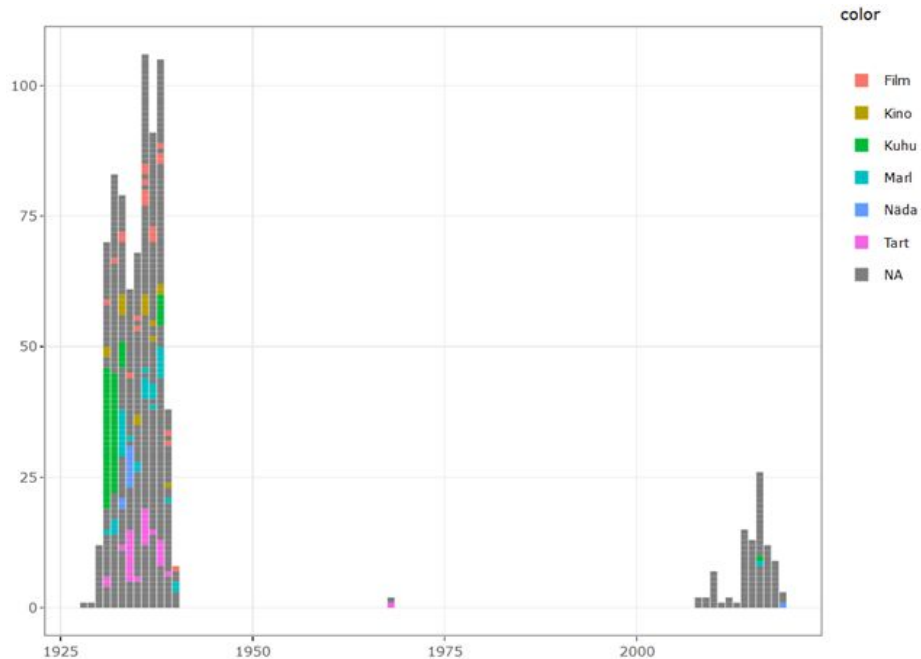
#Technically, each article is in a separate file and in a single row. The same row also contains
system(paste0("rm '", filename, "', for file in ", preloadedcollectionname, "/").snp do unzip -o $file &

#Writing the query out on command line would look like this.
system(paste0("rm search2.txt for file in preloadedcollection/*.snp do unzip -o $file | grep
system(paste0("rm search2.txt for file in preloadedcollection/*.snp do unzip -o $file | grep

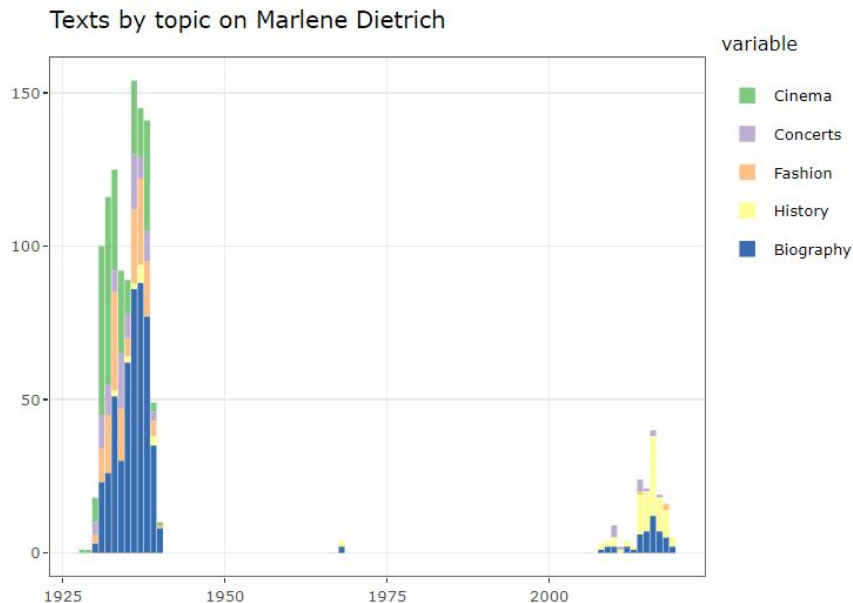
# You can also simply search for nothing, and get all the texts in the collection
system(paste0("for file in preloadedcollection/*.snp do unzip -o $file | grep -i ' ' >> all.txt
```

<p>moemaailmast.</p><p>pressiballöö tähe all<...</p><p>pildinäitaja.</p><p>„pildinäitaja“ ruumis. Väl...</p><p>(seip)habe suur (amotbtti eeskaiva. kaks löökfi...</p><p>jffoefvmas</p><p>paris toodab ka näomood...</p><p>central> teisipäeval, 28. aprillist alates ja edas...</p><p>kino.</p><p>„fiesto“b“t“ orkestrijuhi juubel.</p><p>kuhu minna õhtul</p><p>„estoni a““ teatris ...</p><p>kolywoodiski ott kriis.</p><p>„firma „ühendat...</p><p>laupäev raadios.</p><p>„tallinn. 15.30: päeva...</p><p>kuhu minna õhtul</p><p>„gloria palace“ „d...</p><p>kuhu minna õhtul</p><p>„gloria-palace“ „el...</p><p>naaju.</p><p>friedrich suur jalutas kord berli...</p><p>kuhu minna õhtul</p><p>„gloria-palace“ „g...</p><p>puhu minna õhtul</p><p>„gloria palae“ „p...</p><p>paevalehtew19310209.2.25 19310209 paevalehtew 1931-02-09 2 1931</p><p>nooltatu19310321.2.14 19310321 nooltatu 1931-03-21 3 1931</p><p>postimeesew19310331.2.7.1 19310331 postimeesew 1931-03-31 3 1931</p><p>paevalehtew19310420.2.23 19310420 paevalehtew 1931-04-20 4 1931</p><p>postimeesew19310428.2.7.1 19310428 postimeesew 1931-04-28 4 1931</p><p>paevalehtew19310503.2.25 19310503 paevalehtew 1931-05-03 5 1931</p><p>sonumedi19310516.2.25 19310516 sonumedi 1931-05-16 5 1931</p><p>kaja19310613-1.2.50 19310613 kaja 1931-06-13 6 1931</p><p>sonumedi19310613.2.66 19310613 sonumedi 1931-06-13 6 1931</p><p>sonumedi19310625.2.21 19310625 sonumedi 1931-06-25 6 1931</p><p>sonumedi19310626.2.19 19310626 sonumedi 1931-06-26 6 1931</p><p>nooltatu19310627.2.39 19310627 nooltatu 1931-06-27 6 1931</p><p>sonumedi19310627.2.20 19310627 sonumedi 1931-06-27 6 1931</p><p>sonumedi19310708.2.29 19310708 sonumedi 1931-07-08 7 1931</p>

Rakendused

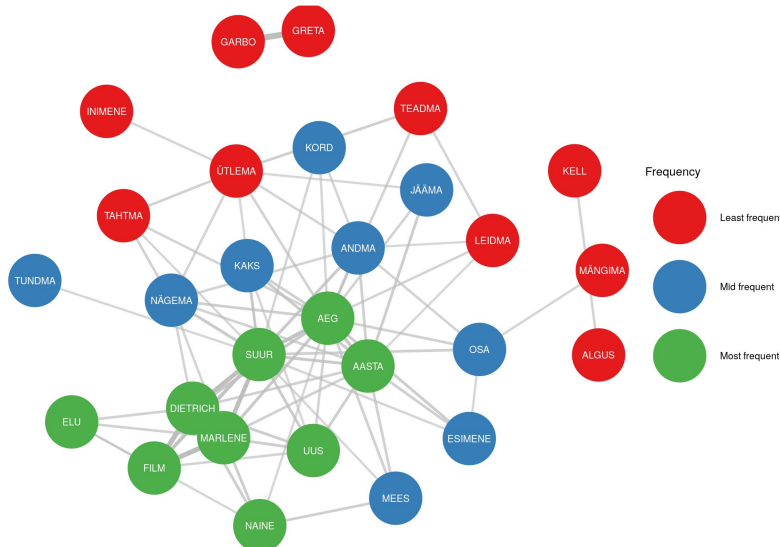


Rakendused



Rakendus

Co-occurrence network of top words



Nüüd-varsti ligipääsetav

~6.3 miljonit artiklit

~3.6 miljonit lk

~390 tuhat numbrit

~2200st väljaandest

Avatud kood,
avatud andmed

